Supplementary Information

Frequentist Model Specification

Hypothesis 1a:
*lmer(averageCorrect~reward*delay +(1|Sub), data)*

Hypothesis 1b and 1c:
pairs(emmeans(H1aMod, c('delay_effect','reward_effect')))

Hypothesis 2a:
*lmer(correctedRecognition~reward*delay+(1|Sub),data)*

Hypothesis 2b and 2c:
pairs(emmeans(H2aMod, c('delay_effect','reward_effect')))

Hypothesis 3:
*lmer(zScore~memoryTest*reward+(1|Sub), data), subset = (delay == long))*

Bayesian Model Specification

Hypothesis 1a:
*brm(averageCorrect ~ reward*delay + (1 | subject), data, iter = 4000, save_all_pars =TRUE, seed = 111, sample_prior = TRUE)*

Hypothesis 1b and 1c:
Pairs(emmeans(H1aMod, c('delay_effect','reward_effect')))

Hypothesis 2a:
*brm(correctedRecognition ~ reward*delay + (1 | subject), data, iter = 4000, save_all_pars =TRUE, seed = 111, sample_prior = TRUE)*

Hypothesis 2b and 2c:
Pairs(emmeans(H2aMod, c('delay_effect','reward_effect')))

Hypothesis 3:
*brm(zScore ~ memoryTest*reward + (1 | subject), data, iter = 4000, seed = 111, sample_prior = TRUE)*

ROPE Power Simulation

As outlined in our Sampling Plan, we will assess evidence for the null for all hypotheses that do not reach significance using the Bayesian Region of Practical Equivalence (ROPE)[45] procedure. Evidence for a null effect can be approximated as 95% of the Highest Density Interval (HDI) that falls inside the ROPE (Kruscke, 2018), which encompasses a small band around 0. There are not, however, consistent rules for determining the width of the ROPE (i.e., for determining how close to zero is practically equivalent to zero). Some guidelines recommend defining the ROPE as 0 ± a small fraction (e.g., 0.1) multiplied by the standard deviation of the dependent variable (Kruschke, 2018). We wanted a principled approach to defining a ROPE that would be sensitive to evidence for the null within our high-powered design, so we simulated null findings and determined the proportion of true nulls that would be interpreted as such.

To this end, we simulated 100 datasets by bootstrapping participants from our pilot data. We then simulated null effects of reward in each dataset by randomly permuting reward conditions across trials, separately at each delay (immediate and delayed tests). We repeated this procedure three times to generate null effects in bootstrapped datasets with each of our proposed sample sizes (n = 90, n = 165, n = 240). We then ran the Bayesian models specified above with the permuted reward variable. Next, we determined the percentage of the HDI inside ROPEs ranging from 0 ± 0.10 to 0.22 times the standard deviation of the dependent variable in increments of 0.01. Finally, for each ROPE, we determined how many of the simulations contained 95% of the HDI inside the ROPE, which would constitute evidence in favour of the null hypothesis. Importantly, our sample size requirement was based on having

95% power to detect hypothesized positive effects with frequentist statistics (see Main Text, Power Analyses). However, we additionally wanted to determine our power to detect evidence for our hypothesized null effects (effect of reward on item and associative memory in the immediate test; Hypothesis 1b and 2b). We also wanted to determine whether we had power to detect null evidence for the simple effect of reward on item recognition in the delayed test (Hypothesis 2c), because we were agnostic as to whether we would find a positive or null effect (see Main Text, Table 1). As such, we repeated the simulations described above to determine power to find evidence in favour of the null for each of these hypotheses (H1b, H2b, H2c) at each sample size in our adaptive design (n = 90, n = 165, n = 240). The results of the simulation revealed that 0 ± 0.17 times the standard deviation of the dependent variable would provide 95% power to detect a null effect for all of these hypotheses (H1b, H2b, H2c) at 165 and 240 participants (see **Supplementary Figure 1)**.

We also conducted the same simulation for an interaction between reward and delay on item recognition since we were also agnostic as to whether we would find a positive or null effect for this analysis. We found that this interaction would require unfeasible sample sizes to achieve 95% power for detecting null evidence (**Supplementary Figure 1d**). Therefore, it is important to note that we will only achieve around 50% power to detect null interaction effects (H1a, H2a), but we are well powered to detect null simple effects (H1b, H1c, H2b, H2c) with our maximum sample size.

Pilot Data

To validate our approach, we collected pilot data on the proposed experiment from 36 participants through CloudResearch, the same platform that will be used for final data

collection. First, 18 participants completed an identical task to the one described in the *Methods* section, but with the addition of a short experimental decision-making task designed to test reward sensitivity following the memory tests in the first session. As the reward sensitivity task did not pass our checks for validity, we removed it and ran another pilot experiment with 18 additional participants. There were no significant interactions between the two groups and key variables in any of our analyses (all p's > 0.69), suggesting that removing the reward sensitivity task did not change the pattern of results. As such, we collapsed across these two groups for a total of 36 pilot participants (Gender (self-report in textbox) = 17 identified as female, 19 identified as male, Mean Age = 25.857, Mean Years of Education = 14.629, Geographic location = 100% from United States). Race and ethnicity data were not recorded for the pilot sample due to a technical error. This error will be fixed and our demographics questionnaire expanded (see Supplementary Appendix 1) to ensure detailed demographics information can be reported for the final sample. We then analyzed the pilot data in accordance with our analysis plan described in the *Methods* section.

We first examined the time delay between sessions one and two. Participants had an average delay of 25.62 hours between the first and second sessions (range: 14 hours – 36 hours). We next excluded participants based on performance in the clock game and response rate in the flanker task. No participant had less than a 75% success rate across both reward conditions in the clock game (average clock game success: 92.85%, range: 81.25%-100%) or responded less than 50% of the time on the flanker task (average missed response rate: 0.006%, range: 0%-9.37%). As such, no exclusions were necessary due to task compliance issues. Next, we assessed floor or ceiling effects in memory performance as the proportion of

participants who received perfect or below chance scores in the associative and item memory tests in either the immediate or delayed condition. No participant achieved a perfect score on either the associative memory tests (defined as 100% accuracy) or the item recognition tests (defined as hit rate of 1 and false alarm rate of 0), indicating that there were no ceiling effects (i.e., 0% of participants performed at ceiling). No participant (0%) was at floor in the associative memory test in either the immediate or delayed condition (defined as accuracy < 33%) and only one out of 36 participants (2.78%) was at floor in the item recognition test (defined as corrected recognition < 0). This participant was excluded from analyses, resulting in a final sample of 35 pilot participants for analysis. After exclusions, performance on all memory tests was therefore in an acceptable range to conduct our analyses of interest (Immediate Test: corrected recognition = 0.736, range = 0.364-0.948; associative memory accuracy average = 77.2%, range = 45.83%-95.83%; Delayed Test: corrected recognition = 0.331, range = 0.031-0.687; associative memory accuracy = 52.5%; range = 33.33%-77.08%).

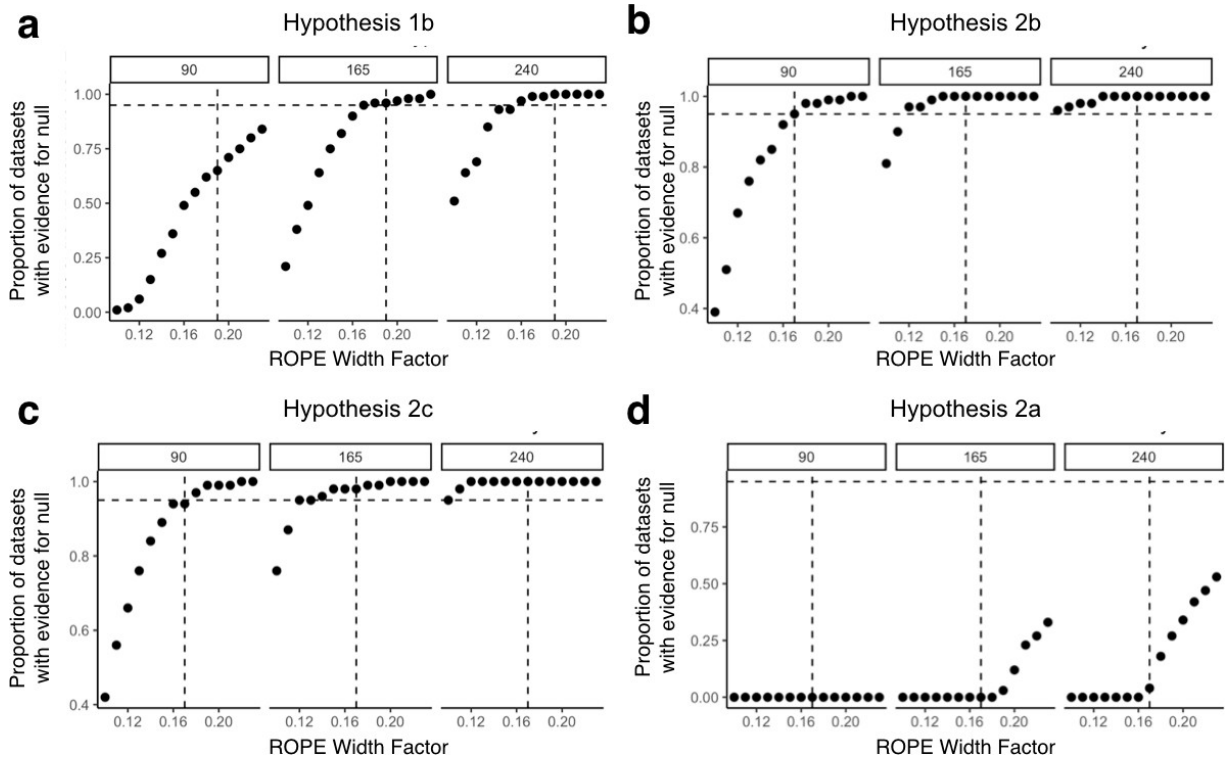
We also determined that our reward manipulation was effective. Clock game success rates were significantly higher on reward trials compared to no reward trials ($t(34) = -3.271$, $p = 0.002$). Participants overall had a 98.21% success rate on the clock game during reward trials, with 87.5% of participants earning the maximum possible reward ($8 in bonus pay) and the remaining participants losing only one reward ($7 in bonus pay). In contrast, participants had an 88.57% success rate on the clock game during no reward trials. Participants were also faster on reward vs no reward trials of the clock game ($t(34) = 2.7667$, $p = 0.009$), with an average response time decrease of 71.7ms on reward trials. Together, this suggests that participants were motivated to earn the reward during the clock game.

We then performed the planned statistical analyses described in *Methods*. Although our pilot data are underpowered, our results are generally consistent with our hypothesized effects. Firstly, there was a trend for a delay dependent effect of reward on associative memory (beta = 0.067, 95% CI =[-0.01, 0.14], p = 0.082, d = 0.47, % in ROPE = 17.10% Hypothesis 1a; see **Supplementary Figure 2a** and **Supplementary Table 1**), such that associative memory was marginally better for reward (vs. no reward) in the delayed vs immediate test. Associative memory did not meaningfully differ between the reward and no reward conditions in the immediate test (beta = -0.007, 95% CI = [-0.08, 0.06], p = 0.994, d = 0.05), but we did not have evidence in favour of the null either (% in ROPE = 78.84%; Hypothesis 1b; see **Supplementary Figure 2a** and **Supplementary Table 1**). Crucially, as hypothesized, reward retroactively enhanced associative memory in the delayed test (beta = 0.075, 95% CI = [0.03, 0.12], p = 0.002, d = 0.57; Hypothesis 1c; see **Supplementary Figure 2a** and **Supplementary Table 1**).

There was no such delay-dependent effect of reward on item memory (beta = 0.02, 95% CI = [-0.04, 0.09], p = 0.525, d = 0.21; Hypothesis 2a; see **Supplementary Figure 2b** and **Supplementary Table 1**), although we were not able to accept the null either (% in ROPE = 74.78%). Item memory in the immediate test did not reliably differ between the reward and no reward conditions (beta = 0.01, 95% CI = [-0.05, 0.07], p = 0.96, d = 0.12), although we did not detect evidence in favour of the null (% in ROPE: 92.82 %; Hypothesis 2b; see **Supplementary Figure 2b** and **Supplementary Table 1**). In contrast to associative memory, item memory was not reliably influenced by reward in the delayed test (beta = -0.01, 95% CI = [-0.07, 0.05], p = 0.97, d = 0.096), although we again did not detect evidence in favour of the null (% in ROPE: 93.69 %; Hypothesis 2c; see **Supplementary Figure 2c** and **Supplementary Table 1**). Finally,

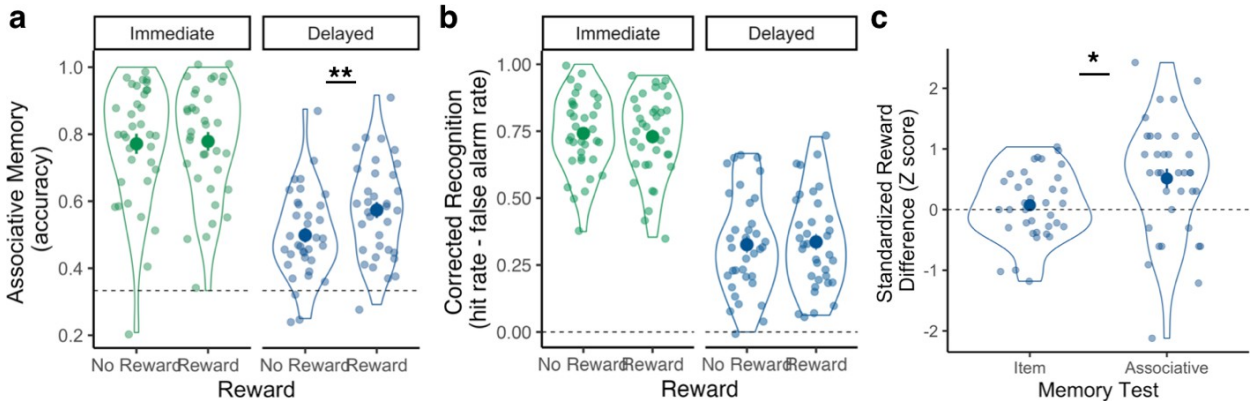
there was a significant interaction between memory test and reward on normalized memory scores in the delayed test (beta = 0.34, 95% CI= [0.06, 0.63], p = 0.019, d = 0.62; Hypothesis 3; see **Supplementary Figure 2c** and **Supplementary Table 1**).

Together, these results suggest that the effect of retroactive reward was greater for associative memory than item recognition in the delayed test. These pilot data provide preliminary evidence to suggest that the retroactive reward effect is more robust for associative memory, compared to item recognition, when tested at a delay, providing preliminary evidence in favour of each of our *a priori* hypotheses regarding associative memory. We did not find preliminary evidence for an effect of reward on item recognition, but were also not able to detect evidence for the null with our relatively small pilot sample size.



Supplementary Figure 1 | Power simulation for a priori null hypotheses

Proportion of simulations in which 95% of the HDI was inside the ROPE for each proposed sample size in our adaptive design (plotted in separate panels) and for a variety of ROPE widths (defining the x-axis). Horizontal dashed lines indicate 95% power and vertical dashed line indicates our selected ROPE width of 0 ± 0.17 times the standard deviation of the dependent variable. We conducted these simulations for **(a)** the effect of reward on associative memory in the immediate test (Hypothesis 1b), **(b)** the effect of reward on item memory in the immediate test (Hypothesis 2b) and **(c)** the effect of reward on item memory in the delayed test (Hypothesis 2c). In **(d)**, we conducted a power simulation for the interaction between delay and reward on item memory and found we will only achieve approximately 50% power to detect a null interaction.



Supplementary Figure 2 | Pilot data results

(a) Associative memory as a function of reward and delay. Associative memory was not enhanced in the reward compared to no reward condition in the immediate test (left, Hypothesis 1b), but was enhanced in the reward compared to no reward condition in the delayed test (right, Hypothesis 1c). **(b)** Item memory as a function of reward and delay. Corrected recognition was not enhanced in the reward compared to no reward condition in either the immediate test (left, Hypothesis 2b) or the delayed test (right, Hypothesis 2c). **(c)** Comparison of item and associative memory as a function of reward in the delayed test. The effect of reward (vs no reward) on associative memory was greater than the effect of reward (vs no reward) on item memory (Hypothesis 3). In all panels, violins indicate the distribution of values across participants. Small translucent points indicate participant averages per condition. Large solid points indicate the mean across participants, with error bars reflecting the standard error of the mean.

Supplementary Table 1 | Statistical analysis of pilot data

Hypothesis	Cohen's D	beta	95% CI	p value	% in ROPE
Hypothesis 1a Is there a delay-dependent effect of reward on associative memory?	0.47	0.067	[-0.01, 0.14]	0.082~	17.10%
Hypothesis 1b Is there a retroactive effect of reward on associative memory in the immediate test?	0.05	-0.007	[-0.08, 0.06]	0.994	78.74%
Hypothesis 1c Is there an effect of reward on associative memory in the delayed test?	0.57	-0.075	[-0.14, -0.004]	0.035*	NA
Hypothesis 2a Is there a delay-dependent effect of reward on item recognition?	0.214	0.02	[-0.04, 0.09]	0.525	74.78%
Hypothesis 2b Is there a retroactive effect of reward on item recognition in the item recognition test in the immediate test?	0.12	0.01	[-0.05, 0.07]	0.96	92.82 %
Hypothesis 2c Is there a retroactive effect of reward on item recognition in the delayed test?	0.096	-0.01	[-0.07, 0.05]	0.97	93.69 %
Hypothesis 3 Is the effect of reward greater for associative than item memory in the delayed test?	0.62	0.34	[0.06, 0.63]	0.019*	NA

* $p<0.05$, ~ $p<0.10$

Supplementary Results

In an unplanned exploratory analysis, we investigated whether retroactive reward effects on memory were related to participants' socioeconomic status (SES). We quantified SES in three ways: 1) a participant's monthly income divided by their monthly expenses, 2) a participant's perceived standing relative to others in their community, and 3) a participant's perceived standing relative to others in their country (measured through self-report, see Supplementary Appendix 1 for Demographics Questionnaire). We then calculated the Spearman's correlation between 1) each measure of SES and 2) their difference in associative memory performance between the rewarded and non-rewarded conditions at the long delay. There was no relationship between associative memory modulation and monthly income (rho = -0.041, p = 0.531), community standing (rho = 0.027, p = 0.675), or country standing (rho = 0.093, p = 0.153).

Supplementary Table 2 | Bayesian Model Outputs for Confirmatory Analyses

Hypothesis	beta	95% CI
Hypothesis 1a Is there a delay-dependent effect of reward on associative memory?	0.02	[-0.01, 0.05]
Hypothesis 1b Is there a retroactive effect of reward on associative memory in the immediate test?	0.005	[-0.02, 0.03]
Hypothesis 1c Is there an effect of reward on associative memory in the delayed test?	-0.013	[-0.03, 0.009]
Hypothesis 2a Is there a delay-dependent effect of reward on item recognition?	0.01	[-0.02, 0.04]
Hypothesis 2b Is there a retroactive effect of reward on item recognition in the item recognition test in the immediate test?	0.01	[-0.009, 0.029]
Hypothesis 2c Is there a retroactive effect of reward on item recognition in the delayed test?	-0.002	[-0.02, 0.02]
Hypothesis 3 Is the effect of reward greater for associative than item memory in the delayed test?	0.06	[-0.05, 0.17]

Supplementary Table 3 | Confirmatory Analyses in High Performing Participants

Hypothesis	beta	95% CI	p value	% in ROPE
Hypothesis 1a Is there a delay-dependent effect of reward on associative memory?	0.025	[-0.01, 0.06]	0.188	57.73%
Hypothesis 1b Is there a retroactive effect of reward on associative memory in the immediate test?	0.006	[-0.03, 0.04]	0.966	97.84%
Hypothesis 1c Is there an effect of reward on associative memory in the delayed test?	-0.018	[-0.05, 0.02]	0.50	77.46%
Hypothesis 2a Is there a delay-dependent effect of reward on item recognition?	0.021	[-0.01, 0.05]	0.177	89.66%
Hypothesis 2b Is there a retroactive effect of reward on item recognition in the item recognition test in the immediate test?	0.019	[-0.01, 0.05]	0.338	100%
Hypothesis 2c Is there a retroactive effect of reward on item recognition in the delayed test?	-0.003	[-0.03, 0.03]	0.995	99.43%
Hypothesis 3 Is the effect of reward greater for associative than item memory in the delayed test?	0.099	[-0.09, 0.29]	0.311	83.03%

Supplementary Table 4 | Confirmatory analyses in participants who were not reward sensitive

Hypothesis	beta	95% CI	p value	% in ROPE
Hypothesis 1a Is there a delay-dependent effect of reward on associative memory?	-0.033	[-0.0782, 0.012]	0.147	51.10%
Hypothesis 1b Is there a retroactive effect of reward on associative memory in the immediate test?	-0.014	[-0.056, 0.028]	0.827	91.50%
Hypothesis 1c Is there an effect of reward on associative memory in the delayed test?	0.019	[-0.022, 0.061]	0.627	83.04%
Hypothesis 2a Is there a delay-dependent effect of reward	0.020	[-0.019, 0.058]	0.316	90.15%

on item recognition?				
Hypothesis 2b Is there a retroactive effect of reward on item recognition in the item recognition test in the immediate test?	0.013	[-0.023, 0.048]	0.789	90.99%
Hypothesis 2c Is there a retroactive effect of reward on item recognition in the delayed test?	-0.007	[-0.042, 0.029]	0.961	82.24%
Hypothesis 3 Is the effect of reward greater for associative than item memory in the delayed test?	-0.124	[-0.315, 0.066]	0.202	54.98%

Supplementary Appendix 1: Demographics Questionnaire

Participants will be asked to report demographic information using the following questionnaire:

1. Sex (assigned at birth) (textbox)

2. Gender (The social and cultural construct that refers to the behaviors, roles, expectations, and norms associated with being male or female in a particular society. E.g., woman, non-binary, etc.) (textbox)

3. Age (textbox)

4. How many years of education do you have? Start counting at Grade 1. (textbox)

5. What is your handedness? (radio buttons with options: left handed, right handed, ambidextrous)

6. Ethnicity (Ethnicity refers to a person's cultural background and identity, which may include shared language, customs, religion, and traditions; e.g., Latinx, Caribbean, Indian) (textbox)

7. Race (Race refers to a person's physical characteristics, such as skin colour, hair texture, and facial features; e.g., White, Black, Asian) (textbox)

8. What country were you born in? (textbox)

9. What country do you currently reside in? (textbox)

10. How well do you use English? [Very well, Well, Fair, Poorly, Very poorly] (multiple choice, set threshold to include well & very well)

11. Please list all the languages you use fluently or near fluently. (textbox)

12. Please list all the languages you consider to be your primary language. (textbox)

13. Eyesight: Do you have normal or corrected vision? For example if you can see normally either with or without glasses or contacts, select yes (radio buttons with options yes or no)

14. Are you color blind? (radio buttons with options yes/no)

15. Have you ever been diagnosed with: "Psychiatric Disorder (e.g. depression, anxiety)", "ADD, ADHD or learning disability", "Neurological disorder (e.g. epilepsy brain damage)", "I have not been diagnosed with any of these" (checkboxes)

16. Please list any medications you take. (textbox)

17. What was your total household income for the past 12 months? This should include your total income across all earners, before taxes and other deductions, from all sources (for example, wages, rent from properties, government benefits, help from relatives including child payments and alimony, interest, retirement, and so on). Please input your household income in your local currency (specified in the next question). (textbox)

18. Indicate the currency in which you have reported your total household income in the previous question. (text box)

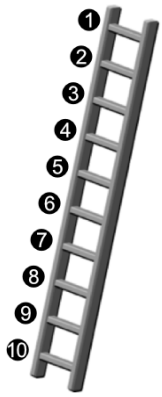
19. What is your postal code? (textbox)

20. If your country does not use postal codes, please name where you currently reside (e.g., by providing the name of your city/town, state/province/prefecture/region/territory, and country). (textbox)

21. Approximately how much does your household spend on essentials each month (e.g., rent, groceries, transportation, childcare, etc.)? Please specify a number in your local currency. (textbox)

22. Think of this ladder as representing where people stand in their communities. People define community in different ways; please define it in whatever way is most meaningful to you. At the top of the ladder are the people who have the highest standing in their community. At the bottom are the people who have the lowest standing in their community. Where would you place yourself on this ladder? Select the number corresponding to the rung where you think you stand at this time in your life, relative to other people in your community. (checkboxes with numbers 1 to 10)

23. Think of this ladder as representing where people stand in the country that you live in. At the top of the ladder are the people who are the best off - those who have the most money, the most education, and the most respected jobs. At the bottom are the people who are the worst off - who have the least money, least education, and the least respected jobs or no job. The higher up you are on this ladder, the closer you are to the people at the very top; the lower you are, the closer you are to the people at the very bottom. Where would you place yourself on this ladder? Select the number corresponding to the rung where you think you stand at this time in your life, relative to other people in your country. (checkboxes with numbers 1 to 10)



References

Kruschke, J. K. Rejecting or Accepting Parameter Values in Bayesian Estimation. *Adv. Methods Pract. Psychol. Sci.* **1**, 270–280 (2018).