The ABC's of Who Benefits from Working with AI: Ability, Beliefs, and Calibration

Online Appendix

Andrew Caplin, David Deming, Shangwen Li, Daniel Martin, Philip Marx, Ben Weidmann, and Kadachi Jiada Ye

A Robustness Analyses

A.1 Demographic Balance and Controls

Table 3 presents summary statistics by participant treatment group. This includes additional demographic information optionally reported by participants to the experimental platform. The table shows that there was balance across control and treated participant characteristics, with marginally significant differences in birthplace and language when we do not account for multiple hypothesis testing. An issue with optional demographic information – which for clarity is omitted from Table 3 – is that this data is incomplete. In particular, 41% of participants do not report education or employment status, though we find no evidence of imbalanced non-reporting across treatment assignment. Table 4 replicates our main Table 2 when we additionally control for the remaining, mostly non-missing demographic indicators (age, gender, ethnicity, birthplace, language) at baseline and interacted with treatment. This confirms that the results of Table 2 are essentially unchanged if we additionally control for participant demographics.

A.2 Calibration and Confidence

In this appendix we consider alternative specifications and implications of net confidence instead of calibration as a covariate. One possibility is to re-estimate our main specifications of Table (2), but distinguishing between under-confident and over-confident miscalibration. In such a specification, we fail to reject ($p = 0.34$) that the coefficients on

Table 3: Subject Summary Statistics

	Control	Treatment	P-value
Total Time (Minutes)	34.14	34.99	0.44
	(13.42)	(14.13)	
Bonus ($)	4.27	4.29	0.90
	(2.16)	(2.08)	
IQ (0-14)	5.03	4.86	0.42
	(2.80)	(2.79)	
Age	40.74	40.45	0.79
	(14.39)	(13.56)	
Female	0.56	0.58	0.69
	(0.50)	(0.49)	
Nonwhite	0.39	0.35	0.35
	(0.49)	(0.48)	
Born Outside the U.S.	0.10	0.06	0.09
	(0.30)	(0.24)	
First Language Not English	0.06	0.03	0.02
	(0.24)	(0.16)	
Student	0.21	0.16	0.19
	(0.41)	(0.37)	
Employed Full-Time	0.55	0.48	0.13
	(0.50)	(0.50)	
Sample Size	239	493	

Note: The values in parentheses are the sample standard deviations. P-values measure the probability of finding a mean difference as large as the one observed, under the null hypothesis of no difference across groups.

under-confident and over-confident miscalibration are equal. However, such a regression is empirically under-powered for under-confidence given the small number of under-confident participants in the experiment, and so the overall results are not reported.

Instead, we explore the mechanism that under-confident individuals may commit too many Type I errors (following the AI when it is incorrect), while over-confident individuals may commit too many Type II errors (failing to follow the AI when it is correct). As also explained in the main text, in Table 5 we repeat our main specification with net confidence instead of calibration and disaggregating by whether the AI prediction was correct. To disagreggate by image attribute (AI correctness), this regression is at the participant-response level with standard errors clustered by participant. (Note that

Table 4: Regression: Skills and Working with AI, with Demographic Controls (Omitted)

	Outcome: Accuracy (Treated Block)			
Skills (Control Block)	(1)	(2)	(3)	(4)
Treatment	5.01	6.06	5.62	6.24
	(1.98)	(1.79)	(1.78)	(1.83)
Accuracy		3.90	3.80	3.44
		(0.41)	(0.49)	(0.49)
Accuracy × Treatment		-1.51	-2.08	-1.77
		(0.53)	(0.61)	(0.61)
Calibration			0.21	0.09
			(0.47)	(0.46)
Calibration × Treatment			1.35	1.42
			(0.56)	(0.57)
IQ				1.29
				(0.45)
IQ × Treatment				-1.01
				(0.54)
Constant	69.09	66.64	66.69	65.90
	(1.67)	(1.49)	(1.49)	(1.54)
Observations	702	702	702	702

Note: Observations are at the subject level. Robust standard errors are in parentheses. Skills are standardized within session.

previous regressions at the participant level with robust standard errors are equivalent to a pooled regression at the participant-response level with standard errors clustered by participant.) Consistent with the preceding Type I/II error explanation, we find that the less confident are relatively more disadvantaged for images where the AI is ex post incorrect, with a standardized treatment effect across net confidence of 1.93 (s.e.=1.34), whereas the more confident are more disadvantaged for images where the AI is ex post correct, with a standardized treatment effect across net confidence of -2.52 (s.e.=0.75). We note that the results where the AI is incorrect are also under-powered, given that the AI is usually correct. Also, a comparison of average baseline performance of 53% and 70% accuracy when the AI is respectively incorrect and correct reveals correlation in image difficulty across human participants and the AI.

Table 5: Regression: Split by Whether AI Correct

	Outcome: Subject Response Correct (Treated Block)	
Skills (Control Block)	(1) AI Incorrect	(2) AI Correct
Treatment	-17.43	13.57
	(1.07)	(0.61)
Accuracy	3.21	4.29
	(0.94)	(0.56)
Accuracy × Treatment	-1.63	-2.63
	(1.30)	(0.82)
Net Confidence	-0.06	0.01
	(0.91)	(0.50)
Net Confidence × Treatment	1.93	-2.52
	(1.34)	(0.75)
Constant	53.14	69.97
	(0.78)	(0.44)
Observations	13,877	44,536

Note: Observations are at the subject-response level. Subject-clustered standard errors are in parentheses. Skills are standardized within session.

A.3 Measurement Error

To address the possibility of measurement error affecting our results, we repeat the preceding OLS analysis of Table 2 with an adaptation of the Obviously Related Instrumental Variables (ORIV) methodology of Gillen, Snowberg and Yariv (2019). Their basic idea is to instrument one imprecise measure with another imprecise measure of the same underlying trait. We adapt this to our experimental setting by randomly drawing a split of the 80 images used for skill estimation into an instrumented and instrumenting set, leading to instrumented and instrumenting estimates of each skill for each individual. To remove the noise introduced by a single split realization, we repeat this splitting procedure 10,000 times. We then repeat our estimation of regression (3) across the 10,000 splits while instrumenting our skill estimates. To correct the standard errors from artificially inflating sample size, we follow Gillen, Snowberg and Yariv (2019) and cluster standard errors at the participant level.

Table 6: Regression: ORIV

	Outcome: Accuracy (Treated Block)			
Skills (Control Block)	(1)	(2)	(3)	(4)
Treatment	6.93	5.45	5.34	5.11
	(0.56)	(0.62)	(0.62)	(0.68)
Accuracy		11.24	11.21	12.17
		(1.65)	(1.80)	(2.37)
Accuracy × Treatment		-3.86	-4.82	-5.02
		(1.98)	(2.11)	(2.73)
Calibration			0.05	0.13
			(0.72)	(0.77)
Calibration × Treatment			1.84	2.02
			(0.84)	(0.89)
IQ				-1.05
				(0.83)
IQ × Treatment				0.01
				(0.94)
Constant	65.46	66.58	66.58	66.74
	(0.48)	(0.53)	(0.54)	(0.60)
Observations	7,320,000	7,320,000	7,320,000	7,320,000

Note: Observations are at the subject-simulation level. Subject-clustered standard errors are in parentheses. Skills are standardized within session and simulated split.

Table 6 repeats the analysis of Table 2 adopting the ORIV estimation approach for ability and calibration; we continue to include IQ as an exogenous control. Consistent with the simple univariate attenuation bias intuition, we find *larger* heterogeneous effects across ability and calibration when accounting for measurement error, which (though less precise) remain significant at conventional levels. For example, in Column (3) including only accuracy and calibration, we find standardized treatment effects across ability and calibration of -4.82 (s.e.=2.11) and 1.84 (s.e.=0.84) respectively, compared to -2.16 (s.e.=0.57) and 1.35 (s.e.=0.54) in the OLS results of Table 2. Interestingly, accounting for measurement error leads to negligible variation in treatment effects across IQ, suggesting that IQ may proxy for imprecisely measured accuracy in the OLS specification. In conclusion, however, accounting for measurement error only strengthens the magnitude of our main results regarding the relationship between calibration and the benefits

of working with AI.

A.4 Alternative Skill Measures

Table 7: Regression: Alternative Skill Measures

	Outcome: Accuracy (Treated Block)			
Skills (Control Block)	(1)	(2)	(3)	(4)
Treatment	8.22	7.34	7.28	7.40
	(0.63)	(0.53)	(0.53)	(0.53)
Accuracy		4.54	4.64	4.36
		(0.40)	(0.61)	(0.62)
Accuracy × Treatment		-1.34	-2.95	-2.71
		(0.52)	(0.73)	(0.73)
Calibration			-0.08	-0.26
			(0.57)	(0.59)
Calibration × Treatment			2.53	2.66
			(0.72)	(0.74)
IQ				1.14
				(0.50)
IQ × Treatment				-0.90
				(0.58)
Constant	70.96	71.62	71.61	71.49
	(0.53)	(0.43)	(0.44)	(0.44)
Observations	732	732	729	729

Note: Observations are at the subject level. Robust standard errors are in parentheses. Skills are standardized within session.

As a final robustness check, we repeat the main analysis of Table 2 using alternative measures of ability and calibration. As an alternative measure of ability, we consider the area under the receiver operator characteristic curve (AUC) (Bamber, 1975; Hanley and McNeil, 1982), a common summary measure of diagnostic ability that extends the Blackwell order (Chan, Gentzkow and Yu, 2022). Intuitively, AUC is interpretable as the probability that a random positive instance has a higher report than a random negative instance. AUC is a useful complement to accuracy because it measures how well predictions are ranked, and not just the performance of an binary classifier induced by the symmetric classification threshold 0.5. On the other hand, changes in accuracy are more

readily interpretable.

For our alternative measure of calibration, we use the coefficient on reported probabilities in a logistic regression of true states S_j on reported probabilities B_{ij} across the images used to estimate skills for each individual i. This corresponds to a one-parameter version of the standard model of Grether (1980) that allows for arbitrary degrees of participant over- and under-confidence. To facilitate interpretation, we take the negative absolute value of the natural log of the aforementioned coefficient as our measure of calibration. The log transformation maintains our earlier interpretation of negative values as under-confident, zero values as calibrated, and positive values as over-confident; further taking the negative absolute value is analogous to our main measure of calibration (2) and allows to interpret higher values with better calibration. Because the logistic function is not defined for reported beliefs of certainty $0, 1$, such responses would drop out of estimation even though they suggest over-confidence; therefore we replace these responses with the nearest interior responses, respectively 0.01 and 0.99. Another complication is that our alternative calibration measure is undefined for participants with a negative coefficient in the logistic regression, corresponding to a negative association between reported probabilities and true states and thus indicative of severe inattention. We drop 3 such participants from evaluation.

Table 7 repeats the analysis of Table 2 adopting the alternative measures for ability and calibration; we continue to include IQ as an exogenous control. While the magnitude of coefficients is not directly comparable to Table 2 given the different outcome (AUC instead of accuracy), we can still compare the relative magnitude of standardized coefficients on the interactions of ability, calibration, and IQ with treatment. Doing so, we find a significant and larger effect of calibration. For example, in the full specification (Column 4 of Table 7), we estimate a standardized treatment effect across calibration of 2.66 (s.e.=0.74), which is nearly equal in magnitude to the analogous ability effect of -2.71 (s.e.=0.73). Interestingly, as in the preceding ORIV specification, statistical significance of the IQ effect (coef=-0.90, s.e.=0.58) is not robust to the alternative specification, although the coefficient is similar in magnitude to the analogous estimate in Table 2

(coef=-1.01, s.e.=0.51). In conclusion, the alternative skill specification supports our main results regarding the importance of calibration.

B Experimental Details

B.1 Overview of Caffe Model

The guesses of our AI assistant are based on the "Convolutional Architecture for Fast Feature Embedding," or Caffe, which is a machine learning model specifically developed for deep learning applications, particularly in image recognition tasks. It allows models to learn statistical patterns directly from large sets of labeled images.

The structure of a Caffe model consists of several key components. Images are initially provided as input data. These images then pass through convolutional layers, which systematically identify and extract visual features such as edges, textures, shapes, and other essential details from the input data. After feature extraction, pooling layers reduce the dimensionality of these features by simplifying and summarizing them, thus helping the model to efficiently handle vast amounts of information. Finally, fully-connected layers aggregate and interpret these condensed features to perform accurate predictions or classifications.

During the training phase, the model adjusts and estimates millions of parameters, known as weights, to minimize prediction errors through iterative learning. This extensive parameter adjustment process enables the model to improve its accuracy continuously. Once trained, the Caffe model can effectively classify new images or recognize visual patterns by leveraging its learned parameters, resulting in high precision and performance in image-based tasks.

B.2 Image Selection

From the images in the desired age ranges, we chose images taken between 2010 and 2014 to ensure they had a high enough resolution and so that participants would be familiar with the outfits of the people in the images. From the remaining images, we removed images that were unclear or contained multiple faces. To further refine the set of images, we conducted two pilots with 100 participants each. For the first pilot, we randomly split 200 images into two subsets, and approximately 50 participants were piloted for each subset. Based on this pilot, we removed images if they were too easy, too difficult, or

too certain. We defined the image as too easy if 95% of participants answered correctly; too difficult if fewer than 30% of participants answered correctly; and too certain if more than 30% of participants reported less than 1 % or above 99%. These exclusion criteria reduced the number of images to 124, so we conducted a second pilot. In the second pilot, we added 196 new images to create a set of 320 images, and increased the restrictiveness of our exclusion criteria. An image was deemed too easy if the average accuracy was above 90% while an image was deemed too difficult if the average accuracy was below 20%. Based on the second round, we selected 227 out of 320 images. Finally, we selected 160 of the 227 suitable images so that AI confidence scores would be well calibrated.

B.3 Image Splits

In each session, we fixed whether an image would appear in a control block or treated block. To increase robustness, we changed the way that images were selected to be in treated or control blocks between two sessions. For the first session, we used the following method to split images. First, all images were ordered lexicographically based on the true label and then the AI prediction of being over 21. Following this order, we alternately assigned each image to treated and control blocks. This produced a set of images containing 40 images of each label in both the treated and control blocks. For the second session, we used the same order based on true label and AI confidence score to form adjacent pairs of images. For each pair, we randomly assigned one image to the control block and the other image to the treated block. We then selected a split that achieved two objectives: (1) having a non-decreasing calibration curve for AI scores in the treated group and (2) having the accuracy and confidence levels of humans in the pilot sessions be well-balanced across treated and control images.

B.4 Screenshots

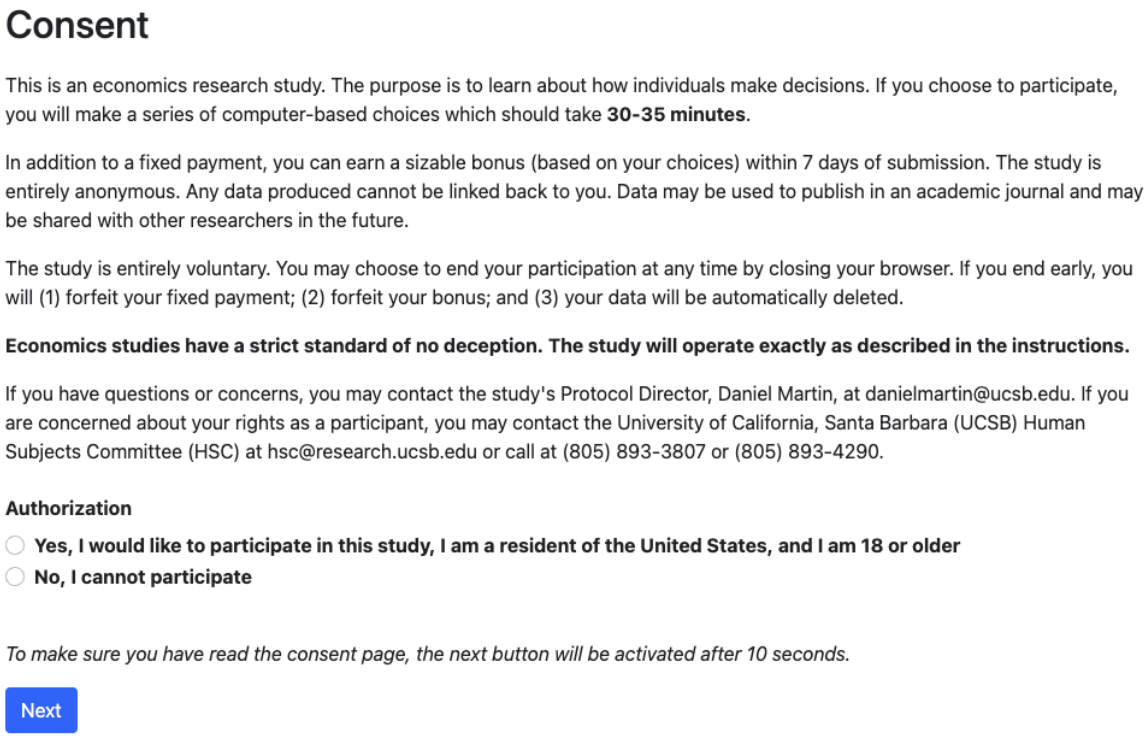


Figure 4: Consent form.

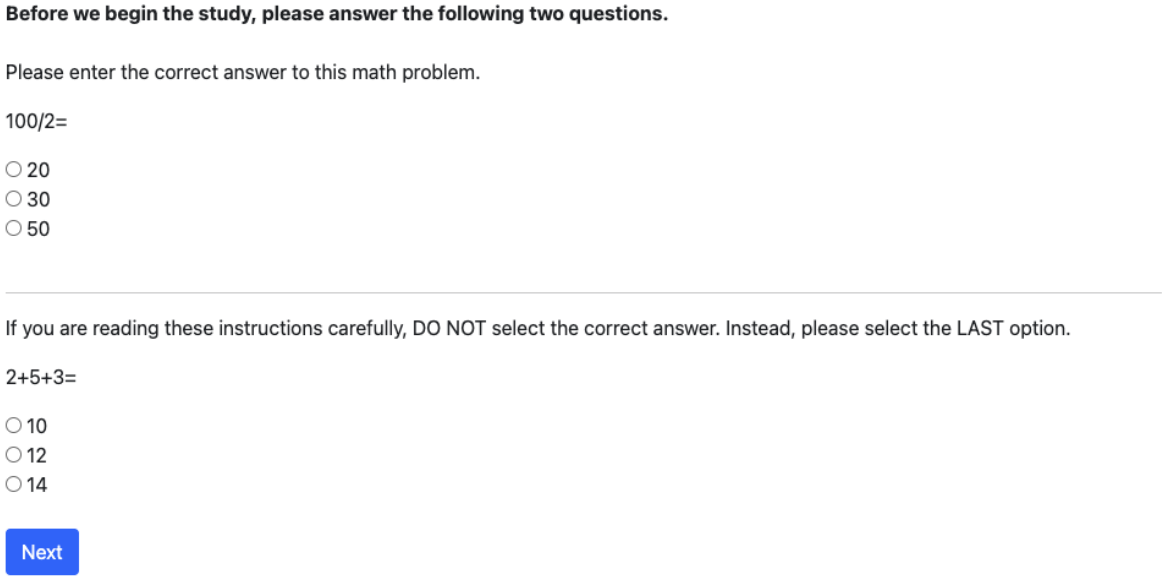


Figure 5: Attention checks.

Welcome

This is a study about classifying images.

The whole session will last for **30 to 35 minutes**. The participation payment you will receive for completing the session is **$5**, and there are two possible additional bonus payments of **$5** and **$1**. Your performance during the session will determine the likelihood you receive the bonus payments.

You will see some instructions over the next few pages, please read them carefully.

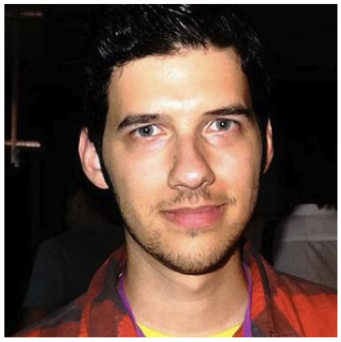
Next

Figure 6: Welcome page.

Instructions

In this session, you will be shown **images** of people one at a time.

For each image, you will be asked the **probability** (from 0% to 100%) that you think the person was over 21 years old when the image was taken. For example, if you were presented with the following image, you would be asked the probability you think the person was over 21 years old.



After seeing 160 images, you will be asked to complete a very brief survey and a short, unrelated additional task.

To make sure you have read the instructions, the next button will be activated after 10 seconds.

Next

Figure 7: Instructions for the experiment.

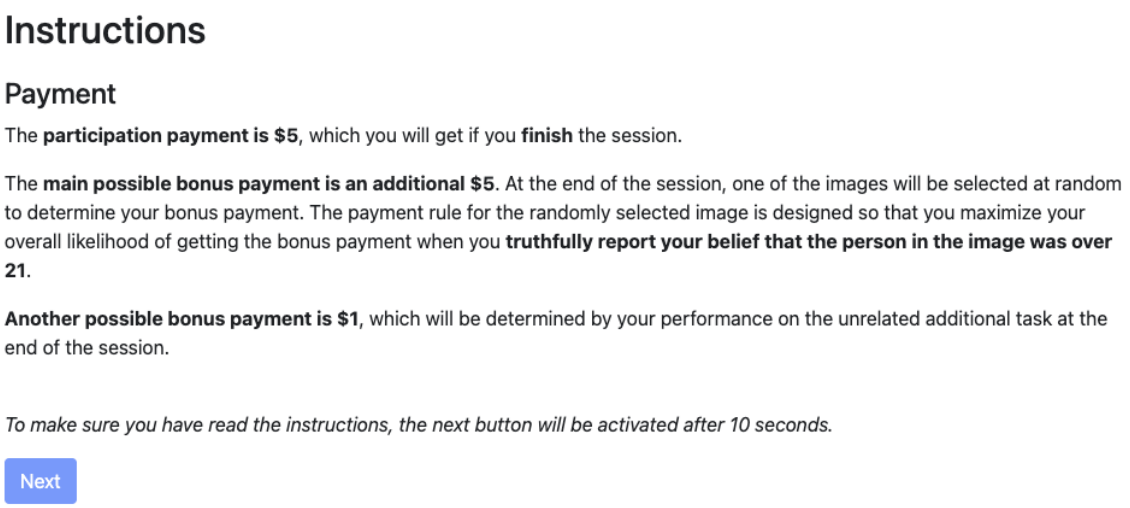


Figure 8: Details about payment.

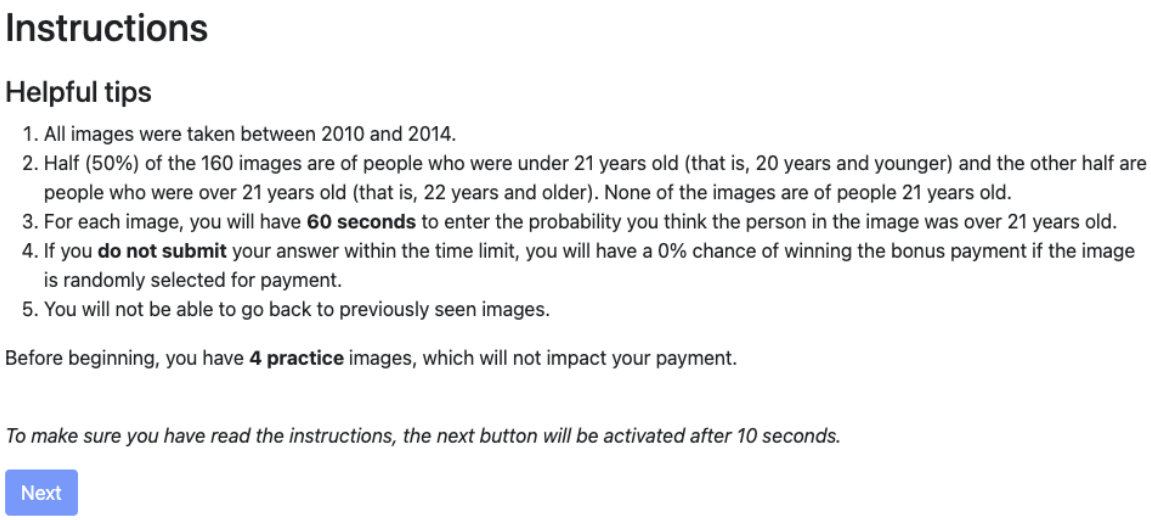


Figure 9: Helpful tips and reminders.

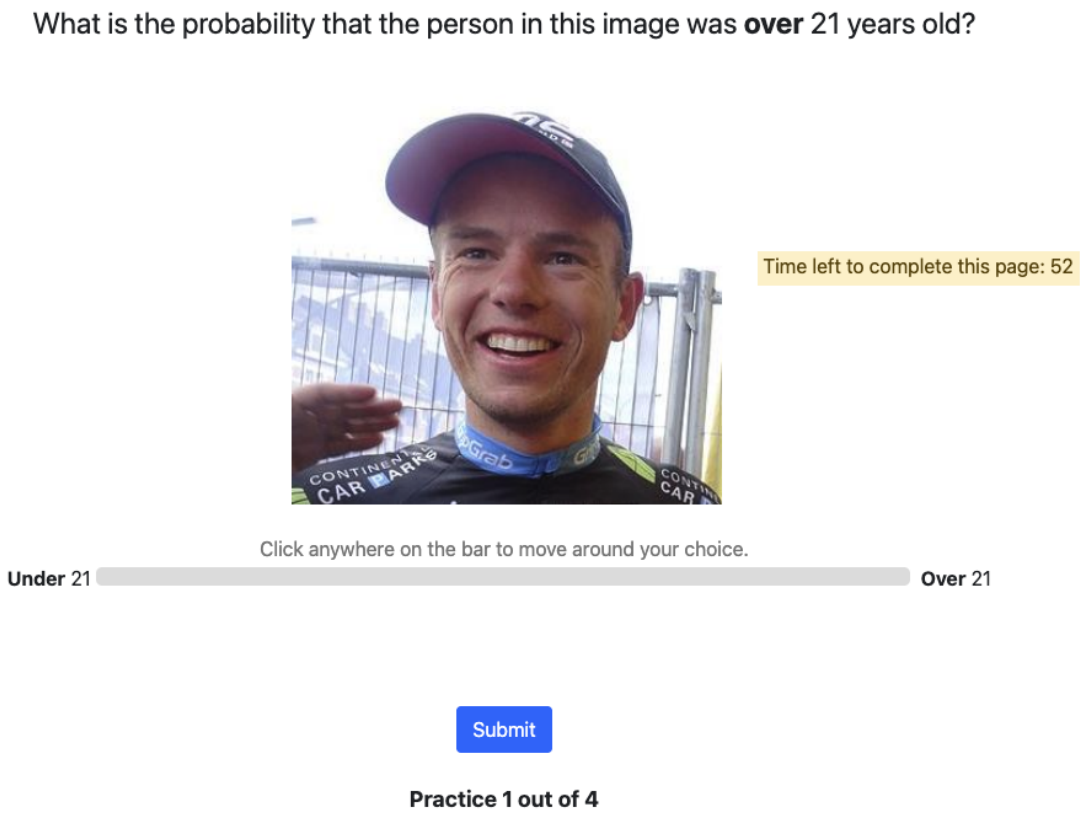


Figure 10: Practice round example (before clicking the slider bar).

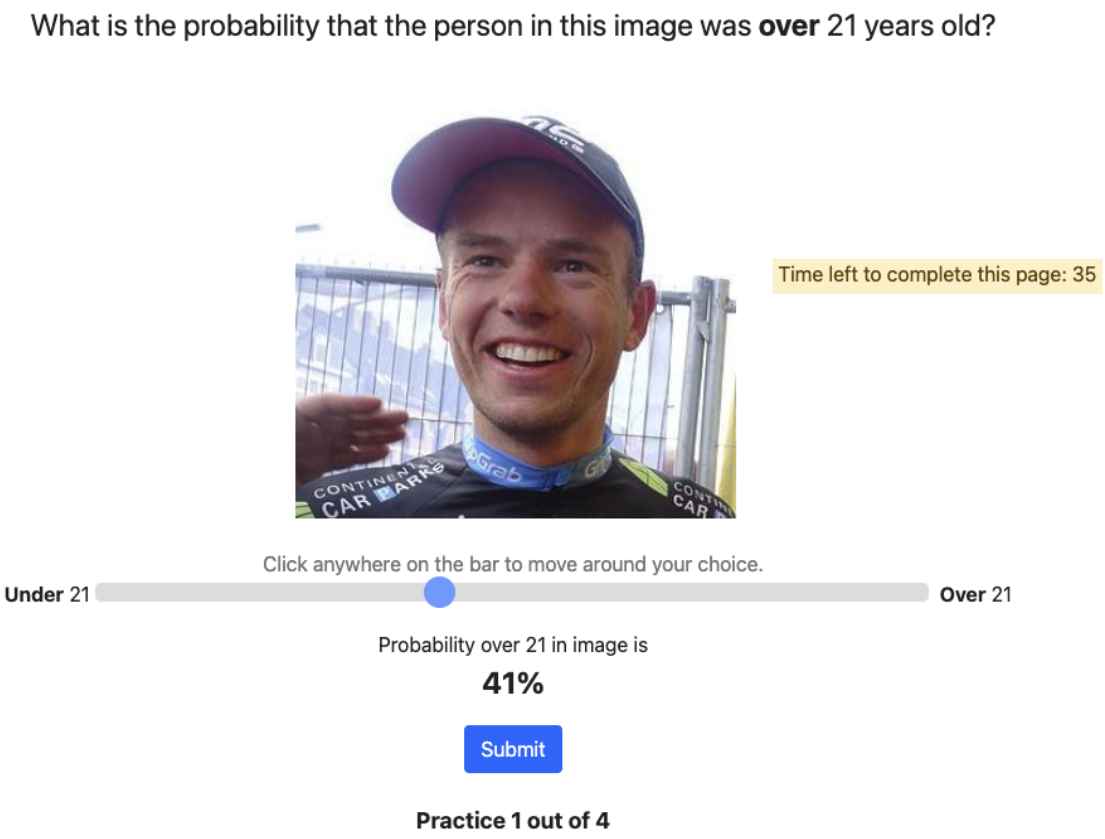


Figure 11: Practice round example (after clicking the slider bar).

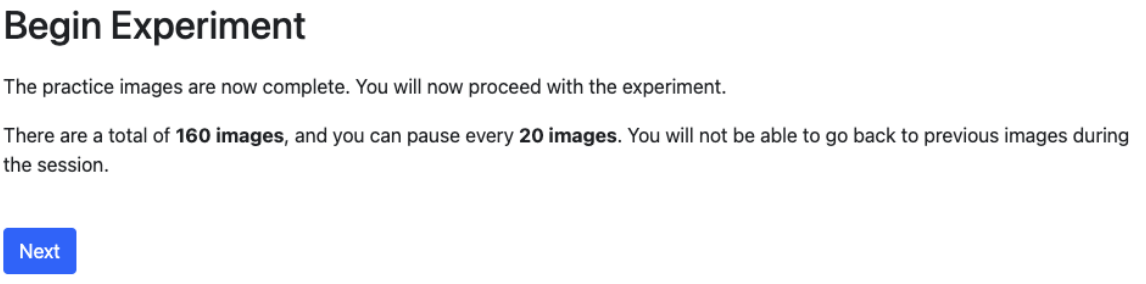


Figure 12: Before incentivized rounds.

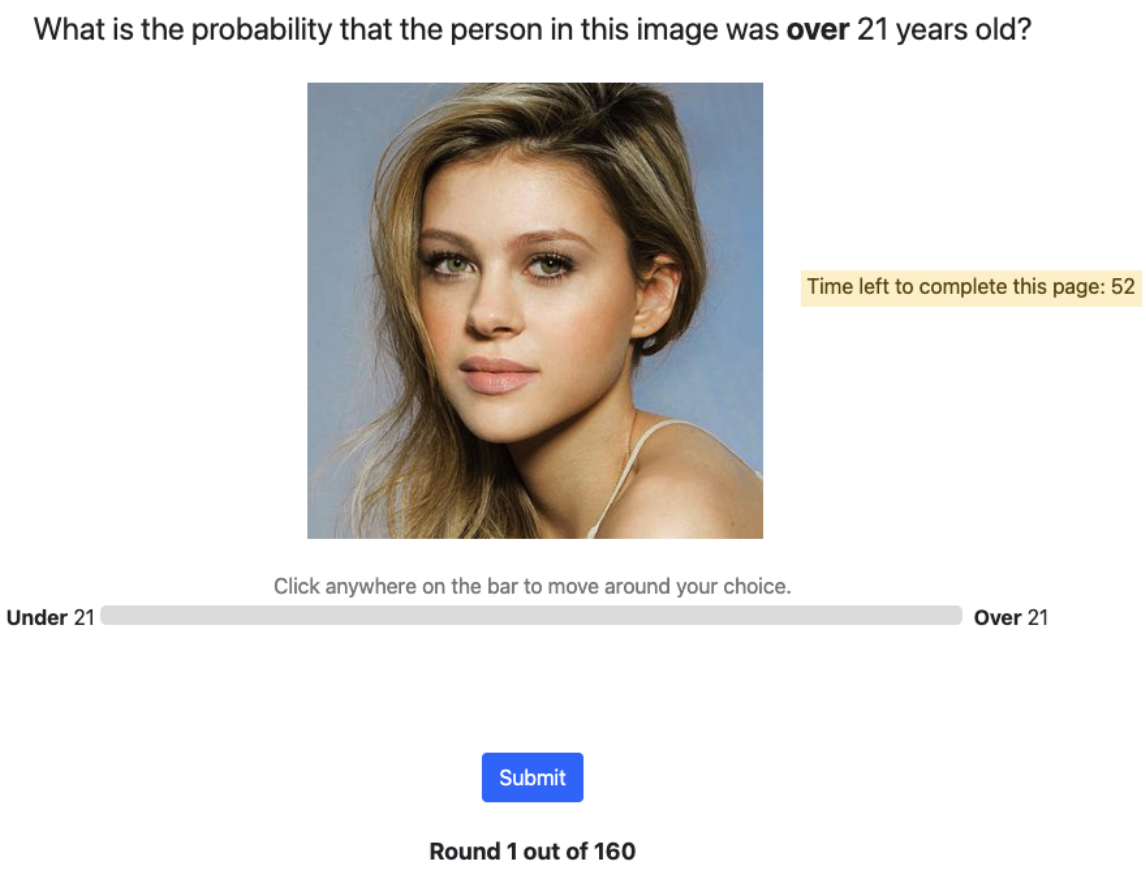


Figure 13: Incentivized round example (before clicking the slider bar).

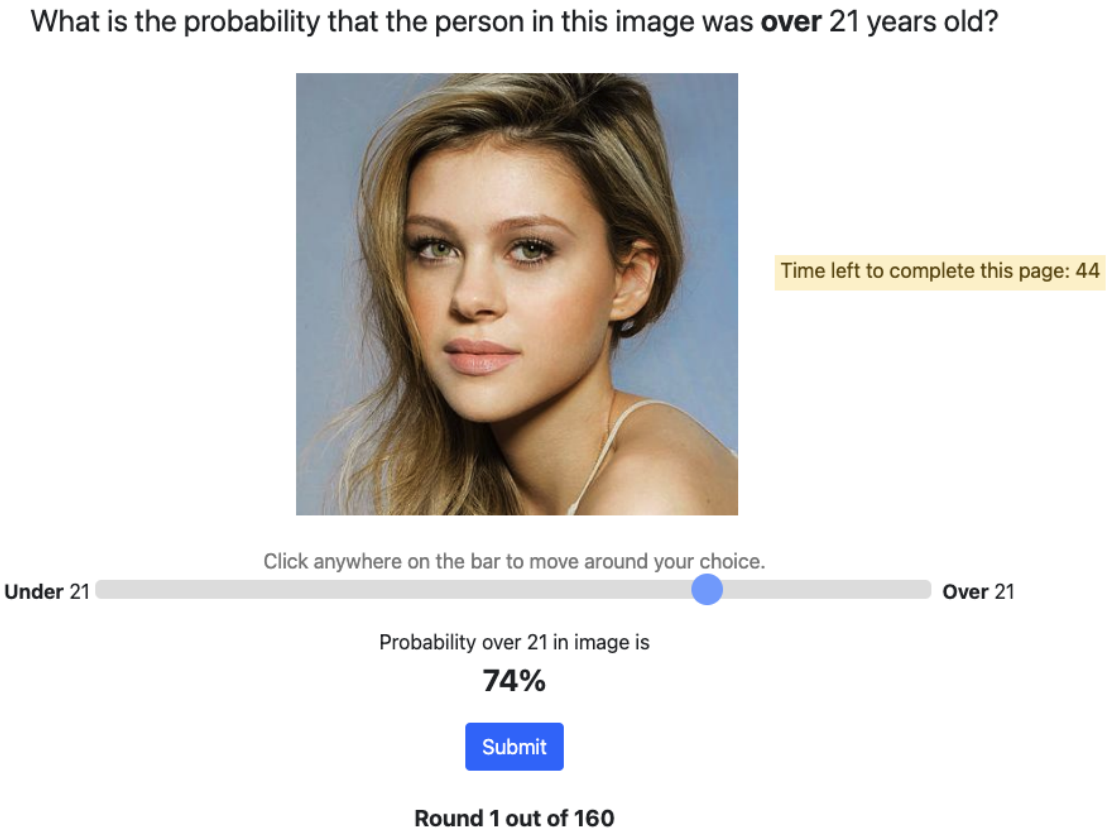


Figure 14: Incentivized round example (after clicking the slider bar).

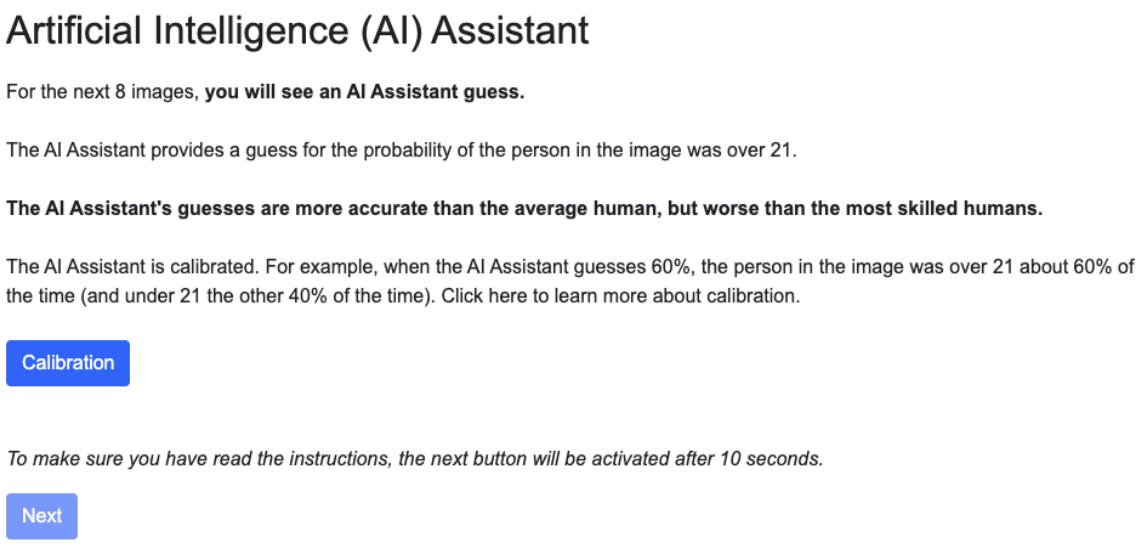


Figure 15: Introduction to AI Assistant.

Calibration

For each image, you will receive a probability from the AI Assistant that was trained using similar images. For example, on a particular image you might be told: "AI assistant's guess is **72%**."

The probability that an image was over 21 for a given probability is given by the figure below. For example, for a score between 50 and 60, approximately 53% of the people in the images were above 21.

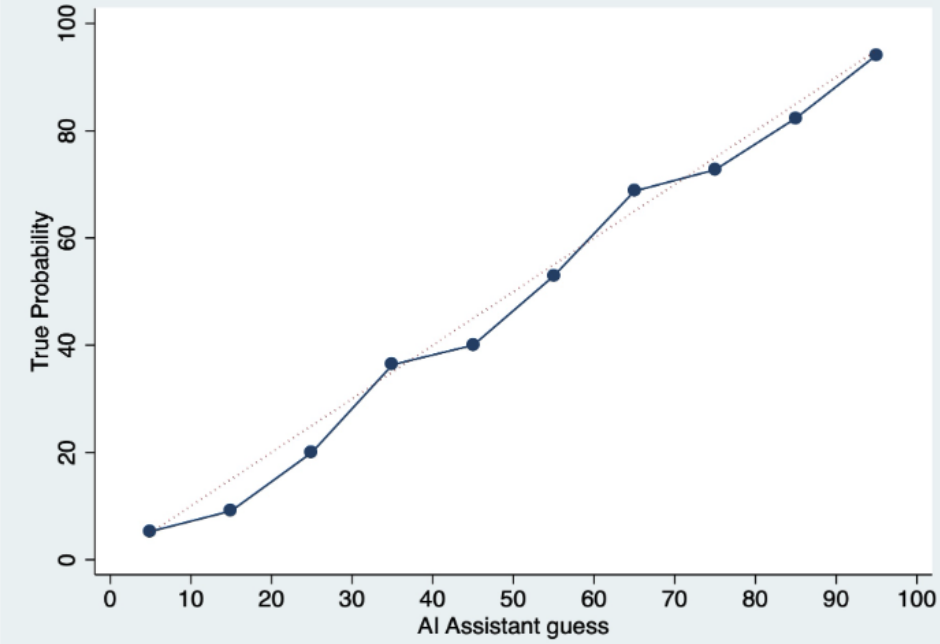


Figure 16: Explanation of calibration.

You have completed 20 out of 160 images

For the next set of 20 images, you will see an **AI Assistant guess.**

Recall that the **AI Assistant's guesses are more accurate than the average human, but worse than the most skilled human.**

Click next to continue.

Next

Figure 17: Page shown to treatment participants before treated blocks begin.

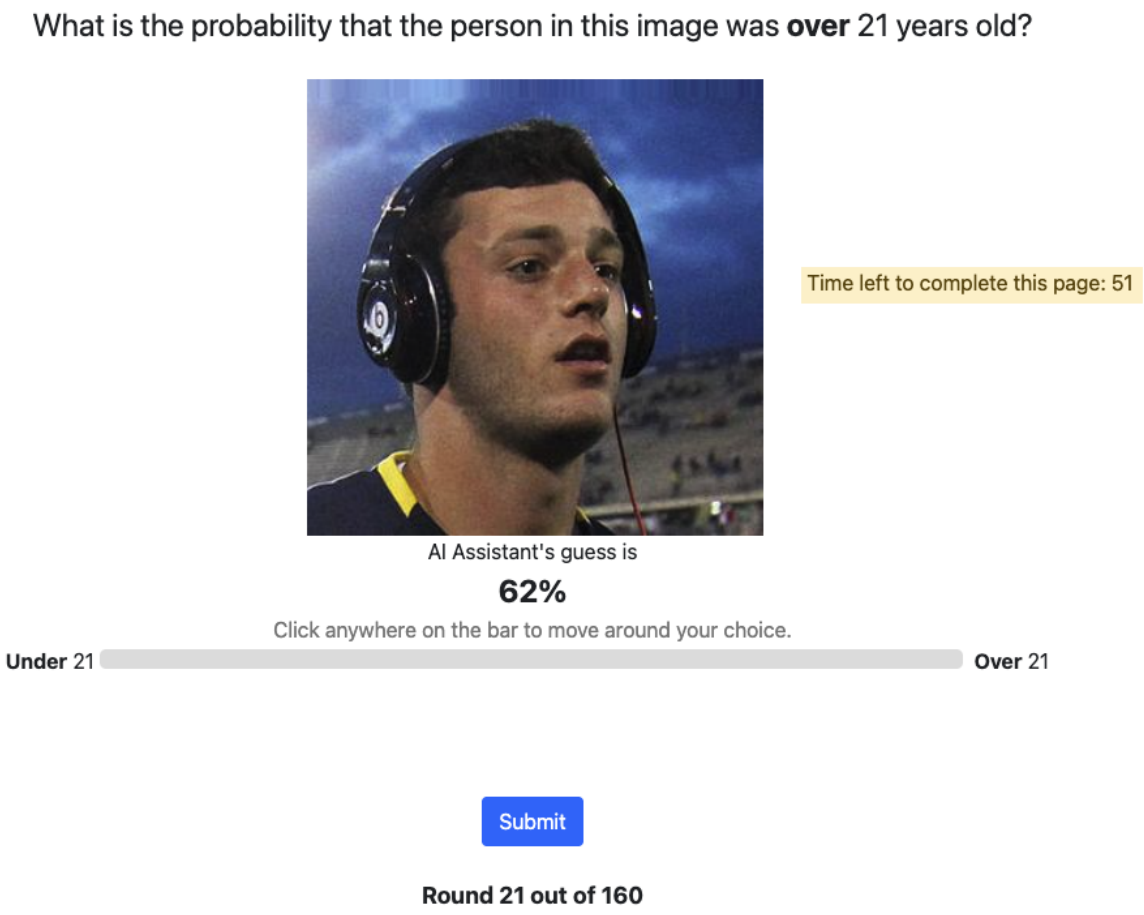


Figure 18: Incentivized round example in treated block for treatment participants (before clicking the slider bar).

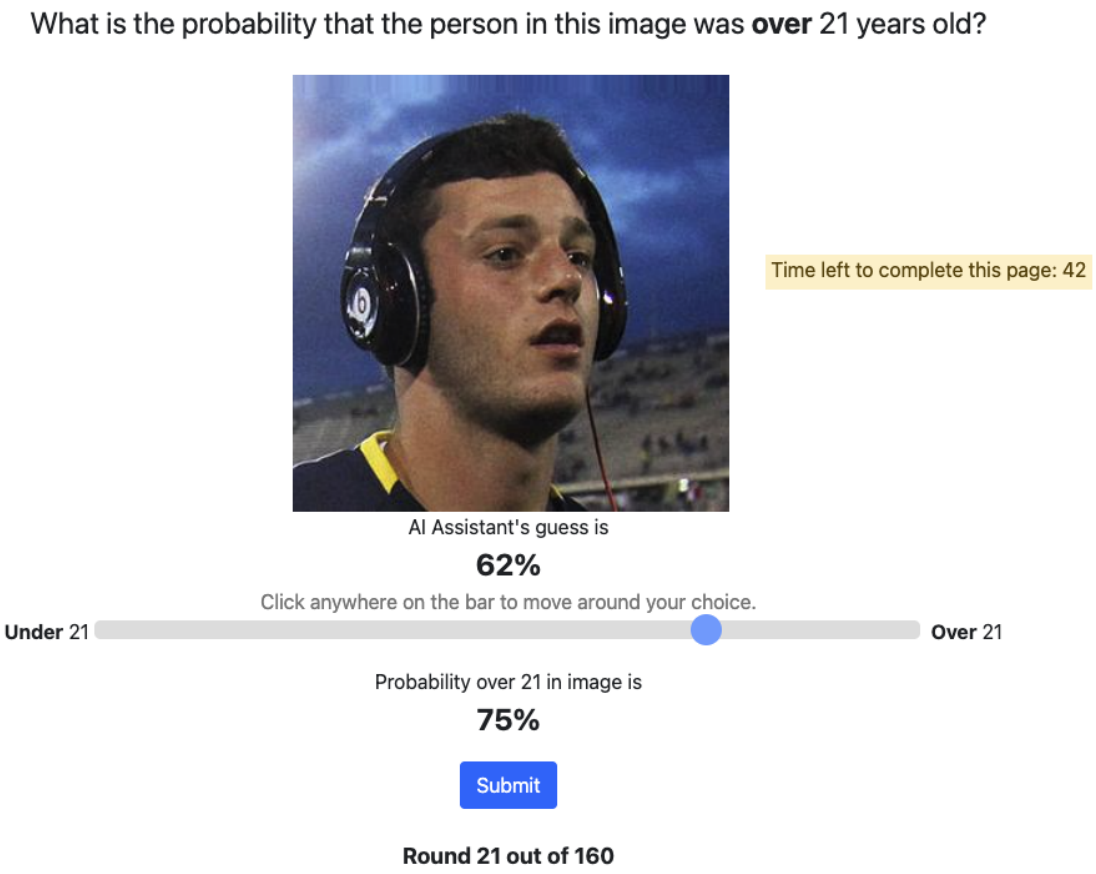


Figure 19: Incentivized round example in treated block for treatment participants (after clicking the slider bar).

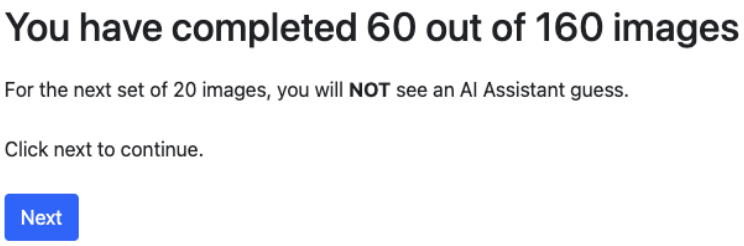


Figure 20: Page shown to treatment participants before control blocks begin.

Figure 21: Break between blocks of 20 images (control treatment).

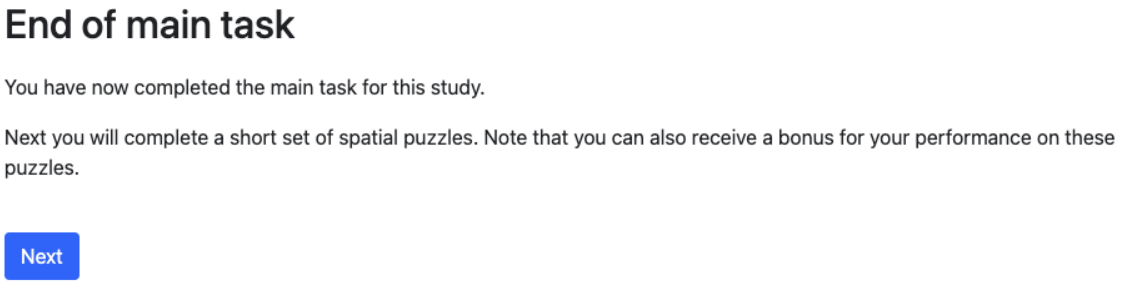


Figure 22: Page shown to all participants after 160 rounds.

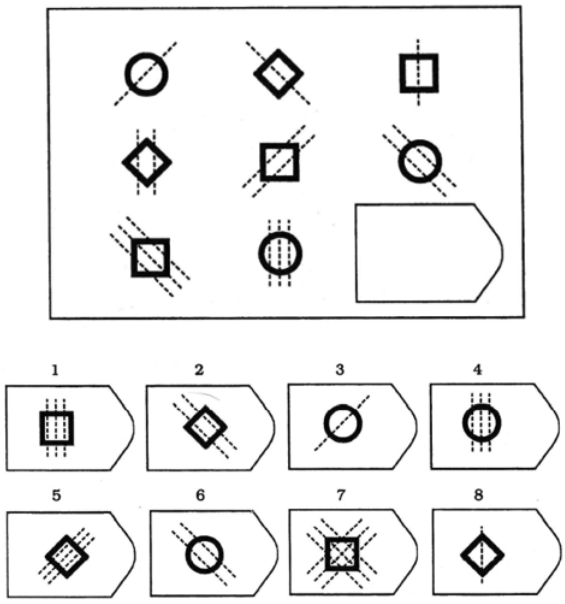
Instructions and example of the puzzles

The last part of the study is a short set of puzzles.

Look at the picture below. There is a missing piece in the bottom right corner. Your job is to figure out which piece fits best.

There are several patterns here. Each row has a square, a circle and a diamond. This is also true of each column. The missing piece must therefore be a diamond.

Also, notice that the shapes on the top row have 1 dotted line through them. Those in the middle row have 2 dotted lines, and in the bottom row it's 3. The missing piece must have 3 dotted lines. Looking at the options, the only diamond with three dotted lines is number 5. This is the missing piece.



Next

Figure 23: Instructions for puzzle task (1/2).

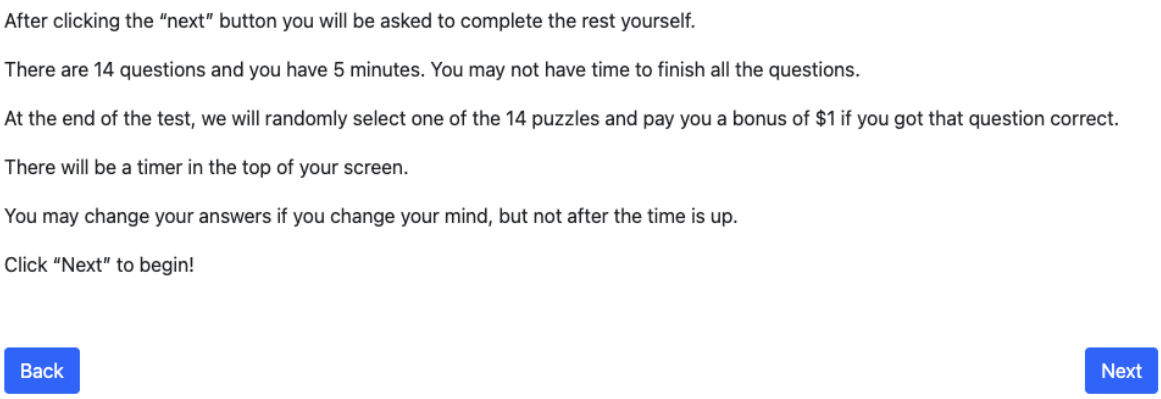


Figure 24: Instructions for puzzle task (2/2).

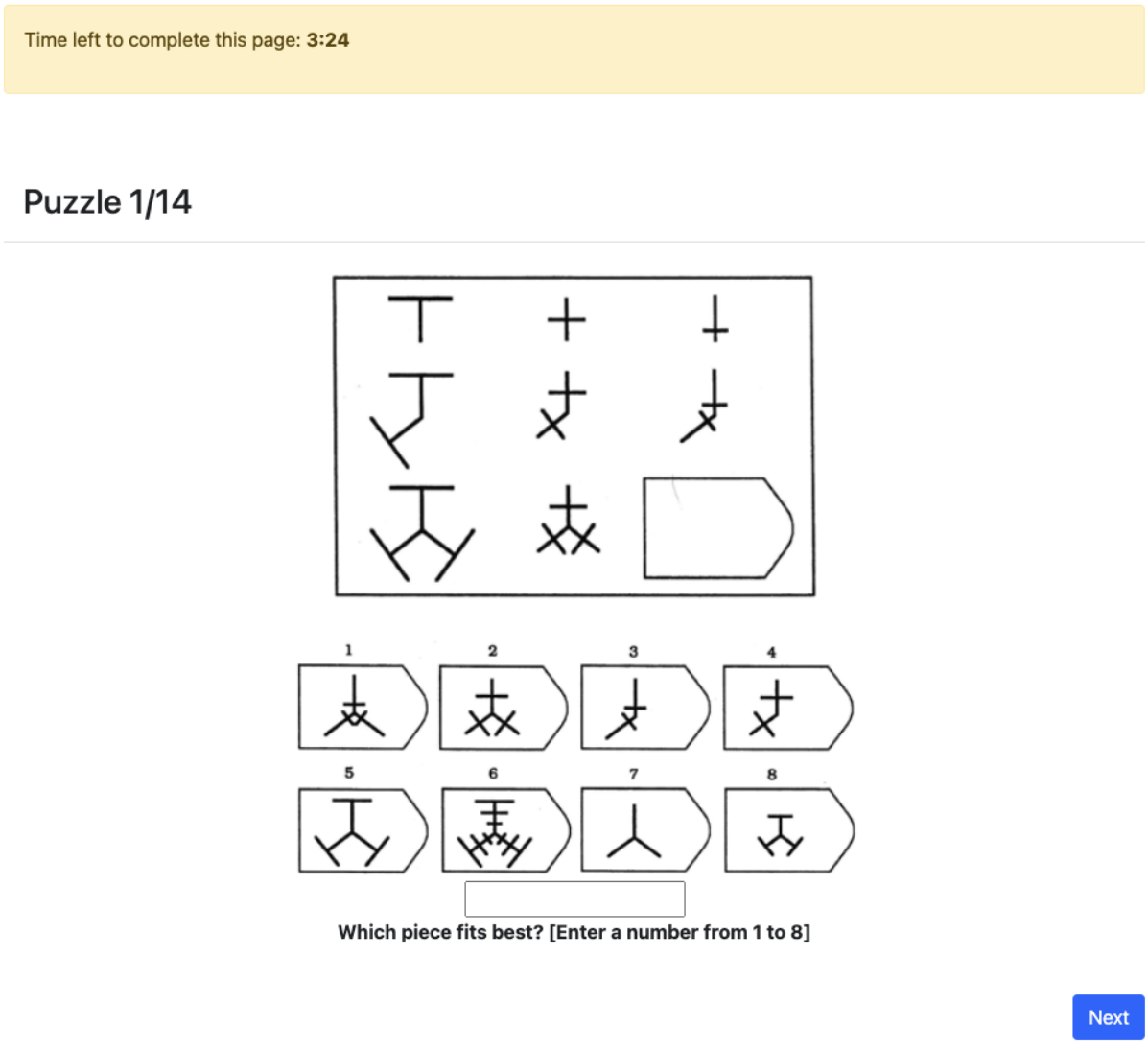


Figure 25: Puzzle task example.

End of Study

The study ends here.

Now you will be asked to complete a very brief survey. Your answers to these questions will not affect your payment in any way. Nevertheless, please answer these questions truthfully since they are very valuable for our research.

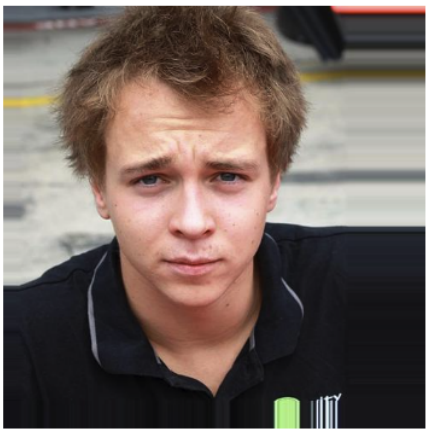
Next

Figure 26: Page shown to all participants after puzzle task.

Main Results

You have reached the end of the session. Thank you for your participation. In this page, one of the images will be randomly chosen from the 160 images, and you will see its results.

Image 157 was randomly selected for payment. Here is the image, and the person was under 21.



As a reminder, for this image you chose that the probability the person was over 21 is 74.0% and the probability the person was under 21 is 26.0%. Therefore, **your probability of getting the bonus payment of $5 given the payment rule we are using is 45.24%**.

If you click the button below, you can see why truthfully telling your belief of the probability leads to maximizing your likelihood of getting the bonus payment.

Payment Rule

Please go to the next page to determine your bonus payment.

Next

Figure 27: Main results page.

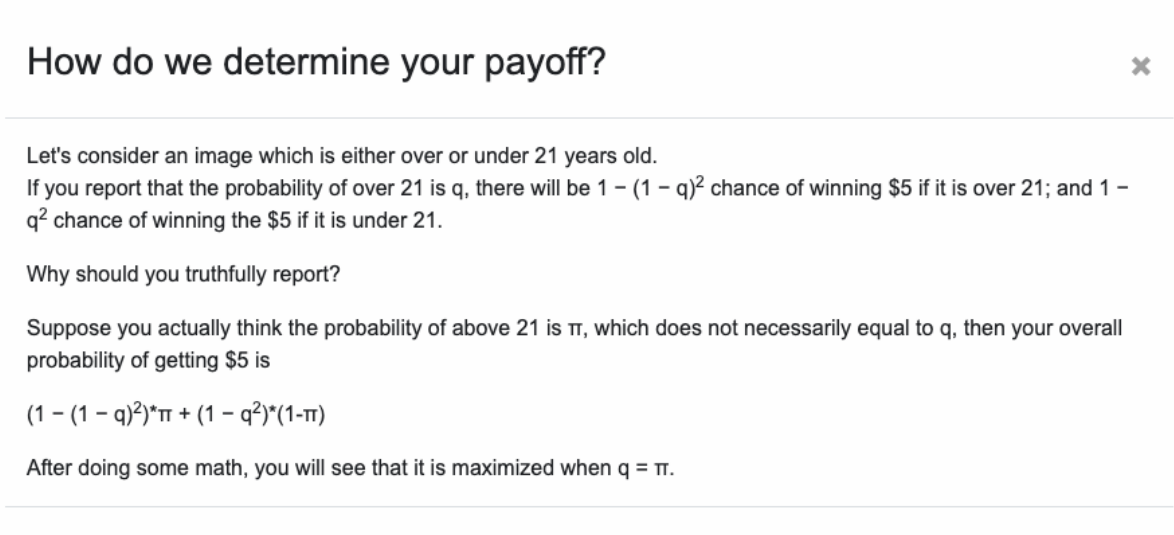


Figure 28: Payment rule description.

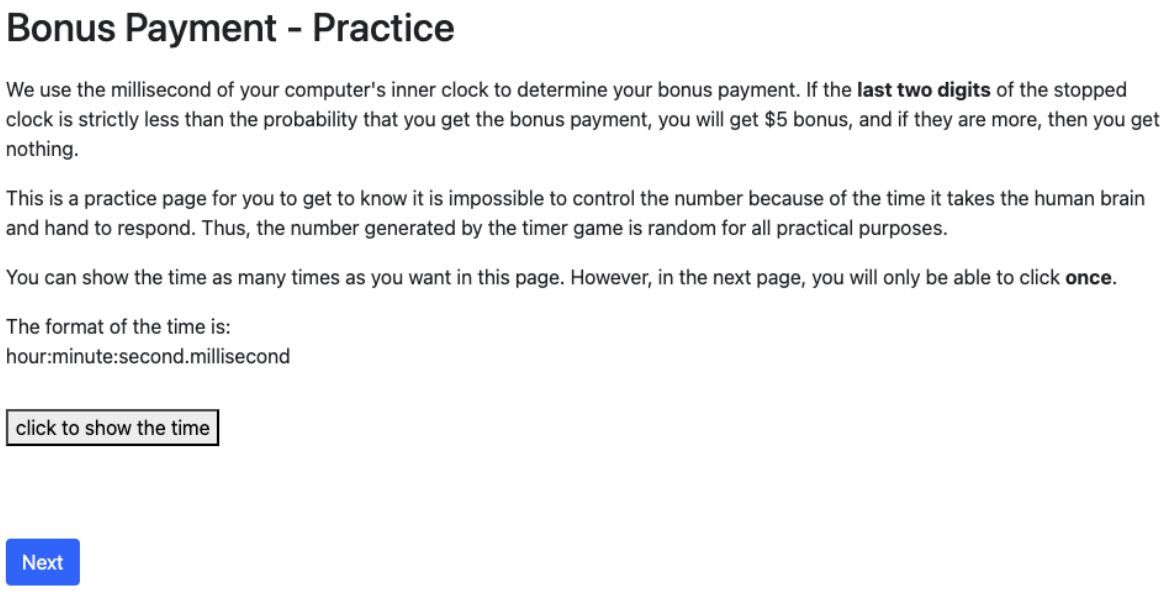


Figure 29: Practice of bonus payment (before clicking the button).

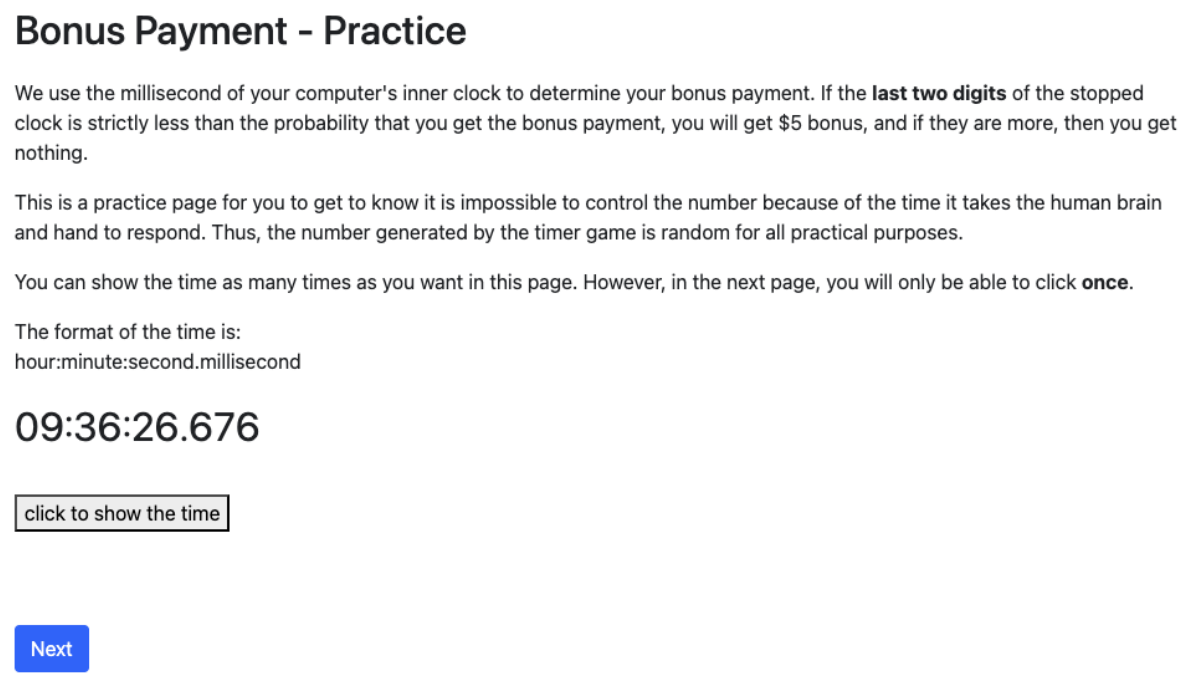


Figure 30: Practice of bonus payment (after clicking the button).

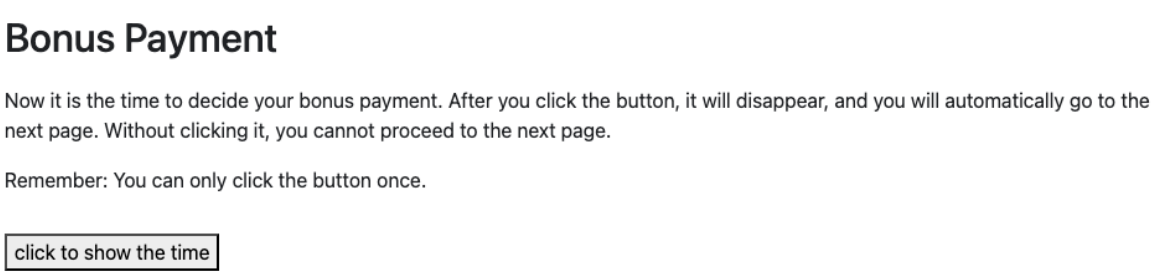


Figure 31: Bonus payment screen.

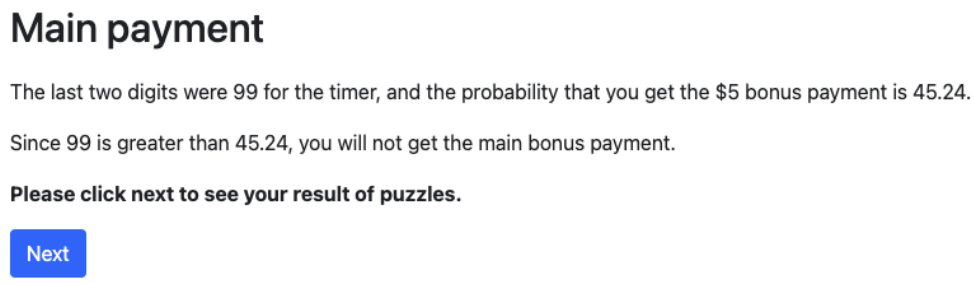
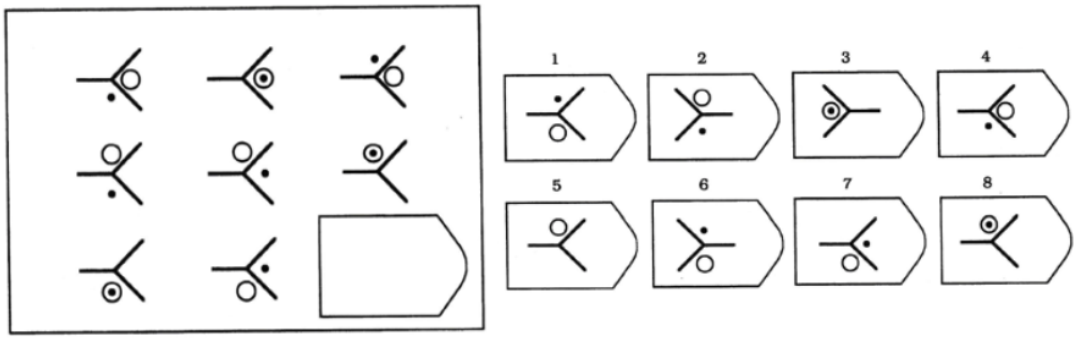


Figure 32: Main payment screen.

Final payment

In this page, one of the puzzles will be randomly chosen from the 14 puzzles, and you will see its results.

Puzzle 7 was randomly selected for payment. Here is the puzzle, and the answer is 1.



As a reminder, for this puzzle you chose None, which is not a correct answer. Therefore, you will get no bonus payment from the puzzle.

In total, you will get $0.00 of bonus payment. It will be distributed through Prolific within 7 days.

Thank you for your participation!

Please click next to finish the experiment and get your participation payment.

Next

Figure 33: Final payment screen.

References (Appendix)

Bamber, Donald. 1975. "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph." *Journal of Mathematical Psychology*, 12(4): 387–415.

Chan, David C, Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostick Skill: Evidence from Radiologists." *Quarterly Journal of Economics*, 137(2): 729–783.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with measurement error: Techniques with applications to the caltech cohort study." *Journal of Political Economy*, 127(4): 1826–1863.

Grether, David M. 1980. "Bayes rule as a descriptive model: The representativeness heuristic." *The Quarterly journal of economics*, 95(3): 537–557.

Hanley, James A, and Barbara J McNeil. 1982. "The meaning and use of the area under a receiver operator characteristic (ROC) curve." *Radiology*, 143(1): 29–36.