The effect of cognitive load on economic decision-making: a replication attempt*

Sheryl Ball[†] Benjamin Katz[‡] Flora Li[§] Alec Smith[¶]

December 26, 2022

Abstract

When cognitive load exceeds cognitive capacity, individuals may make poorer decisions, especially when substantial deliberation is required. Deck and Jahedi's (2015) influential work on cognitive load found that individuals whose arithmetic performance is most impacted by high cognitive load become more risk averse, less patient and more subject to the anchoring effect. Since results of cognitive load manipulation studies are mixed, replication of influential studies is essential to strengthen our understanding of the effects of cognitive load. In this paper, we attempt to closely replicate Experiment 1 in Deck and Jahedi (2015). Though we observe similar effects of cognitive load on arithmetic performance, we fail to replicate their overall results on risky choice and impatience. While we are unable to clearly identify the reasons for this non-replication, the evidence points to subtle differences in the allocation of attention and effort across subject pools.

Keywords: replication, cognitive load, risky choice, patience

*Acknowledgements: We thank Cary Deck and Salar Jahedi for giving us access to their experiment code, data, and analysis scripts. Funding was provided by the Virginia Tech Department of Economics.

[†]Department of Economics, Virginia Tech. School of Neuroscience, Virginia Tech. sball@vt.edu.

[‡]Department of Human Development and Family Science, Virginia Tech. katzben@vt.edu

[§]Corresponding author. Economics Experimental Lab, Nanjing Audit University. florali@nau.edu.cn

[¶]Department of Economics, Virginia Tech. School of Neuroscience, Virginia Tech. alecsmith@vt.edu.

1 Introduction

People have limited capacity to process information when making economic decisions. When these capacity constraints bind, decisions may be noisy, biased, or both. Dual systems theory (e.g. Kahneman, 2011) suggests that when information processing is capacity-constrained, decisions may be influenced by rapid and more intuitive (System I) processing, which is also thought to be more prone to bias and impulsivity. Cognitive load may be especially high among the poor, and the link between cognitive load and economic decision-making may be important for understanding poverty traps (Shah et al., 2012; Dean et al., 2019). Identifying mechanisms by which cognitive load affects economic choices can help to explain the links between behavioral biases and bounded rationality.

Because it is closely related to the use of working memory (Colom et al., 2004), researchers often study cognitive load using memorization tasks. A typical manipulation involves asking subjects to remember a multi-digit number before a decision task, and then to report the number after the decision. While the results of these studies have broadly been interpreted as consistent with dual-systems theory, the effects of manipulating cognitive load are mixed. For instance, Benjamin et al. (2013) find that cognitive load does not affect performance on an arithmetic task, whereas in Deck and Jahedi (2015) cognitive load affects performance on multiplication but not addition problems. Shiv and Fedorikhin (1999) and Hinson et al. (2003) find that cognitive load makes people more impulsive and impatient, but the same results are not observed in other studies (Benjamin et al., 2013; Deck and Jahedi, 2015). Carpenter et al. (2013) show that cognitive load reduces strategic sophistication, whereas in Allred et al. (2016), the relationship between cognitive load and strategic sophistication is not consistent. Drichoutis and Nayga (2020) find that cognitive load affects performance on a reasoning task, but does not affect economic rationality. While Whitney et al. (2008), Benjamin et al. (2013), and Deck and Jahedi (2015) observe that cognitive load increases small-stakes risk aversion, Blaywais and Rosenboim (2019) find that cognitive load increases risk taking in the form of higher bids for lottery tickets involving non-negative payoffs. Since the literature on cognitive load has not always produced a consistent result, more research is clearly needed to understand the source of this inconsistency.

A sound starting point is to replicate significant studies, keeping the experimental design as close as possible to the original, since some of this inconsistency may be due to differences in experimental design. While conceptual replications attempt to verify the underlying hypothesis of an earlier experiment, direct replications (e.g. Camerer et al., 2016) involve

the repetition of an experimental procedure, to verify that the results of the experiment are independent of the time, place, or persons involved (Schmidt, 2009). Maniadis et al. (2014) demonstrate that even a few numbers of independent replications could dramatically decrease the chance of type-I error or "false positive" results. We contribute to the literature by performing a direct replication of Experiment 1 of Deck and Jahedi (2015, henceforth DJ).

DJ study the effect of induced cognitive load on economic decisions in an incentivized laboratory experiment involving a number of different tasks. They find that cognitive load decreases arithmetic performance, increases risk aversion, and makes people more susceptible to anchoring. Their results, in particular the increase in risk aversion under cognitive load, are largely driven by the fraction of their sample whose performance on the arithmetic task is most affected by increased cognitive load. DJ suggest that each of these results are consistent with the dual-system framework, where cognitive load increases the workload of the reasoning system and therefore leads to more intuitive, impulsive, or biased decisions. This suggests that the effect of cognitive load may also vary according to cognitive ability or capacity. A recent model of information processing with capacity constraints suggests that small-stakes risk aversion may result from noisy perceptions of payoffs (Khaw et al., 2021), consistent with the notion that cognitive load diminishes information processing capacity.

We conducted a preregistered direct replication of DJ, using their procedures and experiment and data analysis code, obtained from the authors. To study whether effects of cognitive load are related to cognitive ability, we followed the DJ replication with a working memory task. Our design was powered to replicate the result that elevated cognitive load increases risk aversion (see DJ Table 1a). This sample size also provides more than 80% power for replications of DJ's results linking time preference and anchoring effects to cognitive load, and to replicate their result that increased cognitive load decreases arithmetic performance, which we view as a manipulation check. One contribution of DJ's design is that it measures the systematic effect of cognitive load across multiple tasks. However, these multiple measures increase the difficulty of determining what constitutes a successful replication. For example, DJ argue that their result demonstrating that cognitive load decreases impatience results from the fact that the intertemporal choice task was not incentivized in their Experiment 1, unlike the other tasks. They also conduct a second experiment, the results of which reverse the intertemporal choice from the Experiment 1. Nevertheless, in our preregistration statement, we defined a successful replication as showing significant results for the arithmetic, risk-taking, and intertemporal choice tasks. We did not include the anchoring result in our definition of a successful replication, because DJ's results do not

show a significant effect of increased cognitive load on anchoring, though we report on those results as well.

Although our experiment used the same instructions, experiment software, and incentives as DJ, we failed to replicate their overall results on risky choice and intertemporal choice. We do find an effect of cognitive load on performance in the arithmetic task, indicating that the manipulation was successful. While participants in our sample show a strong anchoring effect, we do not observe a significant effect of high cognitive load on anchoring.

DJ's results are largely driven by the fraction of participants for whom cognitive load has the strongest effect on arithmetic performance. While our replication study fails to find a main effect of increased cognitive load on risk taking, a plausible conjecture from DJ's paper is that this failure to replicate might result from a failure of the cognitive load manipulation or from subject pool differences in the effect of cognitive load in general. That is, the effects of cognitive load might be different in our two samples, due possibly to differences in cognitive abilities. We explore this conjecture using 1) DJ's method of splitting the sample according to the median of the performance decline for arithmetic, using the median from our sample; 2) splitting our sample using DJ's median; 3) Splitting our sample using participants' performance on the follow-up working memory task; and 4) Splitting our sample using performance on the Cognitive Reflection Test. As we demonstrate, none of these analyses demonstrates a significant relationship between cognitive load, cognitive ability, and overall risk-taking in our data.

DJ's risk-taking task involves involves both gain and loss frames. Their main result on risk taking aggregates these measures. When we look separately at our subjects' behavior for both gains and losses, we find that in the pooled data participants are more risk averse under cognitive load for gains, but less risk averse for losses. However, our aggregated results appear to be driven by the sample composition; analyses using fixed-effects regressions fail to find significant effects of cognitive load in our sample in either the gain or loss domain (see Section 4.3).

An alternative explanation for the variation in results may be that cognitive load manipulations force experiment participants to choose how they allocate their scarce cognitive resources. Variation in results from study to study may reflect how participants make this effort allocation decision. Some support for this view is that the participants in our replication study perform better on the memorization task than those in DJ's study. We return to this notion of subject-pool differences in effort allocation in the discussion.

Section 2 of our paper describes our preregistered replication of Experiment 1 from DJ. The results of the pre-registration are presented in Section 3. We present the exploratory analyses that use different strategies for splitting the sample into sensitive and insensitive groups in Section 4. In Section 5 we explore whether subject pool differences might explain our inability to reproduce DJ's results. Section 6 contains a summary of this exercise and conclusions.

2 Experiment

The replication study was preregistered at the Open Science Framework. The target sample size was determined by power analyses.[1] To best replicate DJ, we use a relatively similar sample which consisted primarily of American university students. DJ recruited participants on the campus of the University of Arkansas (UA sample, hereafter), while we recruited participants at Virginia Tech (VT sample, hereafter). Both schools are research-intensive land grant universities. We recruited a total of 218 subjects grouped into 11 sessions for our two-hour study, and report results for 198 participants.[2] The study was reviewed by Virginia Tech's institutional review board prior to data collection. All participants provided informed consent. As in DJ, participants were paid for exactly one of their choices, so including a $10 payment for participation, they earned an average of $23.13.

The goal of this project was to directly replicate Experiment 1 in DJ, so we obtained the original computer program from the authors.[3] We then reproduced the visual stimuli used in the program so that they were compatible with the monitors in our lab. Screenshots can be found in Figure A2. Following DJ, each session contained 80 trials divided into two blocks. Forty trials were completed under low cognitive load with the rest under high cognitive load. The order of the blocks was randomly determined for each participant. Within each block each participant saw a unique series of randomly generated decision problems. At the beginning of each session, participants were seated at randomly selected networked computers, and completed an interactive set of instructions (identical to DJ) to make sure they could identify correctly the payoff for each task type. At the start of each

[1] We conduct separate power analyses for each of the results we are trying to replicate. The minimum requirement for 90% power is 36 for arithmetic problems, 191 for risk tasks, and 186 for intertemporal tasks. Hence we recruited approximately 200 participants.

[2] Session 1's data (16 subjects) was dropped due to a significant, unintended time lag during the experiment, and 4 other participants were dropped for incomplete data.

[3] The program was developed in z-tree (Fischbacher, 2007).

trial, either a 1-digit (low cognitive load mode) or 8-digit number (high cognitive load mode) appeared on the screen for 3 seconds.[4] After each arithmetic, risk, time, or anchoring task, participants were asked to recall the number they saw. Participants earned $22 for correctly recalling the number, an amount DJ intentionally set much higher than what could be earned from a decision making task so that participants would prioritize the working memory task.

As in DJ Experiment 1, there were four categories of individual tasks: (1) arithmetic problems, (2) choices involving risk, (3) choices involving intertemporal substitution, and (4) choices with anchoring effects. Stimuli for each trial were generated by a random method described in DJ Table 2. The arithmetic problems included both addition and multiplication of a one-digit and a two-digit number, and the payoff for the correct answer was $12. The risk task involved gambles in either the gain or loss domain where the payoff depended on the outcome of the participants' choices. Each trial in the risk task required choosing between a sure payment on the left and a risky gamble on the right. In the gain domain, participants were given an initial endowment of $2, then chose between a sure gain of s and a 50/50 gamble of receiving $2s + 2$ or 0; in the loss domain, participants received an endowment of $2s + 4$, and chose between a sure loss of $s + 2$ and a 50/50 gamble of losing $2s + 2$ or 0, where s was an integer drawn from $\sim U\{8, 15\}$. Intertemporal choices were hypothetical, and each choice earned a flat payoff of $12. In the intertemporal choice task, participants chose either a smaller sooner option of $100 at time T on the left or a larger later option $\$F \in U\{105, 110, 115, 125\}$ at time $T + t$ on the right. T was equally likely to be today, in one week, or in one month, and t was equally likely to be one day, one week, or one month. When $T =$ one month, then t could only be one day or one week. For the anchoring effect task, participants first picked a number between 20 and 80. On the next screen, this number was displayed while participants counted how many S's were in a 10x10 matrix of 5's and S's. If the participant's count was within five of the accurate count, they earned $12.

Since time pressure increases the difficulty of tasks in involving working memory (Matthews and Campbell, 2010), all phases of each trial of the experiment were timed. Participants saw the one- or eight-digit number they were asked to remember for three seconds prior to completing the arithmetic or decision task and had nine seconds to enter the remembered number afterwards. They were allowed ten seconds to solve addition or multiplication tasks and eight seconds for the risk or intertemporal choice problems. In the anchoring task, participants were allowed eight seconds to enter a number, followed by ten seconds to count and input the number of S's. If time expired before the participant submitted required inputs,

[4]For the one-digit memorization task, participants could potentially "cheat" by placing one of their fingers at the keyboard, which would further reduce cognitive load. This was possible both for our study and DJ.

they earned $0 for that trial. Following each trial participants reviewed the outcome for five seconds before the next trial began.

Once the tasks from DJ's Experiment 1 concluded, our participants completed two additional tasks: a working memory task and an exit survey. The working memory task contained both forward and backward digit-span subtasks. In the digit-span task, participants were presented with a series of digits one at a time on the screen. In the forward digit-span task, participants entered the digits in the order in which they appeared, whereas in the backward digit-span task they were entered in the reverse order of appearance. The backward digit-span task is one of the most prevalent task in evaluating working memory capacity (Ramsay and Reynolds, 1995), which is an important component of general intelligence, and is included in the Wechsler Adult Intelligence Scale - Fourth Edition (WAIS-IV; Wechsler, 2008). The exit survey included the Cognitive Reflection Test (Frederick, 2005), a demographic questionnaire, and selected questions from the Global Preferences Survey (Falk et al., 2018).

3 Summary Statistics and Preregistered Replication

In the pre-registration document, we defined a successful replication as: "... a statistically significant effect (p-value less than 0.05) for math accuracy, risky choice, and intertemporal choice following Table 4 column 1a, 2a and 3a in the original paper." Because they provide a useful way to compare data sets, however, section 3.1 reproduces Table 3 in DJ, where we report summary statistics on task performance under high and low cognitive load for both UA and VT samples. The results of the preregistered hypotheses are found in section 3.2. Since DJ did not find an effect of cognitive load on anchoring in their Experiment 1 we did not preregister an anchoring hypothesis and will not discuss these results extensively. Anchoring results are, however, included in our tables for completeness.

3.1 Summary Statistics

We begin by looking at participants' performance on the digit memorization task (Table 1). Similar to the UA sample, the VT sample had more trouble remembering the 8-digit number than the 1 digit number (rank-sum test $p<0.001$). Comparing the performance across samples, however, shows that while the UA participants accurately recalled 43.3% of the numbers, VT participants recalled only 35.8%, a significant decrease (rank-sum test

p-values: low 0.006, high 0.000). So while participants from both samples perform worse as cognitive load increases, the VT participants were somewhat worse at the memorization task overall.

Table 1. Summary statistics (Table 3 in DJ)

	Low			High			Low vs. High		Diff-in-Diff	
	(A) UA	(B) VT	(C) P	(D) UA	(E) VT	(F) P	(G) UA	(H) VT	(I) mean	(J) P
Digit memorization	96.3%	95.3%	0.006	43.3%	35.8%	**0.000**	**0.000**	**0.000**	-0.009	**0.000**
	4480	7920		4480	7920					
Correct addition	97.8%	97.7%	0.951	96.9%	94.6%	0.105	0.5307	**0.000**	-0.003	0.191
	312	955		294	986					
Correct multiplication	71.6%	73.9%	0.328	55.9%	59.4%	0.189	**0.000**	**0.000**	0.002	0.738
	634	854		639	794					
Risky choice (gains)	59.5%	61.4%	0.444	52.7%	56.8%	0.099	0.013	0.040	0.003	0.531
	662	933		662	1012					
Risky choice (loss)	45.7%	44.6%	0.663	43.9%	48.3%	0.088	0.517	0.111	0.008	0.132
	588	946		640	951					
Early option	35.8%	36.6%	0.792	31.6%	33.7%	0.451	0.175	0.244	0.002	0.737
(today vs. future)	452	735		459	736					
Early option	30.2%	28.9%	0.507	25.9%	29.1%	0.107	0.050	0.887	0.007	0.107
(later vs. future)	871	1316		815	1277					
S-Count within	50.1%	39.9%	**0.000**	40.9%	31.6%	**0.000**	**0.000**	**0.000**	0.001	0.740
anchoring range	799	1690		876	1754					

Note: Numbers in percentage on top, and # of responses on bottom. Column C lists p-values of UA and VT comparison for low cognitive load condition, and column F lists p-values of UA and VT comparison for high cognitive load condition. Column G lists p-values of low and high cognitive load comparison for UA, and column H lists p-values of low and high cognitive load comparison for VT. P-values in columns C, F, G, H are obtained using Wilcoxon rank-sum test. Bolded p-values are those who survive Bonferroni correction at $\alpha = 0.05$ (adjusted $\alpha = 0.00125$). Column I lists estimations for difference-in-difference treatment effect, and column J lists p-values. UA results are reproduced using DJ's data.

Task performance between the samples looks similar in the risk and impatience tasks, but shows some differences in the anchoring task (Table 1 Columns C and F). The UA sample has significantly higher accuracy in the anchoring task, for both the low (50.1% vs. 39.9%, rank-sum test $p<0.001$) and high (40.9% vs. 31.6%, rank-sum test $p<0.001$) cognitive load trials. Overall, the VT sample looks similar to the UA sample in the arithmetic, risk, and intertemporal choice tasks.

There are, however, some between-sample differences in the statistics on the effect of increased cognitive load. While the VT sample performs better in the addition task under low cognitive load (rank-sum test $p<0.001$), the UA participants were not significantly affected ($p=0.105$). UA participants are less patient under low cognitive load in the later vs. future intertemporal choice task (rank-sum test $p=0.050$), but there is no difference in the VT sample ($p=0.886$). Regression analyses (Table A2) are mostly consistent with these results. Overall, our initial look at summary statistics suggests that the samples are largely comparable.

We next explore whether there are significant differences in the effect of cognitive load on the two samples by comparing the difference in performance on each task (Table 1, columns I and J). We find that the only significant difference between the two samples is in success at digit memorization, where the UA participants outperform those from VT (difference-in-difference estimation, mean=-0.009, $p<0.001$).

3.2 Results of the Preregistered Hypotheses

We now turn to our preregistered hypotheses, which were evaluated using DJ's data analysis scripts. We report analyses of both the UA and VT data in Table 2. As hypothesized, we can successfully replicate the result that cognitive load affects performance on arithmetic tasks (Column A). On the other hand, while UA participants' choices were both more risk averse and more patient when cognitive load increased, we find neither of these results in the VT data (Columns D and G). A strict interpretation of these results is that our planned replication failed. In the remainder of the paper, we explore the reasons for this failure to replicate.

Table 2. Effect of cognitive load manipulation on behavior in both samples.

UA Sample												
	Correct at arithmetic			Risky choice chosen			Earlier option chosen			Guess for Anchoring task		
	A (All)	B (-)	C (+)	D (All)	E (-)	F (+)	G (All)	H (-)	I (+)	J (All)	K (-)	L (+)
8-Digit number	-0.122***	0.033	-0.262***	-0.054**	-0.035	-0.071**	-0.041**	-0.057**	-0.027	0.159	2.391	-2.356
	(0.023)	(0.022)	(0.026)	(0.024)	(0.036)	(0.032)	(0.018)	(0.026)	(0.026)	(2.293)	(3.311)	(3.204)
Anchor										0.068*	0.050	0.094*
										(0.037)	(0.056)	(0.048)
Anchor x 8 digit										0.034	0.006	0.065
										(0.050)	(0.070)	(0.070)
Number of S's										0.708***	0.688***	0.729***
										(0.025)	(0.040)	(0.029)
1-Digit average	0.802	0.786	0.817	0.530	0.560	0.505	0.321	0.263	0.377			
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.232	0.289	0.220	0.306	0.317	0.289	0.315	0.318	0.289	0.675	0.662	0.689
Observations	1879	882	997	2552	1213	1339	2597	1270	1327	1675	824	851
VT Sample												
	Correct at arithmetic			Risky choice chosen			Earlier option chosen			Guess for Anchoring task		
	A (All)	B (-)	C (+)	D (All)	E (-)	F (+)	G (All)	H (-)	I (+)	J (All)	K (-)	L (+)
8-Digit number	-0.069***	0.021*	-0.160***	-0.005	-0.019	0.008	-0.007	0.004	-0.018	-3.573*	-3.060	-3.926
	(0.012)	(0.011)	(0.016)	(0.016)	(0.024)	(0.022)	(0.016)	(0.025)	(0.021)	(2.059)	(2.820)	(3.047)
Anchor										0.089***	0.099**	0.082*
										(0.031)	(0.043)	(0.044)
Anchor x 8 digit										0.062	0.047	0.074
										(0.043)	(0.060)	(0.062)
Number of S's										0.616***	0.653***	0.578***
										(0.023)	(0.034)	(0.029)
1-Digit average	0.865	0.832	0.898	0.530	0.524	0.534	0.316	0.338	0.296			
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.157	0.188	0.153	0.310	0.299	0.321	0.310	0.294	0.324	0.568	0.593	0.544
Observations	3589	1800	1789	3842	1870	1972	4064	2003	2061	3444	1703	1741

Note: Dependent variables vary and are listed at the top of each column. OLS regression, standard errors clustered by participant. (All) for the whole sample, (-) for cognitive load insensitive individuals, and (+) for cognitive load sensitive individuals. The row labeled "1-Digit average" corresponds to the mean performance of the 1-digit treatment group. For the VT sample, we follow DJ's method to median split the sample. Using this method, we can fully recover results for the UA sample. UA results are reproduced using DJ's data.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

4 Median Split Analysis

One of the interesting findings in DJ is that some participants are more affected by cognitive load across a range of tasks. In their analysis, DJ calculate the difference in success at multiplication between the low and high cognitive load conditions, then conduct a median split of the data. Participants whose scores changed the most are referred to as "sensitive" to cognitive load, with the least affected called "insensitive." DJ find that sensitive participants are 7.1% less likely to choose the risky option, a result consistent with previous findings that cognitive ability and risk preferences are related (Dohmen et al., 2010). They also find that insensitive individuals become more patient, a result they explore further in a later experiment. We next conduct an exploratory analysis of whether sensitive VT participants become more risk averse under high cognitive load.

4.1 Median Split Analysis of VT Data

We first divide the VT sample by performing a median split of the sample based on the VT participants' multiplication task performance (as in DJ). The participants whose choices changed the most in the multiplication task when the cognitive load increased are sensitive (+), and the others are not sensitive (-), see Table 2. Our median split perfectly mimics DJ's method (correlation Spearman's rho = 1, p-value < 0.001).

The results of this exercise are mixed for the VT sample. As in DJ, sensitive subjects perform worse in both types of arithmetic problems with high cognitive load ($p<0.001$). This demonstrates that the median split successfully separates the individuals with different cognitive load tolerance. Unlike the UA sample, however, we do not find that cognitive load impacts risk preference (sensitive $p=0.727$, insensitive $p=0.448$) or time preference (sensitive $p=0.392$, insensitive $p=0.885$), for either of the sensitivity classifications. This (null) result reinforces the results of section 3 that the cognitive load manipulation did not affect the VT participants' overall decision making in the risk or time preference tasks.

An assessment of the success of the replication should not rely solely on p-values, but also on the research prior and the statistical power of the experiment (Maniadis et al., 2017). We use a Bayesian replication test (Verhagen and Wagenmakers, 2014, henceforth VW) to quantify the evidence that the data provide for replication success or failure. The analysis compares two competing hypotheses for the data from the replication study: an idealized proponent's view, which assumes that the effect of cognitive load on task performance is

distributed according to the results from DJ (VW call this the proponent's replication hypothesis, H_r); and a skeptical one, which assumes that the effect of cognitive load on task performance is equal to 0 (H_0). We conduct 12 replication tests, matching the 12 regression results in Table 2 for the effect of the 8-digit number on performance in the arithmetic, risk, time, and anchoring tasks, overall and for the cognitive load insensitive and sensitive groups. The input and output of the test is listed in Table A1, and the test results are visualized in Figure A3. We use the t-statistics for the coefficients on the 8-Digit number and the numbers of observations from the UA and VT samples to compute the Bayes factor $\left(BF_{r0} = \frac{p(Y_{rep}|H_r)}{p(Y_{rep}|H_0)}\right)$, a ratio that measures the relative likelihood of the replication data Y_{rep} to occur under H_r (the hypothesis that the effect sizes from the VT sample are equal to the ones from the UA sample) versus H_0 (the hypothesis that the effect sizes of the VT sample are equal to 0). Following VW, for the idealized proponent's hypothesis H_r, the posterior distribution of the effect size from the UA sample is used as the prior for computing the likelihood. A typical rule of thumb is that Bayes factors greater than 3 constitute a successful replication (Rouder et al., 2017). Focusing on our preregistered hypotheses for the cognitive load sensitive group, the Bayes factor BF_{r0} is equal to $2.86e+18, 0.105, 0.874$, and 1.3 for the Arithmetic, Risk, Time, and Anchoring tasks, respectively (Table A1, Figure A3). Again, these analyses are consistent with the conclusion that the VT experiment replicates the result that cognitive load affects arithmetic task performance, but not risk and time preferences or anchoring.

4.2 Median Split Analysis Using the DJ Sample Median

Our failure to reproduce the sensitivity result could be due to a difference in the median of the absolute performance drop in multiplication between the two samples (UA median: -0.17, VT median: -0.07). In this section, we attempt to reproduce Table 4 in DJ using the median of the UA data to split the VT sample.

This creates an unbalanced sample for cognitive load insensitive and cognitive load sensitive groups: more VT subjects are cognitive load insensitive (2013 insensitive vs. 1576 sensitive observations in arithmetic problems). As shown in Table 3, we successfully partition the VT sample so that cognitive load sensitive individuals perform worse under high cognitive load in arithmetic task ($p<0.001$). The results for risk (sensitive $p=0.931$, insensitive $p=0.628$) and time preferences (sensitive $p=0.562$, insensitive $p=0.902$) are still similar to what we found in section 4.1.

Table 3. Effect of cognitive load in VT sample using DJ median to perform median split.

VT Sample												
	Correct at arithmetic			Risky choice chosen			Earlier option chosen			Guess for Anchoring task		
	A (All)	B (-)	C (+)	D (All)	E (-)	F (+)	G (All)	H (-)	I (+)	J (All)	K (-)	L (+)
8-Digit number	-0.069***	0.010	-0.171***	-0.005	-0.011	0.002	-0.007	-0.003	-0.012	-3.573*	-2.420	-4.674
	(0.012)	(0.011)	(0.018)	(0.016)	(0.023)	(0.023)	(0.016)	(0.024)	(0.021)	(2.059)	(2.592)	(3.377)
Anchor										0.089***	0.093**	0.085*
										(0.031)	(0.040)	(0.048)
Anchor x 8 digit										0.062	0.033	0.090
										(0.043)	(0.056)	(0.067)
Number of S's										0.616***	0.649***	0.573***
										(0.023)	(0.032)	(0.031)
1-Digit average	0.865	0.832	0.907	0.530	0.524	0.536	0.316	0.333	0.297			
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.157	0.187	0.150	0.310	0.299	0.323	0.310	0.291	0.333	0.568	0.593	0.539
Observations	3589	2013	1576	3842	2049	1793	4064	2210	1854	3444	1866	1578

Note: Dependent variables vary and are listed at the top of each column. OLS regression, standard errors clustered by participant. (All) for the whole sample, (-) for cognitive load insensitive individuals, and (+) for cognitive load sensitive individuals. The row labeled "1-Digit average" corresponds to the mean performance of the 1-digit treatment group. This provides the baseline performance in each task. For the VT sample, we follow DJ's data, and use the exact cutoff point to median split the sample. Using this cutoff point, we again can fully recover results for the UA sample. UA results are reproduced using DJ's data.

* p < 0.1, ** p < 0.05, *** p < 0.01.

We next plot the density and histogram for the absolute performance drop in the multiplication task (Figure 1) to explore the effect of the median split on the two samples. Even though the sample means are not different (two-sided t-test p=0.945), the sample distributions are marginally different (Epps-Singleton test p=0.071), and the standard deviation for the UA sample is significantly smaller than that for the VT sample (Variance comparison test p=0.010). It is possible that we observe inconsistent results in Table 2 due to differences in the absolute performance drop in the multiplication task. In the next subsection we attempt to identify the basis for this inconsistency by considering different median split strategies.

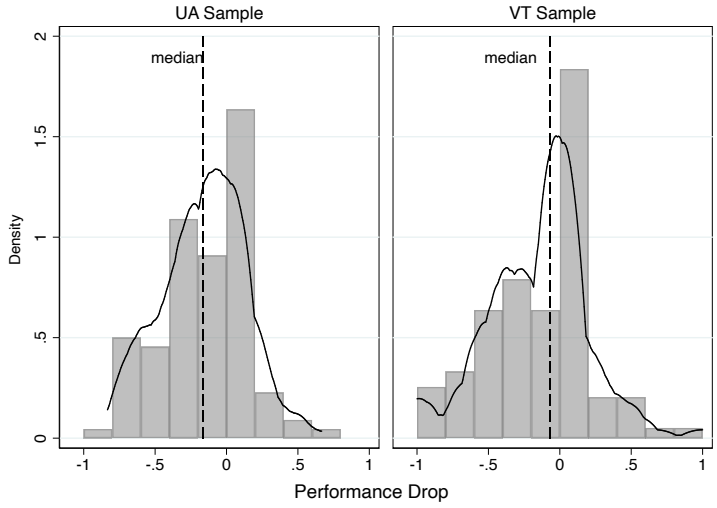


Figure 1. Histogram of Absolute Performance Drop in the Multiplication Task

4.3 Alternate Median Split Analyses

We conjectured that one of the reasons that we failed to reproduce the risk results is that the effect of cognitive load may differ in the gain and loss domains. Table 4 returns to the median split strategy from Section 4.1 that follows DJ but divides participants' risk decisions into gains and losses. In the gain domain, we find that UA but not VT participants are sensitive to cognitive load. In the loss domain we find that neither sample shows an effect of cognitive load. It is possible that participants from different subject pools may have different perceptions of the loss domain. While some participants might perceive the loss condition as intended, for others the choices in the loss condition might still represent gains due to the "house money effect," in which all tasks represent gains relative to the wealth with which participants began the experiment (Cárdenas et al., 2014).

Table 4. Effect of cognitive load on risky gains and losses

	UA Risky Gain			VT Risky Gain			UA Risky Loss			VT Risky Loss		
	A (All) coef./se	B (Not) coef./se	C (Sens) coef./se	D (All) coef./se	E (Not) coef./se	F (Sens) coef./se	G (All) coef./se	H (Not) coef./se	I (Sens) coef./se	J (All) coef./se	K (Not) coef./se	L (Sens) coef./se
8-Digit number	-0.048	0.019	-0.107***	-0.025	-0.027	-0.023	-0.044	-0.071	-0.018	0.002	-0.020	0.021
	(0.031)	(0.048)	(0.040)	(0.021)	(0.031)	(0.029)	(0.035)	(0.051)	(0.049)	(0.023)	(0.033)	(0.032)
Observations	1324	626	698	1945	957	988	1228	587	641	1897	913	984
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.413	0.392	0.431	0.409	0.386	0.425	0.411	0.417	0.400	0.439	0.408	0.458

Note: Dependent variable is the percent of risky choices chosen. OLS regression, standard errors clustered by participant.
* p < 0.1, ** p < 0.05, *** p < 0.01.

We next take advantage of the additional data we collected for the VT sample. We perform three separate median splits of the data based on performance in the CRT, the backward digit span task and the forward digit span task. Since these are all measures of cognitive ability, we conjectured that scores on these tasks might interact with the cognitive load manipulation to produce changes in risk-taking. The CRT is highly correlated with cognitive ability in general (Pennycook and Ross, 2016) and working memory (Stupple et al., 2013) in particular. While participant scores on these three measures are correlated with each other, none are significantly correlated with DJ's sensitivity measure.[5]

For the CRT, the VT sample is partitioned into the bad group (participants who answered all 3 CRT questions wrong) and good group (median of CRT score was 1). For each of the

[5]Insignificant correlation between cognitive load sensitivity and CRT (spearman's rho=-0.112, p=0.116), insignificant correlation between cognitive load sensitivity and digit span backward (spearman's rho=-0.096, p=0.178), insignificant correlation between cognitive load sensitivity and digit span forward (spearman's rho=0.061, p=0.392). Significant correlation between CRT and digit span backward (spearman's rho=0.328, p<0.001), significant correlation between CRT and digit span forward (spearman's rho=0.130, p=0.069), significant correlation between digit span backward and digit span forward (spearman's rho=0.337, p<0.001).

memory task results, the samples are also sorted into the bad group, who remembered fewer digits than the median, and the good group (median in backward digit span task = 9, median in forward digit span task=11). The results on the frequency of selecting the risky option for each of these are found in Table 5. We find no evidence that performance on any of these tasks is associated with participants' choices.

Table 5. Alternative median split methods on risk task.

VT Sample									
	CRT Score			Digit Span Backward			Digit Span Forward		
	A (All)	B (Bad)	C (Good)	D (All)	E (Bad)	F (Good)	G (All)	H (Bad)	I (Good)
8-Digit number	-0.005	-0.007	-0.003	-0.005	-0.026	0.009	-0.005	-0.010	-0.001
	(0.016)	(0.023)	(0.022)	(0.016)	(0.028)	(0.019)	(0.016)	(0.026)	(0.020)
Observations	3842	1718	2124	3842	1575	2267	3842	1848	1994
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.310	0.291	0.318	0.310	0.336	0.292	0.310	0.287	0.328

Note: Dependent variable is the percent of risky choices chosen. OLS regression, standard errors clustered by participant. CRT Score is denoted as good if answers at least one question correctly out of three (median of CRT score is 1). Digit Span Backward/Forward is denoted as good if number of correct recalls is greater or equal to the median performance (median of backward is 9, median of forward is 11).
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

5 Exploring subject pool differences

We sought an explanation for why we were unable to replicate many of DJ's results despite using their experiment code, procedures and data analysis scripts. One plausible candidate was that there are differences, perhaps in cognitive ability, between the participants in the two experiments. Unfortunately, a formal test is not possible as data on cognitive ability is not available for the DJ sample.

As an exploratory analysis, we looked for evidence of differences in the undergraduate student populations, to which most participants belong. To do this, we compared admissions statistics (Common Data Set (CDS) Initiative) between the two universities (University of Arkansas, 2019; Virginia Tech, 2019), focusing on CDS-C (First-Time, First-Year (Freshmen) Admission).[6] More VT first-year enrolled students had high school grade point averages of 3.75 and higher (UA 44.94%, VT 78.32%) with mean high school grade point averages of 3.98 for VT and 3.67 for UA. In addition, on average more VT students rank in the top ten percent of their high school graduating class (UA 25.65%, VT 39.16%), and more UA students rank bottom half of their high school graduating class (UA 14.83%, VT 2.28%). Since the experimental tasks are numerical in nature, we next looked at standardized test

[6]CDS data is extracted from the university website. Since VT and UA experiments were run in different years, we use 5 years CDS during the data collection period (from 2014-2015 to 2018-2019).

scores to explore differences in quantitative ability.[7] Over the last 5 years, 21.75% VT students scored higher than 700 in SAT Math compared to 6.95% for UA students (Figure A4(a)). For students who took the ACT, 31.30% VT students score above 30 in Math, compared to 12.19% for UA students (Figure A4(b)).[8] Based on the above descriptive differences in the two sample pools, VT appear to be slightly better on average academically, which could translate into a difference in cognitive ability that makes them less affected by the experiment's cognitive load treatment. On the other hand, without collecting new data there is no way to determine if this difference is sufficient to explain our results.

We next looked within the VT sample in the hope of finding some evidence that participant differences influenced the effects of cognitive load. We created two sub-groups of participants from within the VT sample: the first reported having majors in our College of Engineering (Eng) and the second reported having majors in our College of Liberal Arts and Human Sciences (LAHS). Note that in the current VT university organization, majors such as mathematics, statistics, physics and economics are in neither of these colleges. VT engineering majors are very math intensive, requiring two full years of college math at the calculus level and beyond. On the other hand, most LAHS majors are not required to take calculus. Table 6 presents task performance with these subsets of the VT data. We find significant performance differences between the two groups in the mathematics and risk tasks. While intriguing, these results are not sufficient to establish subject pool differences as the cause of our replication failure.

[7]VT started to collect ACT scores from 2017-2018. Students have different preferences for submitting exam scores. For UA, 25% enrolled students submitted SAT scores, and 92% submitted ACT scores. For VT, 87% enrolled students submitted SAT scores, and 38% submitted ACT scores.

[8]VT students also score higher in SAT Critical Reading, ACT Composite, and ACT English, when comparing percent of students (Figure A4).

Table 6. Task Performance Differences in College Majors (VT Sample)

	Low			High			Low vs. High		Diff-in-Diff	
	(A) LAHS	(B) Eng	(C) P	(D) LAHS	(E) Eng	(F) P	(G) LAHS	(H) Eng	(I) mean	(J) P
Digit memorization	95.4%	95.4%	0.927	31.6%	45.1%	**0.000**	**0.000**	**0.000**	-0.019	**0.000**
	1160	1680		1160	1680					
Correct addition	97.0%	99.1%	0.139	92.1%	96.7%	0.047	0.075	0.087	-0.004	0.356
	132	219		151	215					
Correct multiplication	74.6%	85.1%	0.021	54.87%	78.2%	**0.000**	0.002	0.088	-0.018	0.068
	122	194		113	174					
Risky choice (gains)	62.8%	54.2%	0.114	60.7%	49.3%	0.033	0.712	0.329	0.004	0.713
	145	192		150	207					
Risky choice (loss)	45.9%	32.3%	0.009	43.5%	39.2%	0.419	0.685	0.161	-0.013	0.221
	157	201		147	189					
Early option	44.6%	35.6%	0.158	44.1%	34.9%	0.119	0.952	0.892	0.000	0.970
(today vs. future)	101	146		111	172					
Early option	32.6%	34.6%	0.651	34.9%	26.6%	0.056	0.643	0.041	0.015	0.098
(later vs. future)	190	280		195	274					
S-Count within	36.5%	40.9%	0.275	28.8%	33.3%	0.239	0.068	0.035	-0.000	0.978
anchoring range	252	362		243	369					

Note: Numbers in percentage on top, and # of responses on bottom. Eng stands for students major in engineering, and LAHS stands for students major in Liberal Arts and Human Sciences. Column C lists p-values of Eng and LAHS comparison for low cognitive load condition, and column F lists p-values of Eng and LAHS comparison for high cognitive load. Column G lists p-values of low and high cognitive load comparison for LAHS, and column H lists p-values of low and high cognitive load comparison for Eng. P-values in columns C, F, G, H are obtained using Wilcoxon rank-sum test. Column I lists estimations for difference-in-difference treatment effect, and column J lists p-values. Bolded p-values are those who survive Bonferroni correction at $\alpha = 0.05$ (adjusted $\alpha = 0.00125$).

6 Conclusion

We conduct a preregistered replication of Deck and Jahedi (2015) Experiment 1 which explores the effect of cognitive load on one's ability to do arithmetic and make economic decisions. In our sample we are able to replicate the effect of high cognitive load on participants' ability to correctly do multiplication problems, which serves as a manipulation check for the digit memorization procedure. Our data fails to confirm the other two preregistered hypotheses about the effect of cognitive load on risky choice and impatience.

In a sense, a failure to fully replicate DJ's cognitive load results is not surprising. In fact, the original authors present a nice summary of previous results on cognitive load which shows they are often mixed. Kessler and Meier (2014), for example, fail to replicate their own result on how cognitive load affects charitable giving, attributing the replication failure to the order in which the cognitive load task occurred in a series of tasks. While this explanation does not apply to DJ's experiments, it suggests that cognitive load manipulations may be relatively more sensitive to details of the experimental design than other experimental treatments. In general, experimenters who wish to maximize the chance of successful replication should, therefore, use the same instructions, procedures, task order and presentation, and subject pool. By using DJ's experiment code, we have done our best to address the first three concerns.

We made several attempts to identify whether differences in the participant pool are responsible for our failure to replicate. Since DJ did not report demographic data on partic-

ipants, we instead review university admissions data and program rankings and find some significant differences in admissions statistics between the student populations at UA and VT. We hypothesized that differences in cognitive ability between our subject pools might provide some insight into the differences we observe. If the differences were great enough, for example, we might expect the participants from the school with higher admissions standards to be less risk averse, in keeping with previous work on the relationship between numeracy and risk aversion (Dohmen et al., 2010; Benjamin et al., 2013, e.g.). That VT students were generally somewhat (but not significantly) more willing to take risks suggests a path for future research.

Distributional analyses might have helped to determine whether these differences hold for subsamples of the subject populations. DJ's median split analysis reveals that, among the UA participants, cognitive load changes risk and impatience decision making only for those who are found to be sensitive to its effects based on multiplication accuracy. However, this result does not hold for VT participants. On the other hand, looking at the summary statistics for the risk task in the loss domain, for example, there is some evidence that increasing cognitive load made VT students slightly more willing to choose risky gambles, whereas the UA students' behavior was almost unchanged. This is consistent with the theory that cognitive load amplifies the reflection effect, wherein people are risk averse in the gain domain and risk seeking in the loss domain. Our additional explorations involved splitting the VT subject pool using the DJ median split point and scores on both memory tasks and the Cognitive Reflection Test. We were ultimately unsuccessful in finding a way to split the VT sample to recover the UA sensitivity result.

The most notable difference between the subject pools is that the UA participants performed better than VT participants on the digit memorization task in both the high and low cognitive load treatments. This could be evidence that UA participants have better memories. It is also possible that UA and VT students used different multitasking strategies with UA students giving higher priority to the memorization task in order to increase the likelihood of earning $22 for correct memorization. Therefore, future research should consider the possibility that people strategically multitask to avoid cognitive resource exhaustion, and might try to identify the extent to which memorization tasks deplete cognitive resources. Also, future research into whether participant differences matter might involve repeating this experiment at two universities with different student populations and collecting subject level data on numeracy, memory, the Cognitive Reflection Test and relevant socioeconomic data.

Recently, Deck et al. (2021) examined the robustness of the effect of different cognitive load manipulations on task performance and decision-making in math problems, lottery tasks, logic puzzles, and distributional (sharing) decisions in a study conducted at Chapman University. They found that cognitive load decreased performance on the math problems and logic puzzles, increased risk aversion in the lottery tasks, and had no effect on distributional choices. In this study, as in DJ, subjects for whom number memorization had the most detrimental effect on math performance exhibit a greater increase in risk aversion under cognitive load. In addition, subjects who perform well on the Cognitive Reflection Test are the most affected by cognitive load. The authors argue that participants who exert cognitive effort (as opposed to those who are more impulsive) are more susceptible to the effects of cognitive load on task performance. As noted above, we also measured performance on the CRT, but did not find a relationship between CRT performance and the effect of cognitive load on risk-taking. Understanding why these results vary across subject populations will (of course) require more studies, both original and replications.

In summary, our results add to the complex literature on the relationship between working memory load and economic decision making. Our replication study not only generates an additional set of data testing the existing paradigm but also explores how subject pool differences might influence the effect of cognitive load manipulations. The inconsistent results from the VT and UA samples suggest that researchers should pay special attention to strategic multitasking under high cognitive load and to individual differences in cognitive constraints. Cognitive load and cognitive ability clearly play a role in economic decision-making; however, many questions remain. In particular, future work on cognitive load and risk-taking should seriously consider how to separate increased noise from changes in preferences. Our work also highlights that replication studies, though seemingly straightforward, still require judgment in design and interpretation. We do observe a relatively strong effect of cognitive load on anchoring, using DJ's "within-range" approach. While a strict interpretation is that we failed to replicate DJ's results, our preferred conclusion is more nuanced. Some of our experiment results are consistent with DJ's dual-process motivation, and more research is needed to understand how people allocate attention and effort under cognitive load. We hope that our study motivates future research probing the fundamental mechanisms that govern the effect of cognitive load and the allocation of cognitive resources during economic decision-making.

Appendices

Table A1. Bayesian Replication Test Results

	Correct at arithmetic			Risky choice chosen			Earlier option chosen			Guess for Anchoring task		
	A (All)	B (-)	C (+)	D (All)	E (-)	F (+)	G (All)	H (-)	I (+)	J (All)	K (-)	L (+)
t_{org}	-5.43	1.48	-10.2	-2.28	-0.998	-2.21	-2.28	-2.22	-1.03	0.0694	0.722	-0.736
N_{org}	1879	882	997	2552	1213	1339	2597	1270	1327	1675	824	851
t_{rep}	-5.76	1.82	-9.69	-0.312	-0.762	0.35	-0.444	0.145	-0.861	-1.74	-1.09	-1.29
N_{rep}	3589	1800	1789	3842	1870	1972	4064	2003	2061	3444	1703	1741
$p\left(Y_{rep}\vert H_r\right)$	0.139	0.225	0.0147	0.0739	0.239	0.0395	0.0804	0.0468	0.241	0.131	0.109	0.226
$p\left(Y_{rep}\vert H_0\right)$	2.62e-8	0.0756	5.11e-21	0.38	0.298	0.375	0.362	0.395	0.275	0.0886	0.221	0.174
Bayes Factor	5.31e+6	2.98	2.87e+18	0.194	0.802	0.105	0.222	0.119	0.875	1.48	0.493	1.3

Note: t-statistics and number of observations from the original study and the replication study ($t_{org}, N_{org}, t_{rep}, N_{rep}$) are used as inputs of the Bayesian replication test. The Bayes Factor BF_{r0} is the output of the test, which is the ratio of the likelihood of the replication data Y_{rep} under H_r ($p\left(Y_{rep}\vert H_r\right)$) to the likelihood of the the replication data Y_{rep} to occur under H_0 ($p\left(Y_{rep}\vert H_0\right)$). A typical rule of thumb is that Bayes factors greater than 3 constitute a successful replication (Rouder et al., 2017).

Table A2. The effect of cognitive load on sub-task.

UA Sample

	Correct at arithmetic		Risky choice chosen		Earlier option chosen		Within range
	Addition	Multiplication	Risk (gains)	Risk (losses)	Now vs. future	Later vs. future	Anchoring
8-Digit number	-0.008	-0.157***	-0.068**	-0.018	-0.043	-0.043*	-0.092***
	(0.019)	(0.031)	(0.033)	(0.035)	(0.029)	(0.022)	(0.027)
Constant	0.978***	0.716***	0.595***	0.457***	0.358***	0.302***	0.501***
	(0.016)	(0.028)	(0.032)	(0.036)	(0.032)	(0.026)	(0.022)
Observations	606	1273	1324	1228	911	1686	1675
R^2	0.001	0.027	0.005	0.000	0.002	0.002	0.009

VT Sample

	Correct at arithmetic		Risky choice chosen		Earlier option chosen		Within range
	Addition	Multiplication	Risk (gains)	Risk (losses)	Now vs. future	Later vs. future	Anchoring
8-Digit number	-0.031***	-0.146***	-0.046*	0.037	-0.029	0.003	-0.082***
	(0.009)	(0.023)	(0.023)	(0.025)	(0.025)	(0.021)	(0.019)
Constant	0.977***	0.739***	0.614***	0.446***	0.366***	0.289***	0.399***
	(0.005)	(0.023)	(0.025)	(0.027)	(0.026)	(0.021)	(0.017)
Observations	1941	1648	1945	1897	1471	2593	3444
R^2	0.006	0.024	0.002	0.001	0.001	0.000	0.007

Note: Dependent variables vary and are listed at the top of each column. OLS regression, standard errors clustered by participant.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A3. Effect of cognitive load manipulation on math performance.

VT Sample												
	Cognitive Load Sensitivity			CRT Score			Digit Span Backward			Digit Span Forward		
	A (All)	B (-)	C (+)	D (All)	E (Low)	F (High)	G (All)	H (Bad)	I (Good)	J (All)	K (Bad)	L (Good)
8-Digit number	-0.069***	0.021*	-0.160***	-0.069***	-0.097***	-0.049***	-0.069***	-0.096***	-0.053***	-0.069***	-0.038**	-0.097***
	(0.012)	(0.011)	(0.016)	(0.012)	(0.021)	(0.014)	(0.012)	(0.023)	(0.013)	(0.012)	(0.017)	(0.016)
Observations	3589	1800	1789	3589	1526	2063	3589	1359	2230	3589	1679	1910
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.157	0.188	0.153	0.157	0.161	0.148	0.157	0.141	0.161	0.157	0.123	0.192

Note: Dependent variable is the accuracy of math problems. OLS regression, standard errors clustered by participant. Cognitive Load Sensitivity defined same as in Table 2. CRT Score is denoted as high if answers at least one question correctly out of three. Digit Span Backward/Forward is denoted as good if number of correct recalls is greater or equal to the median performance.

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A4. Effect of cognitive load manipulation on impatient choices.

VT Sample												
	Cognitive Load Sensitivity			CRT Score			Digit Span Backward			Digit Span Forward		
	A (All)	B (-)	C (+)	D (All)	E (Low)	F (High)	G (All)	H (Bad)	I (Good)	J (All)	K (Bad)	L (Good)
8-Digit number	-0.007	0.004	-0.018	-0.007	-0.024	0.006	-0.007	-0.010	-0.005	-0.007	-0.023	0.007
	(0.016)	(0.025)	(0.021)	(0.016)	(0.025)	(0.021)	(0.016)	(0.026)	(0.021)	(0.016)	(0.023)	(0.023)
Observations	4064	2003	2061	4064	1802	2262	4064	1578	2486	4064	1926	2138
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.310	0.294	0.324	0.310	0.324	0.296	0.310	0.295	0.320	0.310	0.308	0.312

Note: Dependent variable is the percent of impatient choices. OLS regression, standard errors clustered by participant. Cognitive Load Sensitivity defined same as in Table 2. CRT Score is denoted as high if answers at least one question correctly out of three. Digit Span Backward/Forward is denoted as good if number of correct recalls is greater or equal to the median performance.

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A5. Effect of cognitive load manipulation on anchoring effect.

VT Sample												
	Cognitive Load Sensitivity			CRT Score			Digit Span Backward			Digit Span Forward		
	A (All)	B (-)	C (+)	D (All)	E (Low)	F (High)	G (All)	H (Bad)	I (Good)	J (All)	K (Bad)	L (Good)
8-Digit number	-3.573*	-3.060	-3.926	-3.573*	-2.527	-3.900	-3.573*	-4.866	-2.946	-3.573*	-7.116**	-0.411
	(2.059)	(2.820)	(3.047)	(2.059)	(3.577)	(2.408)	(2.059)	(3.705)	(2.463)	(2.059)	(2.975)	(2.773)
Anchor	0.089***	0.099**	0.082*	0.089***	0.136***	0.064	0.089***	0.048	0.112***	0.089***	0.096**	0.069
	(0.031)	(0.043)	(0.044)	(0.031)	(0.045)	(0.041)	(0.031)	(0.052)	(0.038)	(0.031)	(0.043)	(0.043)
Anchor x 8 digit	0.062	0.047	0.074	0.062	0.040	0.062	0.062	0.092	0.047	0.062	0.093	0.036
	(0.043)	(0.060)	(0.062)	(0.043)	(0.072)	(0.051)	(0.043)	(0.075)	(0.053)	(0.043)	(0.061)	(0.060)
Number of S's	0.616***	0.653***	0.578***	0.616***	0.519***	0.692***	0.616***	0.595***	0.629***	0.616***	0.602***	0.629***
	(0.023)	(0.034)	(0.029)	(0.023)	(0.036)	(0.026)	(0.023)	(0.033)	(0.031)	(0.023)	(0.035)	(0.029)
Observations	3444	1703	1741	3444	1479	1965	3444	1336	2108	3444	1658	1786
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.568	0.593	0.544	0.568	0.494	0.636	0.568	0.557	0.574	0.568	0.538	0.601

Note: Dependent variable is the guess for the anchoring task. OLS regression, standard errors clustered by participant. Cognitive Load Sensitivity defined same as in Table 2. CRT Score is denoted as high if answers at least one question correctly out of three. Digit Span Backward/Forward is denoted as good if number of correct recalls is greater or equal to the median performance.

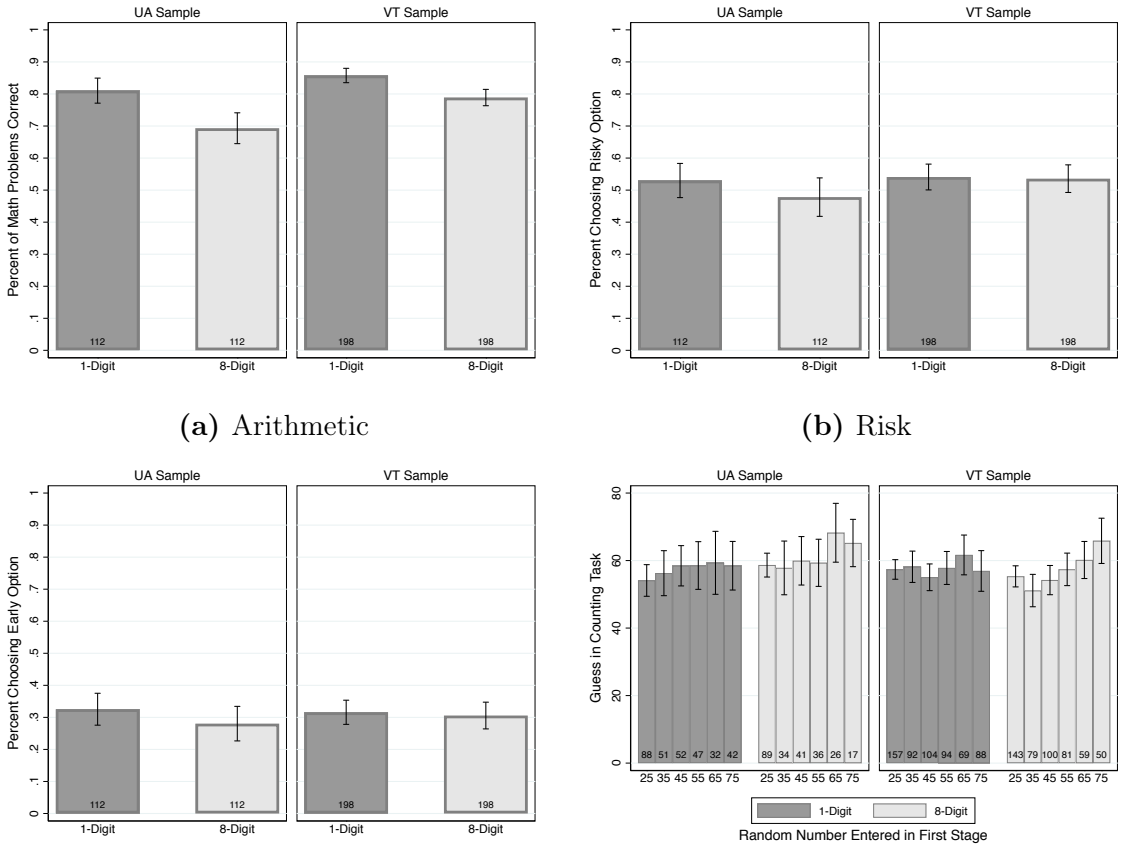
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A6. Effect of cognitive load manipulation on behavior with 3 groups

UA Sample												
	Correct at arithmetic			Risky choice chosen			Earlier option chosen			Guess for Anchoring task		
	A (All)	B (-)	C (+)	D (All)	E (-)	F (+)	G (All)	H (-)	I (+)	J (All)	K (-)	L (+)
8-Digit number	-0.122***	0.108***	-0.336***	-0.054**	-0.065	-0.123***	-0.041**	-0.043	-0.011	0.159	0.388	-2.743
	(0.023)	(0.036)	(0.030)	(0.024)	(0.046)	(0.037)	(0.018)	(0.029)	(0.035)	(2.293)	(5.830)	(5.243)
Anchor										0.068*	0.102	0.158**
										(0.037)	(0.115)	(0.063)
Anchor x 8 digit										0.034	0.036	0.057
										(0.050)	(0.120)	(0.104)
Number of S's										0.708***	0.703***	0.702***
										(0.025)	(0.058)	(0.044)
Observations	1879	416	583	2552	558	804	2597	607	789	1675	360	496
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.232	0.147	0.248	0.306	0.342	0.274	0.315	0.307	0.325	0.675	0.673	0.652
VT Sample												
	Correct at arithmetic			Risky choice chosen			Earlier option chosen			Guess for Anchoring task		
	A (All)	B (-)	C (+)	D (All)	E (-)	F (+)	G (All)	H (-)	I (+)	J (All)	K (-)	L (+)
8-Digit number	-0.069***	0.082***	-0.185***	-0.005	0.011	-0.015	-0.007	-0.007	-0.028	-3.573*	-0.407	-6.505
	(0.012)	(0.019)	(0.022)	(0.016)	(0.035)	(0.026)	(0.016)	(0.035)	(0.026)	(2.059)	(4.688)	(3.959)
Anchor										0.089***	0.097	0.077
										(0.031)	(0.080)	(0.057)
Anchor x 8 digit										0.062	-0.005	0.126
										(0.043)	(0.095)	(0.079)
Number of S's										0.616***	0.659***	0.528***
										(0.023)	(0.042)	(0.037)
Observations	3589	809	1177	3842	858	1375	4064	899	1386	3444	751	1190
Subject fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.157	0.099	0.136	0.310	0.284	0.341	0.310	0.363	0.337	0.568	0.608	0.510

Note: Dependent variables vary and are listed at the top of each column. OLS regression, standard errors clustered by participant. (All) for the whole sample, (-) for 1/3 of the sample who are most cognitive load insensitive, and (+) for 1/3 of the sample who are most cognitive load sensitive. For the VT sample, we follow DJ's method to median split the sample.

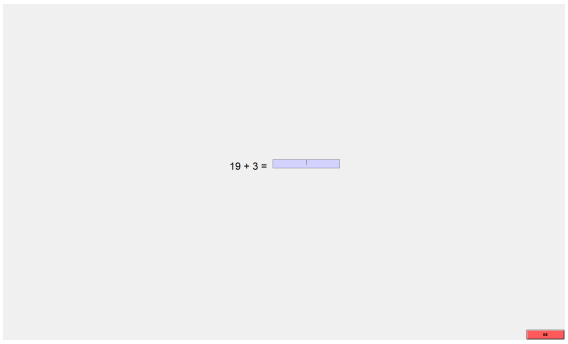
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.



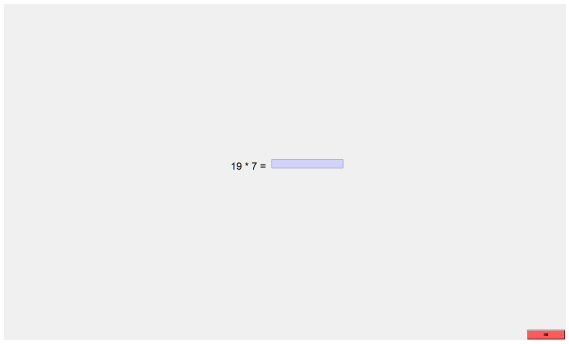
(a) Arithmetic **(b)** Risk

(c) Impatience **(d)** Anchoring

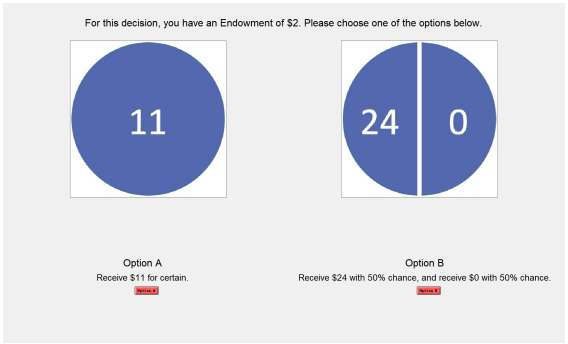
Figure A1. The treatment effect of cognitive load across the four main tasks.



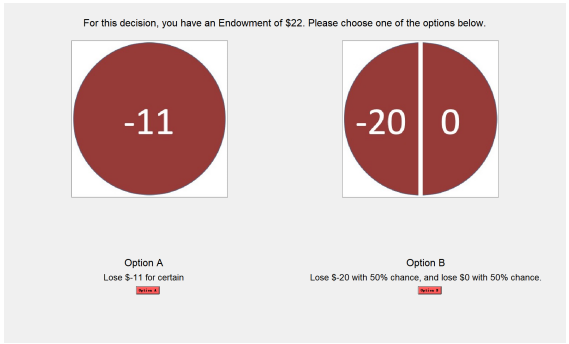
(a) Arithmetic - Addition



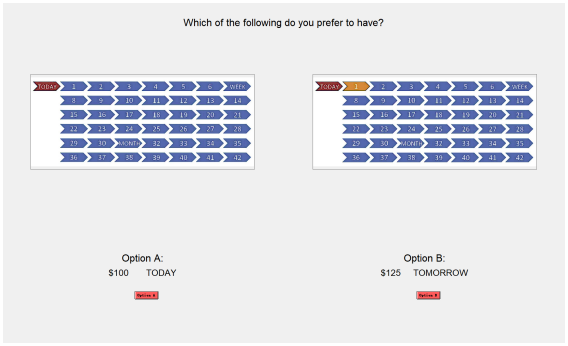
(b) Arithmetic - Multiplication



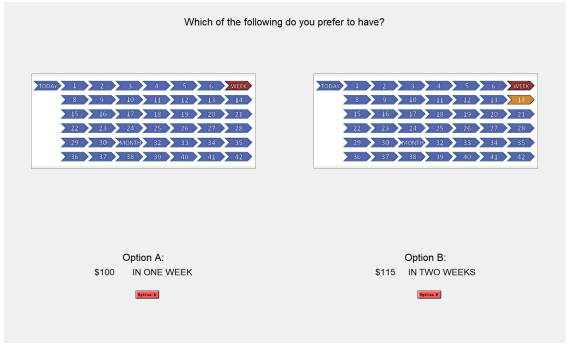
(c) Risk - Gain Domain



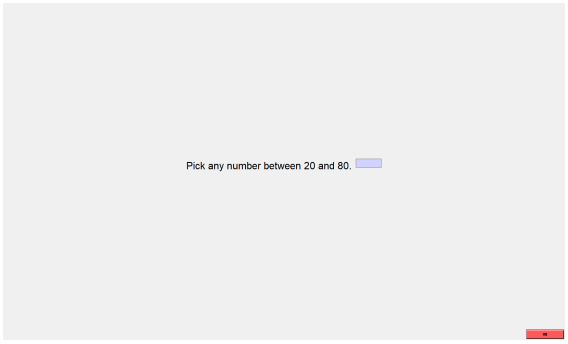
(d) Risk - Loss Domain



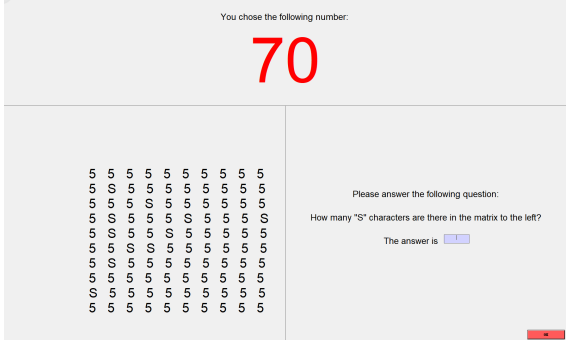
(e) Impatience - Now vs. Future



(f) Impatience - Later vs. Future

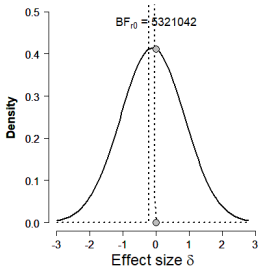


(g) Anchoring - Stage 1

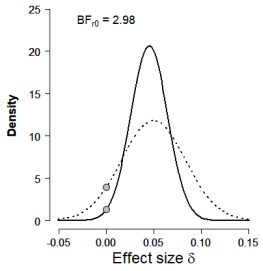


(h) Anchoring - Stage 2

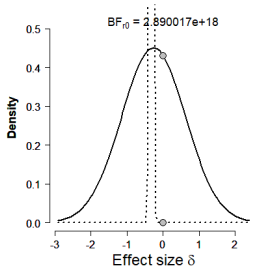
Figure A2. Experimental Screenshots for Individual Tasks



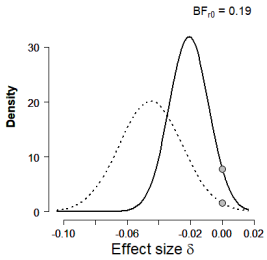
(a) Arithmetic All



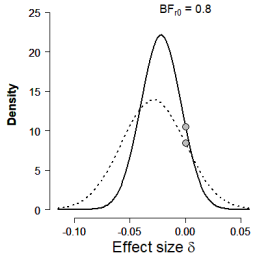
(b) Arithmetic (-)



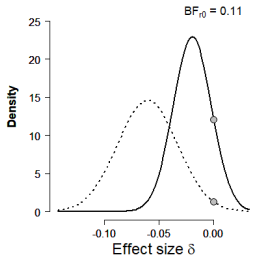
(c) Arithmetic (+)



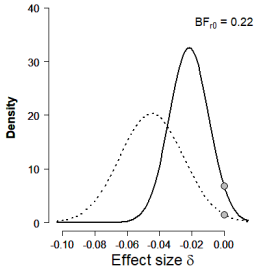
(d) Risk All



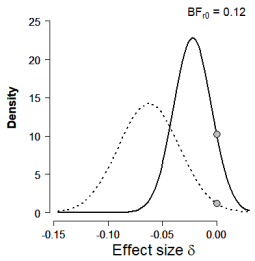
(e) Risk (-)



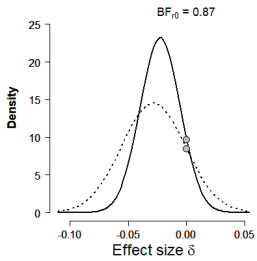
(f) Risk (+)



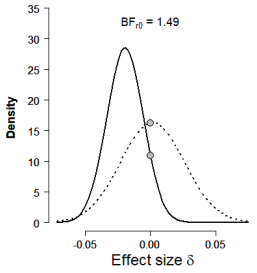
(g) Time All



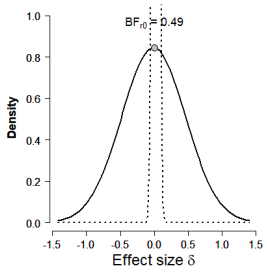
(h) Time (-)



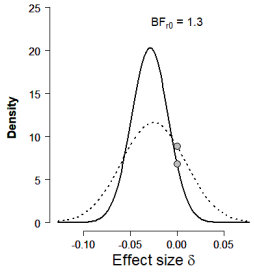
(i) Time (+)



(j) Anchor All

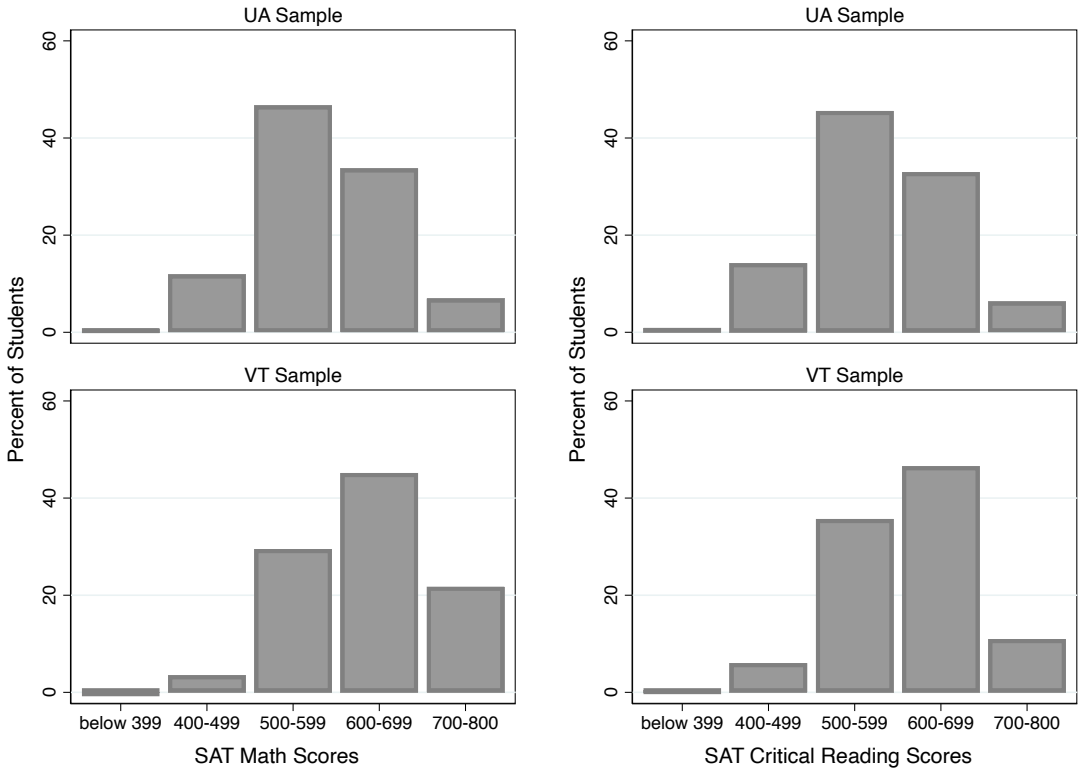


(k) Anchor (-)

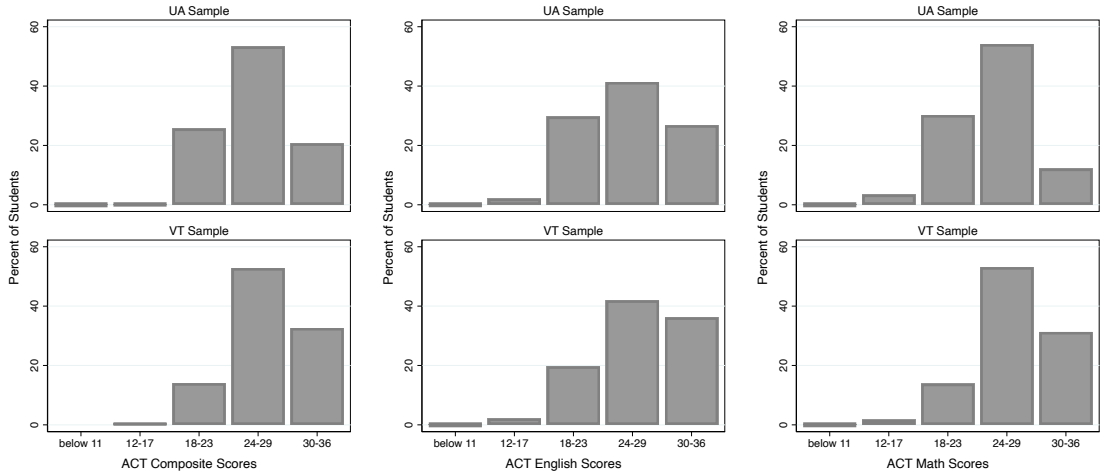


(l) Anchor (+)

Figure A3. Bayesian Replication Test Result. The dotted lines represent the posterior from the original experiment, which is used as prior for effect size in the replication tests. The solid lines represent the posterior distributions after the data from the replication attempt are considered. The gray dots indicate the ordinates of this prior and posterior at the skeptic's null hypothesis that the effect size is zero.



(a) SAT Scores



(b) ACT Scores

Figure A4. Average Standardized Examination Scores over 5 Years

References

Allred, S., Duffy, S., and Smith, J. (2016). Cognitive load and strategic sophistication. *Journal of Economic Behavior & Organization*, 125:162–178.

Benjamin, D. J., Brown, S. A., and Shapiro, J. M. (2013). Who is 'behavioral'? cognitive ability and anomalous preferences. *Journal of the European Economic Association*, 11(6):1231–1255.

Blaywais, R. and Rosenboim, M. (2019). The effect of cognitive load on economic decisions. *Managerial and Decision Economics*, 40(8):993–999.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Cárdenas, J. C., De Roux, N., Jaramillo, C. R., and Martinez, L. R. (2014). Is it my money or not? an experiment on risk aversion and the house-money effect. *Experimental Economics*, 17(1):47–60.

Carpenter, J., Graham, M., and Wolf, J. (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior*, 80:115–130.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., and Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, 32(3):277–296.

Dean, E. B., Schilbach, F., and Schofield, H. (2019). Poverty and cognitive function. In *The economics of poverty traps*, pages 57–118. University of Chicago Press.

Deck, C. and Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review*, 78:97–119.

Deck, C., Jahedi, S., and Sheremeta, R. (2021). On the consistency of cognitive load. *European Economic Review*, 134:103695.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review*, 100(3):1238–60.

Drichoutis, A. C. and Nayga, Jr., R. M. (2020). Economic rationality under cognitive load. *The Economic Journal*, 130(632):2382–2409.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42.

Hinson, J. M., Jameson, T. L., and Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(2):298.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kessler, J. B. and Meier, S. (2014). Learning from (failed) replications: Cognitive load manipulations and charitable giving. *Journal of Economic Behavior & Organization*, 102:10–13.

Khaw, M. W., Li, Z., and Woodford, M. (2021). Cognitive imprecision and small-stakes risk aversion. *The review of economic studies*, 88(4):1979–2013.

Maniadis, Z., Tufano, F., and List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, 104(1):277–290.

Maniadis, Z., Tufano, F., and List, J. A. (2017). To replicate or not to replicate? exploring reproducibility in economics through the lens of a model and a pilot study. *Economic Journal*, 127(605):F209–F235.

Matthews, G. and Campbell, S. E. (2010). Dynamic relationships between stress states and working memory. *Cognition and emotion*, 24(2):357–373.

Pennycook, G. and Ross, R. M. (2016). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, 7:9.

Ramsay, M. C. and Reynolds, C. R. (1995). Separate digits tests: A brief history, a literature review, and a reexamination of the factor structure of the test of memory and learning (tomal). *Neuropsychology review*, 5(3):151–171.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., and Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, 22(2):304.

Schmidt, S. (2009). Shall we really do it again? the powerful concept of replication is neglected in the social sciences. *Review of general psychology*, 13(2):90–100.

Shah, A. K., Mullainathan, S., and Shafir, E. (2012). Some consequences of having too little. *Science*, 338(6107):682–685.

Shiv, B. and Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of consumer Research*, 26(3):278–292.

Stupple, E., Gale, M., and Richmond, C. (2013). Working memory, cognitive miserliness and logic as predictors of performance on the cognitive reflection test. In *Proceedings of the Annual Meeting of the Cognitive Science society*, volume 35.

University of Arkansas (2019). Common data set initiative 2014-2019. Retrieved Jan 07, 2022 from https://oir.uark.edu/datasets/cds/index.php.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457.

Virginia Tech (2019). Common data set initiative 2014-2019. Retrieved Jan 07, 2022 from https://aie.vt.edu/strategic-analysis/common-data-set.html.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale (4th ed.)*. Pearson, San Antonio, TX.

Whitney, P., Rinehart, C. A., and Hinson, J. M. (2008). Framing effects under cognitive load: The role of working memory in risky decisions. *Psychonomic bulletin & review*, 15(6):1179–1184.