Appendix A.1

The following Appendix A.1 presents additional general information and explanations on the experiments. Appendix A.2 presents additional tables, figures and analyses of experiment #1 we refer to in the robustness checks. Appendix A.3 does the same for experiment #2. In Appendix A.4 we show the results of a lab replication of experiment #2 to justify our usage of Amazon MTurk.

A.1.1 Experiment duration statistics

Table 1: Duration statistics for the experimental phases in all experiments (in minutes)

Duration statistic (min:sec)	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Experiment #1								
Total	335	22:28	10:08	05:54	15:02	20:25	26:59	59:06
Instructions	335	10:56	07:28	01:39	06:20	08:55	12:34	52:03
Experimental stages	335	06:11	04:02	02:02	03:50	05:02	07:17	38:55
Post exp. questionnaire	335	05:20	03:19	01:15	03:23	04:39	06:19	36:33
Experiment #2 - Basic treatment								
Total	131	17:42	08:13	05:53	12:41	15:16	21:08	50:57
Instructions	131	07:38	05:12	02:18	04:17	05:51	09:14	30:18
Experimental stages	131	04:57	03:16	00:45	03:03	04:08	05:38	23:59
Post exp. questionnaire	131	05:07	02:43	01:22	03:23	04:14	05:53	21:13
Experiment #2 - ShowNewDebts treatment								
Total	135	18:47	08:18	04:57	13:02	16:59	22:46	52:03
Instructions	135	08:19	05:55	01:49	04:35	06:25	10:04	34:41
Experimental stages	135	05:22	02:47	01:41	03:29	04:40	06:08	20:48
Post exp. questionnaire	135	05:07	02:39	01:21	03:20	04:30	06:17	15:29
Experiment #2 - ShowSavedMoney treatment								
Total	138	20:36	09:06	08:26	13:55	19:23	25:30	58:08
Instructions	138	08:54	06:28	02:41	05:00	07:08	10:11	48:36
Experimental stages	138	05:51	03:00	01:53	03:44	05:14	06:47	17:27
Post exp. questionnaire	138	05:51	03:20	01:01	03:38	05:03	07:14	21:55

A.1.2 Details on the data collection

We recruit the participants for both our experiments from the crowd-sourcing platform Amazon Mechanical Turk (MTurk). Participants on MTurk (Turkers) are asked to solve individual Human Intelligence Tasks (HIT), which can then be approved or rejected by the requester of that HIT. Each of our experiments is one single HIT. We restrict participation to Turkers with at least 100 completed HITs to screen out throwaway accounts, bots and new Turkers, whom we expect to make more mistakes due to their unfamiliarity with MTurk. We require an approval rate on former HITs of at least 95%, a common threshold that was shown to ensure high data quality (Peer et al., 2014). No participant was allowed to take part in any of our experiments more than once.

To exclude bots, we ask our participants to describe the strategies they have used in the experiment in an open question. Two different researchers analyze if the answers are meaningful for that question. Both agree that this is the case for all our subjects in both experiments. We are therefore confident that our data does not contain any bot. We also include two attention check questions. In the first question, positioned in stage 1 right after the numeracy question, participants have to agree or disagree with the statement "All my friends are from outer space". Whoever agrees is screened out. The second question is included in the financial literacy questionnaire in stage 3. Subjects have to decide between choices we label "First answer" and "Second answer", where we ask them to select "Second answer". We screen out everyone who selects "First answer".

In experiment #1, 468 MTurkers started the experiment, of which 343 finished it. Out of the 125 who did not finish the experiment, 89 dropped out before the basic numeracy question, 27 did not pass the basic numeracy question, and 9 dropped out within the experiment or the post experimental questionnaire. Out of the remaining 343 participants, 335 passed the attention tests; these form our eventual sample.

Concerning experiment #2, 527 MTurkers started our experiment, of which 414 finished it. Out of the 113 who did not finish the experiment, 89 dropped out before the basic numeracy question, 36 did not pass the basic numeracy question, and 15 dropped out within the experiment or the post experimental questionnaire - 4 in each the Basic and the ShowNewDebts treatment, and 7 in the ShowSavedMoney treatment. Out of the remaining 414 participants, 404 passed both attention tests. These 404 participants form our sample. It should be noted that we recruit participants for the individual treatments in separate HITs on MTurk at the same time and using the same wording. As subjects cannot differentiate between the treatments, this should rule out selection effects.

A.1.3 Methodological Discussion

While evidence on methodological "standard objections" tends to be mixed and effects are often quite small (Camerer (2015); Camerer and Hogarth (1999); Dhami (2016); Zizzo (2010)), stake size might be a reasonable objection for our case, since high incentives can induce higher effort and may thereby improve performance in decision making and problem solving (Camerer and Hogarth, 1999). After all, repaying all the money on the low interest rate card reduces bonus payments by only 83 US-cents in our MTurk-based experiment #2. However, due to the change that no money can be left on the checking account, in the lab replication this difference is €10, and we find even more misallocation. The fact that stake size has rather the opposite effect in our study should hence be seen as an argument against this objection.

Another problem with such standard methodological objections is that they do not imply any predictable patterns in the data. We, in contrast, find distinct patterns. In order to explain these as artifacts would hence require specific methodological objections that allow to predict the Cuckoo Fallacy, the Complete Repayment fallacy, Equal Start effects *and* the 1/N strategy in experiment #1. A second example for strong structural effects that are hard to reconcile with methodological objections are the strong differences in the distributions of the chosen options in experiment #1, say between the Everything Equal and the control scenario for the 1/N Heuristic. Subjects that are not responding to our incentives or instructions should not behave that differently between scenarios as they do. In the same vein, we find that the effects of financial literacy on misallocation are stable and negative, which raises the question which objection allows to explain a constant relationship between financial literacy and the vulnerability to experimental effects. The argument that "subjects find the experiment too easy to be true", for instance, predicts a positive correlation between financial literacy and misallocation, since the more literate a person is, the easier the task should appear - but we find the opposite. Relationships between financial literacy and experimenter demand effects, scrutiny or other typical experimental artifacts are also not obvious, so again the question remains which methodological error, or combination of errors, could cause the observed patterns.

One specific experimenter demand effect might exist for some of the fallacies in experiment #1, however, if participants anticipate our hypothesis that fallacy scenarios cause misallocation. For fallacies such as the Cuckoo Fallacy or Complete Repayment, it might be easier in the fallacy scenario to predict which button we "want" our subjects to click than in the control scenario, which might explain why they use it more often. We tried to tackle this problem by ruling out that a fallacy scenario and its control scenario can occur directly after each other to muddle the contrasts somewhat, but ultimately we cannot exclude this explanation. However, we are not sure about the direction of that effect, because we have severe doubts that our participants systematically anticipate our hypothesis, and even if they do, that they are motivated to follow "our demands". Consider what kind of a situation this experimenter demand effect assumes: A subject correctly identifies a fallacy scenario as a trap, and then believes we "want" them to step into the trap, so they do. But we believe it is just as likely - if not more so - that in such a situation a participant thinks we "want" them to identify the trap and avoid it - which would lead to the exact opposite behavior. And to the degree that participant interpret setting up traps as a negative behavior by us, reciprocity might change their motive from "helping" us to "showing" us, which again predicts the opposite effect. In combination with standard objections against experimenter demand effects such as that they go against the incentive structure and that they are less severe in online experiments than in a lab experiment where we as researchers are physically present, we do not think this explanation works for the scenarios.

A final methodological objection that we want to raise is the strong effect size in our data. Set against some parts of the earlier literature, the number of subjects that make at least one non-optimal choice in our experiments is very high: 82% on MTurk and 91% in the control group and the lab replication in experiment #2. Keys et al. (2016), for instance, find that around 20% of US households who could refinance mortgages more cheaply did not, even though this task is clearly more complex than our experiments. Agarwal et al. (2015) show that in a natural experiment where consumers could acquire a credit card, roughly 40% chose the higher interest rate card. In Keys and Wang (2019), only up to 20% of credit card owners are influenced by anchoring due to minimum repayments. Our results appear less outlandish, however, when compared to the literature most closely related to our work. In experiment #1 of Amar et al. (2011) the misallocation is 41% or 49%, depending on the treatment, and virtually no participant finished their 25 rounds game without any misallocation. The two field studies that resemble our work most closely have similar results: While the share of misallocating people is not directly reported in Ponce et al. (2017), Gathergood et al. (2019) indicate that "85 percent of individuals should put 100 percent of their excess payments on the high interest rate card but only 10 percent do so". And the results in the field are even stronger than in our data if we refer to misallocation itself. In both Gathergood et al. (2019) and Ponce et al. (2017), the average observed misallocation is around 50%, while in our data it is around 30%. Even the highest values that we observe - around 44% in the ShowNewDebts treatment when the Cuckoo Fallacy is possible, and around 47% in the Cuckoo Fallacy scenario - are below 50%. In fact, we should have expected the effect to be smallest in a pure, but potentially unmeasurable state, modest in an experimental setting with some methodological problems or a design to provoke misallocation, and highest in the field, where most distractions and a selection effect with respect to the use of credit cards exist. Altogether, we hence admit that some or a combination of methodological objections might influence the effect size, or even specify interaction effects, but we believe that they cannot negate the existence of the reported effects.

A.1.4 Theorem to prove optimal repayment

The following theorem proves that repaying the high interest rate credit card is indeed the debt minimizing way of credit card repayment:

Let there be two credit cards x and y with start balances $b_x,\ b_y \in \mathbb{R}$ and interest rates $i_x > i_y > 0$, such that x is the high interest rate credit card. Let there be n repayment rounds following the procedure as proposed in Section **??**. Furthermore, let $a > 0$ be the money available every round on the checking account and let $r_k \in [0,\ 1]$ for $1 \leq k \leq n$ be the share of money that is repaid on the high interest rate credit card x. Consequently $1 - r_k$ is the share of money that is repaid on y. Then the overall debt after n rounds are minimized if and only if $r_k = 1\ \forall\ 1 \leq k \leq n$.

Proof: The overall balance f is the sum of the balances of the credit cards x and y after n rounds depending on the choices of r_k for $1 \leq k \leq n$. Thus, f is a function

$$[0,1]^n \to \mathbb{R} : \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} \mapsto b_x i_x^n + \sum_{k=1}^{n} \left(a r_k i_x^{n-k+1} \right) + b_y i_y^n + \sum_{k=1}^{n} \left(a(1-r_k) i_y^{n-k+1} \right)$$

Therefore

$$f \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} = b_x i_x^n + b_y i_y^n + a \cdot \sum_{k=1}^{n} \left(r_k i_x^{n-k+1} + (1-r_k) i_y^{n-k+1} \right)$$

and

$$Df = \begin{pmatrix} \frac{\partial f}{\partial r_1} \\ \vdots \\ \frac{\partial f}{\partial r_n} \end{pmatrix} = a \cdot \underbrace{\begin{pmatrix} i_x^n - i_y^n \\ \vdots \\ i_x^1 - i_y^1 \end{pmatrix}}_{>0,\text{ because of } i_x > i_y}.$$

The derivative of f is constant positive, therefore f is strictly increasing in all its components r_k. Thus, f takes on the absolute maximum if and only if $r_k = 1 \ \forall \ 1 \leq k \leq n$. The absolute maximum of the account balances corresponds to a minimum of the debt. Note that this proof also applies for positive account balances, meaning that you maximize your money by investing in an asset with the highest interest rate. □

Appendix A.2 - Experiment #1

A.2.1 Additional Tables and Figures

Table 2: Logistic regression model with random effects[a]

	Dependent variable: Choice of fallacy-implicated repayment option (1 = Chosen, 0 = Not chosen)						
	Cuckoo Fallacy	Equalize Balances	Complete Repayment	Balance Matching	1/N Heuristic	Interest Matching	Equal Start
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fallacy scenario	0.285***	0.015	0.172***	0.030	0.242***	-0.083*	0.216***
Standard error	(0.038)	(0.011)	(0.024)	(0.016)	(0.033)	(0.026)	(0.027)
p-value	[0.000]	[0.177]	[0.000]	[0.061]	[0.000]	[0.001]	[0.000]
Holm adjusted	[0.000]	[1.000]	[0.000]	[0.605]	[0.000]	[0.021]	[0.000]
Financial literacy	0.008	-0.012	-0.000	-0.007	-0.015	-0.015	-0.024
Standard error	(0.011)	(0.006)	(0.012)	(0.008)	(0.011)	(0.016)	(0.015)
p-value	[0.446]	[0.060]	[0.977]	[0.400]	[0.174]	[0.343]	[0.120]
Holm adjusted	[1.000]	[0.605]	[1.000]	[1.000]	[1.000]	[1.000]	[0.960]
Age	-0.004**	-0.001	-0.004**	-0.003*	0.002	-0.001	0.001
Standard error	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
Dummy: Male	-0.016	-0.024	-0.030	-0.030	-0.004	-0.011	-0.021
Standard error	(0.028)	(0.018)	(0.032)	(0.023)	(0.027)	(0.043)	(0.040)
Years of education (yoe)	-0.007	-0.004	-0.010	-0.003	-0.006	0.002	-0.030***
Standard error	(0.006)	(0.004)	(0.007)	(0.005)	(0.006)	(0.010)	(0.009)
Observations	670	670	670	670	670	670	670

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$ for the Holm-adjusted p-values

[a] Reported coefficients are margins. The seven models denote the seven scenario pairs, the differences of control- and fallacy scenario are denoted in the Fallacy scenario coefficients. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for 28 coefficients from two tables: The seven fallacy scenario coefficients from Table 2, the seven fallacy scenario coefficients from Table 3, and the 14 financial literacy coefficients from both tables. Asterisks indicate significance after adjustment.

Table 3: Logistic regression model with random effects[a]

	Dependent variable: Choice of optimal repayment option (1 = Chosen, 0 = Not chosen)						
	Cuckoo Fallacy	Equalize Balances	Complete Repayment	Balance Matching	1/N Heuristic	Interest Matching	Equal Start
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fallacy scenario	-0.104**	-0.009	-0.071	-0.057	-0.353***	0.077*	-0.086*
Standard error	(0.028)	(0.026)	(0.025)	(0.027)	(0.022)	(0.026)	(0.027)
p-value	[0.000]	[0.733]	[0.005]	[0.035]	[0.000]	[0.003]	[0.002]
Holm adjusted	[0.004]	[1.000]	[0.056]	[0.389]	[0.000]	[0.045]	[0.024]
Financial literacy	0.062**	0.079***	0.063**	0.062**	0.071***	0.073***	0.057*
Standard error	(0.017)	(0.016)	(0.017)	(0.018)	(0.016)	(0.017)	(0.018)
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]
Holm adjusted	[0.004]	[0.000]	[0.005]	[0.008]	[0.000]	[0.000]	[0.020]
Age	-0.000	0.005*	0.004	0.005*	0.001	0.004	0.002
Standard error	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Dummy: Male	0.088	0.042	0.089	0.031	0.030	0.050	0.093*
Standard error	(0.045)	(0.044)	(0.047)	(0.048)	(0.042)	(0.047)	(0.047)
Years of education (yoe)	0.001	0.007	0.008	0.009	0.006	0.002	0.014
Standard error	(0.011)	(0.010)	(0.011)	(0.011)	(0.009)	(0.011)	(0.010)
Observations	670	670	670	670	670	670	670

Note: $^{*}p<0.05$;$^{**}p<0.01$;$^{***}p<0.001$ for the Holm-adjusted p-values

[a] Reported coefficients are margins. The seven models denote the seven scenario pairs, the differences of control- and fallacy scenario are denoted in the Fallacy scenario coefficients. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for 28 coefficients from two tables: The seven fallacy scenario coefficients from Table 2, the seven fallacy scenario coefficients from Table 3, and the 14 financial literacy coefficients from both tables. Asterisks indicate significance after adjustment.

Table 4: Number of choices for each repayment option in each scenario.

	All on low	2:1	1:1	1:2	All on high	∅ MA
Scenario 01: Cuckoo Fallacy, Control	12	13	67	95	148	0.26
Scenario 02: Cuckoo Fallacy, Treatment	97	42	38	45	113	0.47
Scenario 03: Equalize Balances, Control	8	16	15	67	229	0.14
Scenario 04: Equalize Balances, Treatment	13	19	14	63	226	0.16
Scenario 05: Complete Repayment, Control	24	20	50	97	144	0.28
Scenario 06: Complete Repayment, Treatment	81	17	33	84	120	0.41
Scenario 07: Balance Matching, Control	8	17	20	92	198	0.18
Scenario 08: Balance Matching, Treatment	20	27	24	85	179	0.23
Scenario 09: 1/N Heuristic, Control	19	15	10	32	259	0.13
Scenario 10: 1/N Heuristic, Treatment	10	20	82	93	130	0.28
Scenario 11: Interest Matching, Control	10	19	14	104	188	0.19
Scenario 12: Interest Matching, Treatment	13	10	22	76	214	0.17
Scenario 13: Equal Start, Control	14	18	66	90	147	0.27
Scenario 14: Equal Start, Treatment	10	9	140	58	118	0.31
Scenario 15: Everything Equal	9	8	274	5	39	-

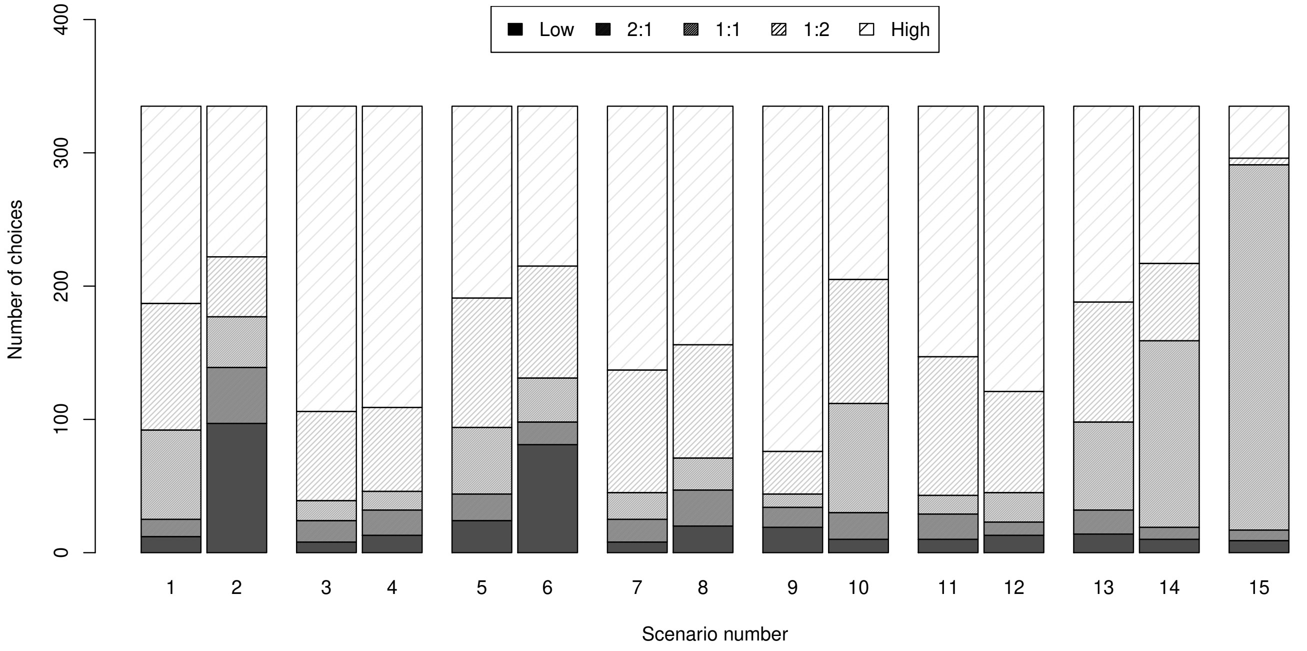

Figure 1: Relative proportion of choices in the scenarios

Table 5: Multinomial Regression analysis[a]

	Dependent variable: Chosen option						
	Cuckoo Fallacy	Equalize Balances	Complete Repayment	Balance Matching	1/N Heuristic	Interest Matching	Equal Start
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fallacy scenario							
Option 1	0.271***	0.015	0.176***	0.037*	-0.024	0.009	-0.011
Standard error	(0.035)	(0.014)	(0.030)	(0.017)	(0.013)	(0.014)	(0.014)
p-value	[0.000]	[0.276]	[0.000]	[-]	[-]	[-]	[-]
Holm adjusted	[0.000]	[1.000]	[0.000]	[-]	[-]	[-]	[-]
Option 2	0.071**	0.009	-0.009	0.030	0.009	-0.027	-0.026
Standard error	(0.022)	(0.017)	(0.016)	(0.019)	(0.014)	(0.016)	(0.015)
p-value	[-]	[-]	[-]	[0.127]	[-]	[-]	[-]
Holm adjusted	[-]	[-]	[-]	[0.975]	[-]	[-]	[-]
Option 3	-0.083***	-0.003	-0.051*	0.012	0.228***	0.024	0.215***
Standard error	(0.024)	(0.015)	(0.024)	(0.019)	(0.035)	(0.017)	(0.031)
p-value	[-]	[-]	[-]	[-]	[0.000]	[-]	[0.000]
Holm adjusted	[-]	[-]	[-]	[-]	[0.000]	[-]	[0.000]
Option 4	-0.143***	-0.012	-0.041	-0.021	0.143***	-0.083	-0.093**
Standard error	(0.028)	(0.030)	(0.033)	(0.034)	(0.029)	(0.034)	(0.030)
p-value	[-]	[-]	[-]	[-]	[-]	[0.014]	[-]
Holm adjusted	[-]	[-]	[-]	[-]	[-]	[0.210]	[-]
Option 5	-0.116*	-0.009	-0.075	-0.057	-0.355***	0.077	-0.084
Standard error	(0.034)	(0.034)	(0.036)	(0.037)	(0.026)	(0.036)	(0.036)
p-value	[0.001]	[0.792]	[0.036]	[0.122]	[0.000]	[0.034]	[0.018]
Holm adjusted	[0.010]	[1.000]	[0.400]	[0.975]	[0.000]	[0.405]	[0.246]
Financial literacy							
Option 1	0.006	-0.012	-0.003	-0.020***	-0.013*	-0.013*	-0.003
Standard error	(0.011)	(0.006)	(0.011)	(0.006)	(0.006)	(0.005)	(0.007)
p-value	[0.608]	[0.031]	[0.796]	[-]	[-]	[-]	[-]
Holm adjusted	[1.000]	[0.406]	[1.000]	[-]	[-]	[-]	[-]
Option 2	-0.013	-0.021***	-0.040***	-0.007	-0.028***	-0.016*	-0.023***
Standard error	(0.009)	(0.006)	(0.007)	(0.009)	(0.007)	(0.007)	(0.007)
p-value	[-]	[-]	[-]	[0.389]	[-]	[-]	[-]
Holm adjusted	[-]	[-]	[-]	[1.000]	[-]	[-]	[-]
Option 3	-0.045***	-0.023***	-0.023*	-0.018*	-0.017	-0.019***	-0.028
Standard error	(0.011)	(0.006)	(0.009)	(0.007)	(0.010)	(0.006)	(0.014)
p-value	[-]	[-]	[-]	[-]	[0.093]	[-]	[0.042]
Holm adjusted	[-]	[-]	[-]	[-]	[0.839]	[-]	[0.417]
Option 4	-0.010	-0.020	0.004	-0.016	-0.008	-0.020	-0.003
Standard error	(0.012)	(0.012)	(0.013)	(0.014)	(0.012)	(0.014)	(0.013)
p-value	[-]	[-]	[-]	[-]	[-]	[0.139]	[-]
Holm adjusted	[-]	[-]	[-]	[-]	[-]	[0.975]	[-]
Option 5	0.062**	0.076***	0.062**	0.061**	0.067***	0.069***	0.057**
Standard error	(0.015)	(0.013)	(0.015)	(0.015)	(0.014)	(0.014)	(0.015)
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.001]	[0.000]	[0.001]	[0.001]	[0.000]	[0.000]	[0.002]
Observations	670	670	670	670	670	670	670
Further control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: $^{*}p<0.05$;** $p<0.01$;*** $p<0.001$ for the Holm-adjusted p-values

[a] Reported coefficients are margins and denote the estimated differences in the probability that a certain option is chosen (rows) for all seven scenarios (columns). The first block "Fallacy scenario" shows the average differences in percentage of chosen options when switching from the control to the corresponding fallacy scenario. The second block "Financial literacy" shows how each correctly answered financial literacy question changes the probability to choose a certain option. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values for the variables we interpret in brackets. The p-values are adjusted to include all the 28 coefficients for which we present adjusted p-values. Asterisks indicate significance after adjustment.

Table 6: OLS of the changes between control and fallacy scenario[a]

	Dependent variable: Option change between control and fallacy scenario						
	Cuckoo Fallacy (All low)	Equalize Balances (All low)	Complete Repayment (All low)	Balance Matching (2:1)	1/N Heuristic (1:1)	Interest Matching (1:2)	Equal Start (1:1)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	-0.952***	-0.069	-0.513***	-0.236***	-0.549***	0.081	-0.218**
Standard error	(0.102)	(0.054)	(0.086)	(0.061)	(0.066)	(0.053)	(0.064)
p-value	[0.000]	[0.207]	[0.000]	[0.000]	[0.000]	[0.126]	[0.001]
Holm adjusted	[0.000]	[0.252]	[0.000]	[0.001]	[0.000]	[0.252]	[0.002]
Observations	335	335	335	335	335	335	335
Further control variables	No	No	No	No	No	No	No
R^2	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$ for the Holm-adjusted p-values

[a] The constant estimates the mean number of options a participant changes between the two scenario types, negative values implicate a change away from the optimal option. The seven models denote the seven scenario pairs. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for all 7 coefficients. Asterisks indicate significance after adjustment.

Table 7: Logistic regression model with random effects (including screened out participants)[a]

	Dependent variable: Choice of fallacy-implicated repayment option (1 = Chosen, 0 = Not chosen)							
	Cuckoo Fallacy	Cuckoo Fallacy	Equalize Balances	Complete Repayment	Balance Matching	1/N Heuristic	Interest Matching	Equal Start
	(1.1)	(1.2)	(2)	(3)	(4)	(5)	(6)	(7)
Fallacy scenario	0.273	0.280^{***}	0.018	0.168^{***}	0.029	0.223^{***}	-0.078^{*}	0.219^{***}
Standard error	(2.584)	(0.037)	(0.011)	(0.024)	(0.017)	(0.031)	(0.026)	(0.026)
p-value	[0.916]	[0.000]	[0.119]	[0.000]	[0.079]	[0.000]	[0.002]	[0.000]
Holm adjusted	[see caption]	[0.000]	[0.835]	[0.000]	[0.713]	[0.000]	[0.032]	[0.000]
Financial literacy	0.009	0.010	-0.015	-0.000	-0.008	-0.018	-0.014	-0.019
Standard error	(0.124)	(0.011)	(0.006)	(0.012)	(0.008)	(0.011)	(0.016)	(0.015)
p-value	[0.943]	[0.348]	[0.021]	[0.978]	[0.338]	[0.085]	[0.383]	[0.206]
Holm adjusted	[see caption]	[1.000]	[0.233]	[1.000]	[1.000]	[0.713]	[1.000]	[1.000]
Observations	686	684	686	686	686	686	686	686
Further control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$ for the Holm-adjusted p-values

[a] Reported coefficients are margins. The seven models denote the seven scenario pairs, the differences of control- and fallacy scenario are denoted in the Fallacy scenario coefficients. The first column shows a model where we believe there was a technical error in the algorithm that prevented the calculations of the standard errors for the Cuckoo Fallacy from succeeding properly. Only when we include a particular person *and* use robust standard errors *and* include the age as a control variable *and* report the margins, we get a standard error of 2.494. If any of these conditions is not fulfilled, the standard error decreases by a factor of around 65. The particular participant shows no anomalies (e.g., she is 28 years old). We do not think this is a "legit" standard error but a problem of the margin calculation (else we would see the problem in the logit calculation as well, but we do not), so we solved the problem by screening out the problematic participant. The results are in column (1.2). For completeness, we still report the erroneous calculation in column (1.1), but do ignore this model for the Bonferroni-Holm correction. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for 28 coefficients from two tables: The seven "legit" fallacy scenario coefficients from Table 7, the seven fallacy scenario coefficients from Table 8, and the 14 financial literacy coefficients from both tables. Asterisks indicate significance after adjustment.

Table 8: Logistic regression model with random effects (including screened out participants)[a]

	Dependent variable: Choice of optimal repayment option (1 = Chosen, 0 = Not chosen)						
	Cuckoo Fallacy	Equalize Balances	Complete Repayment	Balance Matching	1/N Heuristic	Interest Matching	Equal Start
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fallacy scenario	-0.104**	-0.012	-0.067	-0.058	-0.346***	0.076	-0.090*
Standard error	(0.027)	(0.026)	(0.025)	(0.026)	(0.022)	(0.026)	(0.027)
p-value	[0.000]	[0.651]	[0.008]	[0.028]	[0.000]	[0.004]	[0.001]
Holm adjusted	[0.002]	[1.000]	[0.096]	[0.276]	[0.000]	[0.051]	[0.013]
Financial literacy	0.065**	0.081***	0.066**	0.064**	0.074***	0.074***	0.063**
Standard error	(0.016)	(0.015)	(0.017)	(0.017)	(0.016)	(0.017)	(0.017)
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.001]	[0.000]	[0.001]	[0.003]	[0.000]	[0.000]	[0.003]
Observations	686	686	686	686	686	686	686
Further control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: *p<0.05; **p<0.01; ***p<0.001 for the Holm-adjusted p-values

[a] Reported coefficients are margins. The seven models denote the seven scenario pairs, the differences of control- and fallacy scenario are denoted in the Fallacy scenario coefficients. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for 28 coefficients from two tables: The seven "legit" fallacy scenario coefficients from Table 7, the seven fallacy scenario coefficients from Table 8, and the 14 financial literacy coefficients from both tables. Asterisks indicate significance after adjustment.

A.2.2 Exploratory within-subject analyses

Our analysis of the hypothesized fallacies so far treated each individual's decisions independently. However, the within-subject behavior of the participants over the seven fallacies might be interesting in its own - and deliver further insights on decision making processes. We therefore run additional exploratory analyses where we trace each participant's decisions throughout the experiment and compare decisions among participants and across scenarios. We start by counting the optimal answers of each participant and report the results in Figure 2. 60 out of 335 participants (about 17.9%) always chose the optimal option 5, i.e. gave 14 optimal answers. On the other hand, 19 participants (about 5.7%) never chose option 5.

The transition matrices for each fallacy in Table 9 confirm the results from a within-subjects perspective. Each cell in these matrices gives the proportion of participants that switch (or do not switch, in the cells on the main diagonal) from one option in the control scenarios to another in the fallacy scenario. Indeed, many participants switch between the five options comparing control and fallacy scenarios in all seven scenario pairs. However, the table also indicates that for the Cuckoo Fallacy, Complete Repayment, 1/N and Equal Start, more participants switch from any other option in the control scenario to the fallacy-implicated option in the fallacy scenario than the other way around.

This impression is supported by Table 10, where we calculate the proportion of optimal answers (panel (a)) and of fallacy-implicated answers (panel (b)) over all participants. We show the corresponding results for combinations of control (rows) and fallacy (column) scenarios. In panel (a), participants below the main diagonal (in grey) give more optimal answers in the control than in the fallacy scenarios. There are, for example, 18 participants (5.37%) who chose the optimal option twice in the control scenarios but only once in the fallacy scenarios, while there are only 9 people (2.69%) who show the opposite behavior. In line with our main findings, a Wilcoxon rank sum test of differences between optimal answers in control and fallacy scenarios reveals that participants indeed answered more optimally in the control scenarios (p-value: $2.2 \cdot 10^{-16}$). Panel (b) shows the corresponding proportions for the fallacy-implicated options instead. We expect more fallacy-implicated answers in the fallacy scenarios, i.e. higher proportions displayed above the main diagonal than below. This is confirmed by another Wilcoxon test (p-value: $2.2 \cdot 10^{-16}$).

A visualization of Table 10 is given in Figure 3, where each point represents one participant. We adapt the axes to the table, such that the x-axis denotes the count for the fallacy scenarios and the downwards directed y-axis denotes the count for the control scenarios.

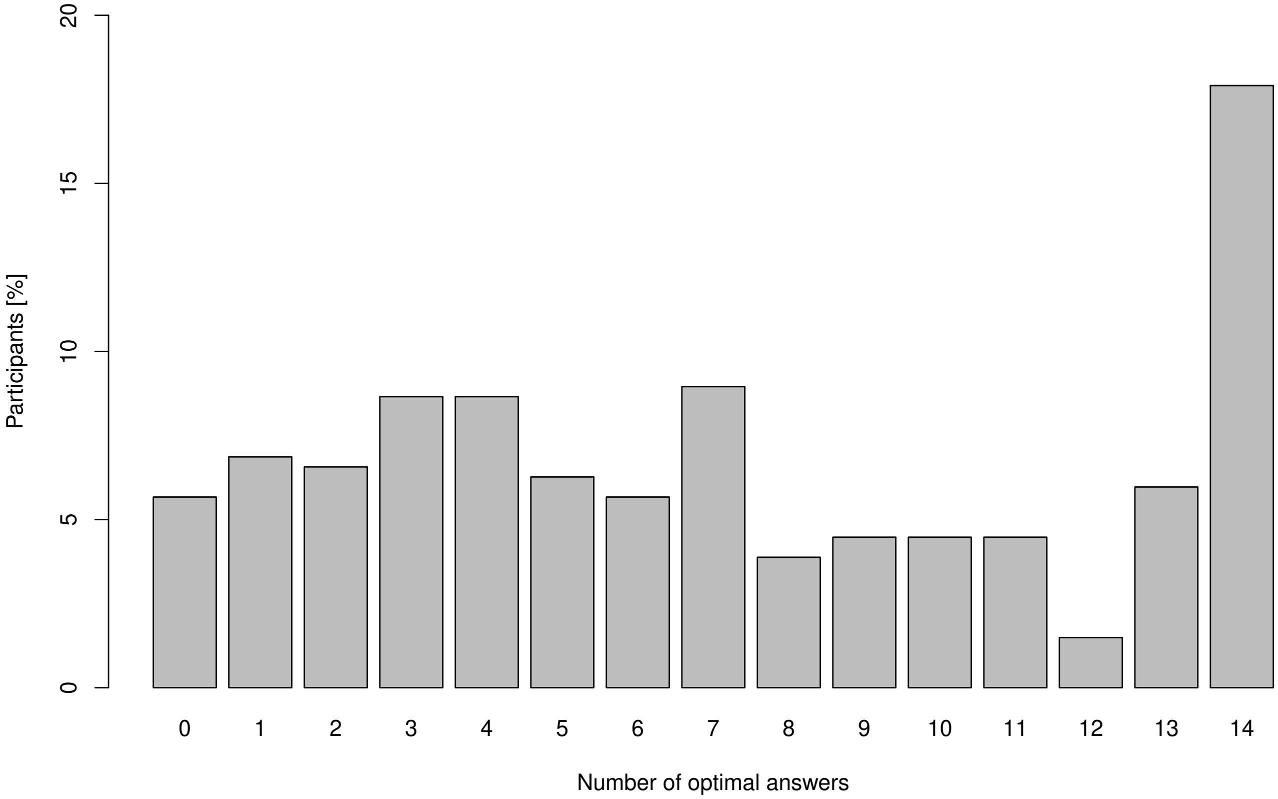

Figure 2: The bars show which proportion of participants gave a certain amount of optimal answers (option 5) in the 14 scenarios (excluding the Everything Equal scenario).

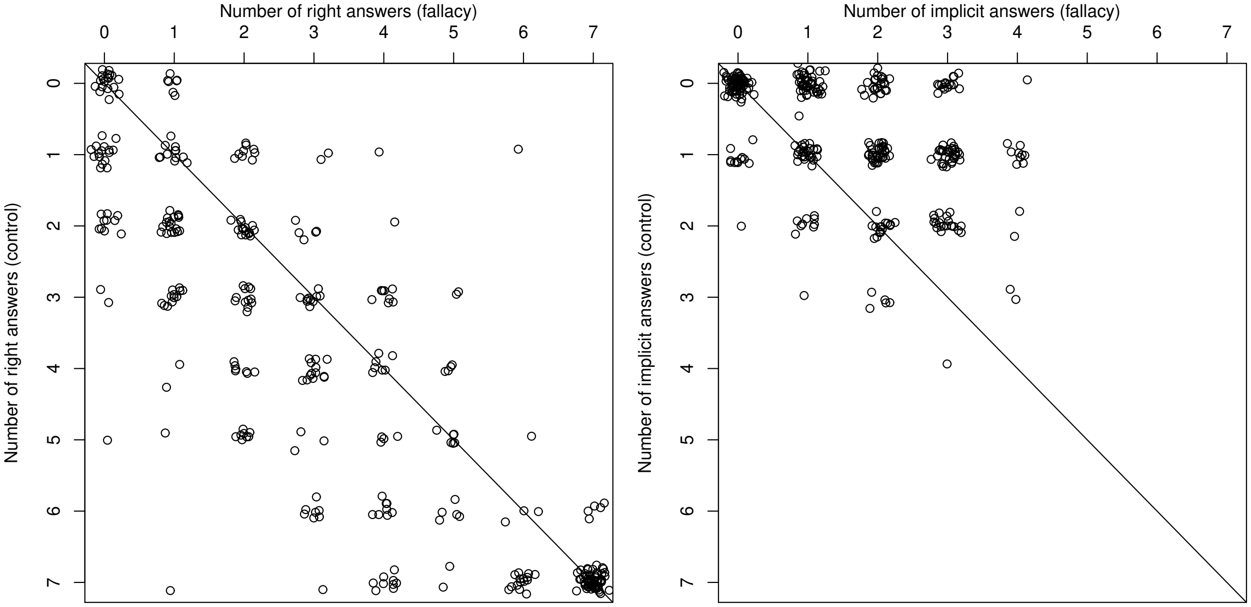

Figure 3: This graph visualizes how many participants give certain answers in control (y-axis) vs fallacy scenarios (x-axis). The left graphic shows the number of optimal answers (option 5), the right graphic shows the number of fallacy-implicated answers.

Table 9: Transition matrices between control and fallacy scenarios[a]

Fallacy	Cuckoo Fallacy					Equalize Balances				
Control	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	1.19%	0.00%	0.60%	0.00%	1.79%	1.19%	0.30%	0.00%	0.60%	0.30%
2	0.90%	0.30%	1.19%	0.30%	1.19%	0.90%	1.19%	0.90%	0.30%	1.49%
3	5.07%	3.28%	4.48%	4.18%	2.99%	0.00%	0.90%	1.19%	1.49%	0.90%
4	8.36%	6.87%	3.88%	6.57%	2.69%	0.90%	1.49%	1.49%	8.06%	56.42%
5	13.43%	2.09%	1.19%	2.39%	25.07%	0.90%	1.49%	1.49%	8.06%	56.42%
Fallacy	**Complete Repayment**					**Balance Matching**				
Control	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	5.07%	0.60%	0.60%	0.30%	0.60%	1.19%	0.30%	0.30%	0.30%	0.30%
2	2.09%	1.49%	0.30%	1.49%	0.60%	0.30%	2.39%	0.30%	1.19%	0.90%
3	3.28%	1.19%	3.88%	5.37%	1.19%	0.90%	0.60%	1.79%	2.09%	0.60%
4	4.18%	0.90%	4.48%	14.33%	5.07%	1.79%	2.99%	2.99%	11.94%	7.76%
5	9.55%	0.90%	0.60%	3.58%	28.36%	1.79%	1.79%	1.79%	9.85%	43.88%
Fallacy	**1/N Heuristic**					**Interest Matching**				
Control	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	1.19%	1.49%	1.79%	0.60%	0.60%	1.19%	0.60%	0.30%	0.90%	0.00%
2	0.00%	1.19%	1.19%	1.49%	0.60%	0.60%	0.60%	1.79%	1.19%	1.49%
3	0.00%	0.90%	1.19%	0.60%	0.30%	0.60%	0.60%	1.49%	0.30%	1.19%
4	0.30%	0.60%	3.28%	4.18%	1.19%	0.30%	0.90%	1.49%	14.93%	13.43%
5	1.49%	1.79%	17.01%	20.90%	36.12%	1.19%	0.30%	1.49%	5.37%	47.76%
Fallacy	**Equal Start**									
Control	**1**	**2**	**3**	**4**	**5**					
1	0.60%	0.30%	0.90%	0.60%	1.79%					
2	1.19%	0.30%	1.79%	1.79%	0.30%					
3	0.60%	1.19%	14.33%	1.79%	1.79%					
4	0.60%	0.30%	14.03%	6.87%	5.07%					
5	0.00%	0.60%	10.75%	6.27%	26.27%					

[a] This table shows the proportion of participants that switch from a certain option in the control scenario (rows) to a certain option in the fallacy scenario (columns) for all seven scenario pairs. Grey cells mark fallacy-implicated options and the participants switching to these option in the fallacy scenarios.

Table 10: Proportion of optimal or fallacy-implicated answers in control- and fallacy scenarios

(a) A participant can choose up to 7 times the optimal solution in the control scenarios (rows), and up to 7 times in the fallacy scenarios (columns). The table shows the proportion of each of these 7 × 7 possible combinations.

Frequency of optimal answers in fallacy scenarios								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	5.67%	2.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1	4.78%	3.58%	2.69%	0.60%	0.30%	0.00%	0.30%	0.00%
2	2.99%	5.37%	4.48%	1.49%	0.30%	0.00%	0.00%	0.00%
3	0.60%	3.58%	3.58%	2.99%	2.39%	0.60%	0.00%	0.00%
4	0.00%	0.60%	2.09%	3.88%	2.09%	1.19%	0.00%	0.00%
5	0.30%	0.30%	2.39%	0.90%	1.19%	1.79%	0.30%	0.00%
6	0.00%	0.00%	0.00%	2.09%	2.39%	1.49%	0.90%	1.49%
7	0.00%	0.30%	0.00%	0.30%	2.69%	0.60%	4.48%	17.91%
Wilcoxon rank sum test of differences in optimal answers control vs treatment: p-value $< 2.2 \cdot 10^{-16}$								

(b) A participant can choose up to 7 times the fallacy implicated solution in the control scenarios (rows), and up to 7 times in the fallacy scenarios (columns). The table shows the proportion of each of these 7 × 7 possible combinations.

Frequency of fallacy-implicated answers in fallacy scenarios								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	21.19%	14.33%	7.46%	4.78%	0.30%	0.00%	0.00%	0.00%
1	3.28%	8.06%	11.34%	9.55%	2.69%	0.00%	0.00%	0.00%
2	0.30%	2.69%	4.78%	5.97%	0.60%	0.00%	0.00%	0.00%
3	0.00%	0.30%	1.49%	0.00%	0.60%	0.00%	0.00%	0.00%
4	0.00%	0.00%	0.00%	0.30%	0.00%	0.00%	0.00%	0.00%
5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
6	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Wilcoxon rank sum test of differences in fallacy-implicated answers control vs treatment: p-value $< 2.2 \cdot 10^{-16}$								

We also investigate if the fallacies are correlated. Table 11 shows the correlation matrix between choices of fallacy-implicated options in the fallacy scenarios, i.e. whether participants who choose a fallacy-implicated option for one particular heuristic tend to also choose the fallacy-implicated option in other fallacy scenarios. For most comparisons we cannot detect any significant dependencies between the fallacies. Only four correlations are significantly positive (between 1/N Heuristic, Interest Matching and Equal Start as well as between Cuckoo Fallacy and Balance Matching), but they are not particularly large (below 0.3). Thus, we cannot confirm clear linear dependencies between the fallacies.

Table 11: Correlation matrix of fallacy-implicated answers in the fallacy scenarios

Correlation	**Cuckoo Fallacy**	**Equalize Balances**	**Complete Repayment**	**Balance Matching**	**1/N Heuristic**	**Interest Matching**	**Equal Start**
Cuckoo Fallacy	1	0.076	-0.068	0.125^{*}	-0.042	-0.016	0.060
Equalize Balances	0.076	1	0.103	0.054	-0.042	0.002	0.049
Complete Repayment	-0.068	0.103	1	0.063	-0.013	-0.040	-0.012
Balance Matching	0.125^{*}	0.054	0.063	1	-0.041	0.023	0.038
1/N Heuristic	-0.042	-0.042	-0.013	-0.041	1	0.156^{**}	0.264^{***}
Interest Matching	-0.016	0.002	-0.040	0.023	0.156^{**}	1	0.220^{***}
Equal Start	0.060	0.049	-0.012	0.038	0.264^{***}	0.220^{***}	1

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

In the next analysis we aim to identify groups of participants with similar answers to investigate whether the results are driven by a particular sub-population. We start by checking for each control scenario whether a participant selects the optimal option, the fallacy-implicated option, or any other non-optimal option combined. We then identify to which of these three possibilities the participant switches to (or stays) in the respective fallacy scenario. This allows us to identify nine distinct "types" of participants, e.g. participants who repay optimally in the control scenario and in the fallacy scenario (type ’optimal->optimal’), or participants who switch from optimal repayment in the control scenario to the fallacy-implicated option in the fallacy scenario (type ’optimal->implic’). In the next step, we count for each heuristic how many participants belong to a specific type and present the results as proportions in Table 12. We compare the number of participants for the ’implic->optimal’ type and the ’optimal->implic’ type with binomial tests and report the p-values in the table. With the exception of Equalize Balances and Balance Matching, we detect large differences between the proportions of participants switching from optimal to fallacy-implicated option (row 4) and of participants that exhibit the reversed behavior (row 2) in all scenario pairs, which is in line with the differences we report in our main analyses (Interest Matching again shows the reversed sign, as it is the only pair where the value in row 4 is larger than in line 2).[1]

Table 12: Behavior in the scenario pairs[a]

Behavior	**Cuckoo Fallacy**	**Equalize Balances**	**Complete Repayment**	**Balance Matching**	**1/N Heuristic**	**Interest Matching**	**Equal Start**
implicated->implicated	1.19%	1.19%	5.07%	2.39%	1.19%	14.93%	14.33%
implicated->optimal	1.79%	0.30%	0.60%	0.90%	0.30%	13.43%	1.79%
implicated->other	0.60%	0.90%	1.49%	1.79%	1.49%	2.69%	3.58%
optimal->implicated	13.43%	0.90%	9.55%	1.79%	17.01%	5.37%	10.75%
optimal->optimal	25.07%	56.42%	28.36%	43.88%	36.12%	47.76%	26.27%
optimal->other	5.67%	11.04%	5.07%	13.43%	24.18%	2.99%	6.87%
other->implicated	14.33%	1.79%	9.55%	3.88%	6.27%	2.39%	16.72%
other->optimal	6.87%	10.75%	6.87%	8.66%	2.39%	2.69%	7.16%
other->other	31.04%	16.72%	33.43%	23.28%	11.04%	7.76%	12.54%
p-value	$\mathbf{3.1 \cdot 10^{-8}}$	0.616	$\mathbf{3.31 \cdot 10^{-7}}$	0.502	$\mathbf{4.15 \cdot 10^{-14}}$	$\mathbf{5.79 \cdot 10^{-4}}$	$\mathbf{3.80 \cdot 10^{-6}}$

[a] This table shows the proportion of participants exhibiting a certain behavior between control and fallacy scenario of a scenario pair. The behavior in the control scenario is denoted on the left before the '->', the behavior in the respective fallacy scenario is denoted on the right after the '->', where 'implicated' means fallacy-implicated option, 'optimal' the optimal option (5) and 'other' every other option. The p-values refer to a binomial test of differences between the numbers of 'implicated->optimal' and 'optimal->implicated' participants for each scenario pair. We report significant p-values below 0.05 in bold to show that the numbers of participants switching from the optimal to the fallacy-implicated option differs from the participants switching the other way round.

In a final step, we use these nine types of participants to identify groups by employing a k-means cluster analysis. The cluster analysis helps us to identify a hypothetical "average" participant per group and use the information from the cluster to describe their typical decision making more closely. We use the elbow criterion (Thorndike, 1953), Akaike's information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz, 1978) to determine the number of clusters. All three criteria are visualized in Figure 4 for numbers of clusters between 1 and 30 (x-axis) and lead us to a choice of four clusters. To stabilize the clusters, we run the k-means algorithm 1000 times with different random starting values. We report the cluster centers of the four clusters in Table 13. Each cluster center stands for the average participant in the respective cluster. The numbers in the cells denote in how many out of seven scenario pairs the average participant exhibits behavior of the respective type. For each column, we print the maximum number in bold as it drives the assignment to this cluster the most.

Analyzing the four clusters, the clearest assignment is to cluster 3 with 88 out of 335 participants. This

[1] Note that while we present switches in the direction from control to fallacy scenarios, the participants might also have answered the fallacy scenario first, depending on the random order of the scenarios.

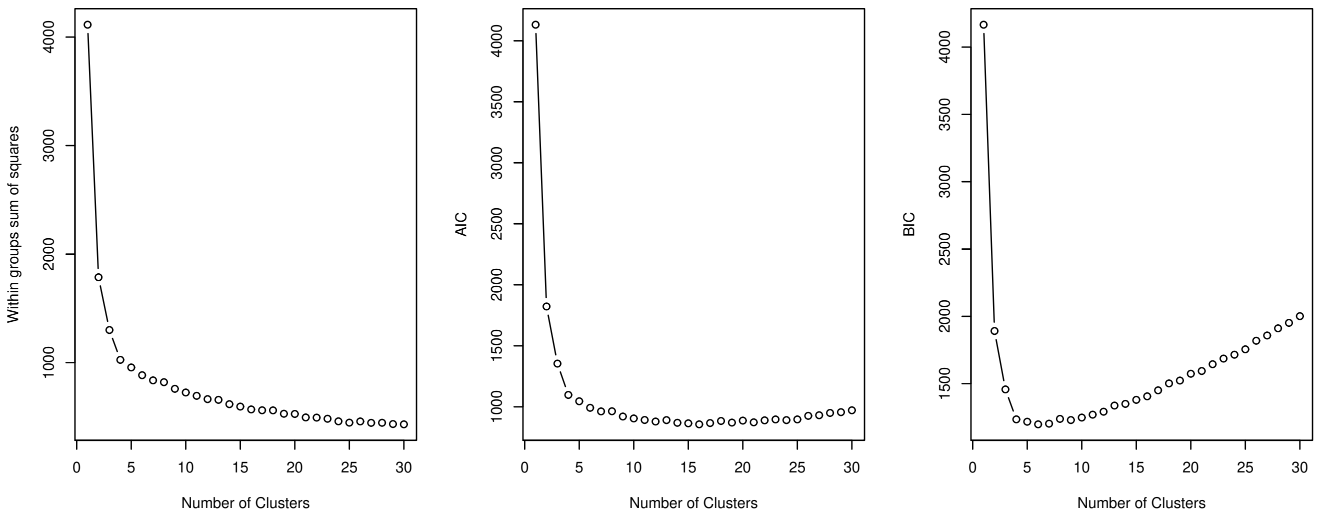

Figure 4: This Figure shows three criteria to determine the number of clusters for a k-means clustering. The x-axes show the number of clusters between 0 and 30. The y-axes show the value of within groups sum of squares (elbow criterion, left figure), the AIC (Akaike's information criterion, middle figure) or the BIC (Bayesian information criterion, right Figure). Considering all three criteria we determined four as an appropriate number of clusters.

cluster contains the optimally choosing participants. They choose the optimal option in the control scenarios in 6.77 (= $0.23+6.59+0.05$) out of seven control scenarios on average, and 6.69 times (= $0.02+6.59+0.08$) in the fallacy scenarios. They keep the optimal answer in 6.59 scenario pairs, and tend to correct the few errors they make. Out of the 0.13 times (= $0.01+0.02+0.00+0.00+0.08+0.02$) they chose any non-optimal option in the control scenarios, they correct this error in 0.1 (= 0.02 + 0.08) times in the fallacy scenario.

In contrast, the 82 participants assigned to cluster 1 seem to have a relatively good grasp on how to repay debts, but are vulnerable to fallacies. One average, they choose the optimal answer 4.81 times in the control scenarios, but only 3.65 times in the fallacy scenarios. They keep the optimal choice, provided they found it in the control scenarios, in only 2.94 times in the fallacy scenarios. Instead, they switch from the optimal to the fallacy-implicated answer in 0.82 times, and to any of the other three option in 1.05 times. They sometimes correct errors from the control scenario in the fallacy scenario (0.71 times in total), but these corrections do not offset the losses. On the other hand, they choose the fallacy-implicated option in only 0.63 control scenarios, but in 1.62 fallacy scenarios.

Cluster 4 (90 participants) seems to be similar to cluster 1, but with a much more erratic behavior, and starting from a lower level of optimality. Its participants choose the optimal option more often in the control scenarios than in the fallacy scenarios (2.64 to 1.9 times) too, and they show a vulnerability to getting

Table 13: Description of cluster means[a]

Behavior	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
implicated->implicated	0.37	0.77	0.01	0.51
implicated->optimal	0.21	0.16	0.02	0.37
implicated->other	0.05	0.23	0.00	0.23
optimal->implicated	0.82	0.35	0.23	0.93
optimal->optimal	**2.94**	0.20	**6.59**	0.53
optimal->other	1.05	0.48	0.05	1.18
other->implicated	0.43	0.87	0.00	0.93
other->optimal	0.50	0.19	0.08	1.00
other->other	0.65	**3.76**	0.02	**1.31**
Cluster size	82	75	88	90
Within_SS	304.37	235.81	69.89	415.53
between_SS / total_SS	75.1 %			

[a] This table shows the cluster means of a k-means clustering with 1,000 random starting points. The columns show how many out of seven times a participants showed a specific behavior on average in each cluster. The behavior in the control scenario is denoted on the left before the '->', the behavior in the respective fallacy scenario is denoted on the right after the '->', where 'implicated' means fallacy-implicated option, 'optimal' the optimal option (5) and 'other' every other option. A number in bold stands for the maximum value in the respective cluster.

distracted from the optimal option as well (0.93 times to the implicated option, 1.18 times to one of the other three). They also choose the fallacy-implicated option more often in the fallacy scenarios (2.37 times, vs. 1.11 times in the control scenarios). However, their erraticism also enables them to find the optimal decision in a fallacy scenario when they failed to do so in the control scenario relatively often (in around 1.37 of the 7 cases, compared to only 0.71 times for cluster 1).

The most striking feature of cluster 2 (75 participants) is that its participants rarely if ever find any optimal solution, be it in the control scenarios (1.03 times) or the fallacy scenarios (0.55 times). They are prone to fallacies and choose the implicated options in 1.99 fallacy scenarios but in only 1.16 control scenarios. Unlike cluster 4 however, they do not show the erraticism that helps them to correct errors (they switch from any of the four non-optimal options to the optimal option in only 0.35 fallacy scenarios).

It stands out that there is no specific cluster that shows switches to the fallacy-implicated option particularly often. While these switches do occur in the clusters 1, 2 and 4, and only cluster 3 seems to not be vulnerable for fallacies, the much more important features for the clusters seem to be optimality and consistency. This leads us to conclude that, while there are participants that choose optimally far more often than others, there is no particular group of participants who regularly choose fallacy-implicated options. At the same time, however, only a minority of participants seems to understand repayment problems well enough to resist fallacies. This supports our main results from Section 2.3 where we have already shown that certain fallacies indeed lead to an increased number of participants choosing the fallacy-implicated option. The auxiliary results from this section do not allow us to pin this behavior to a distinct group of people, but the findings underline that a certain vulnerability to fallacies seems to be the norm rather than the exception.

Appendix A.3 - Experiment #2

A.3.1 Explanation of numbers

We set the starting debts on each credit card to \$2200 and the starting income on the checking account to \$250. One of the credit cards has an interest rate of 3% per round and the other one of 5% per round. In every round the checking account is refilled with \$250. We choose the particular values for account levels and interest rates because they fulfill several conditions:

1. Both credit cards start with the same amount of debts, so the balances do not "favor" any card in the first round.

2. It is not possible to repay one of the cards completely in ten rounds. For our research questions we are only interested in situations where subjects actually have to make a choice between two cards. Therefore ruling out this possibility ensures that we can evaluate every round of each subject.

3. The total new debt on both cards in each round does not exceed the income in the checking account, so subjects would not get the impression of “pointless repayments” due to runaway debts.[2]

[2] If a subject does not repay anything at all, then their total new debts do exceed the checking account deposits in rounds 9 and 10. But if subjects do not repay at all, their “repayment” behavior could not be distracted by any feelings of fatalism anyway.

A.3.2 Additional Figures and Tables

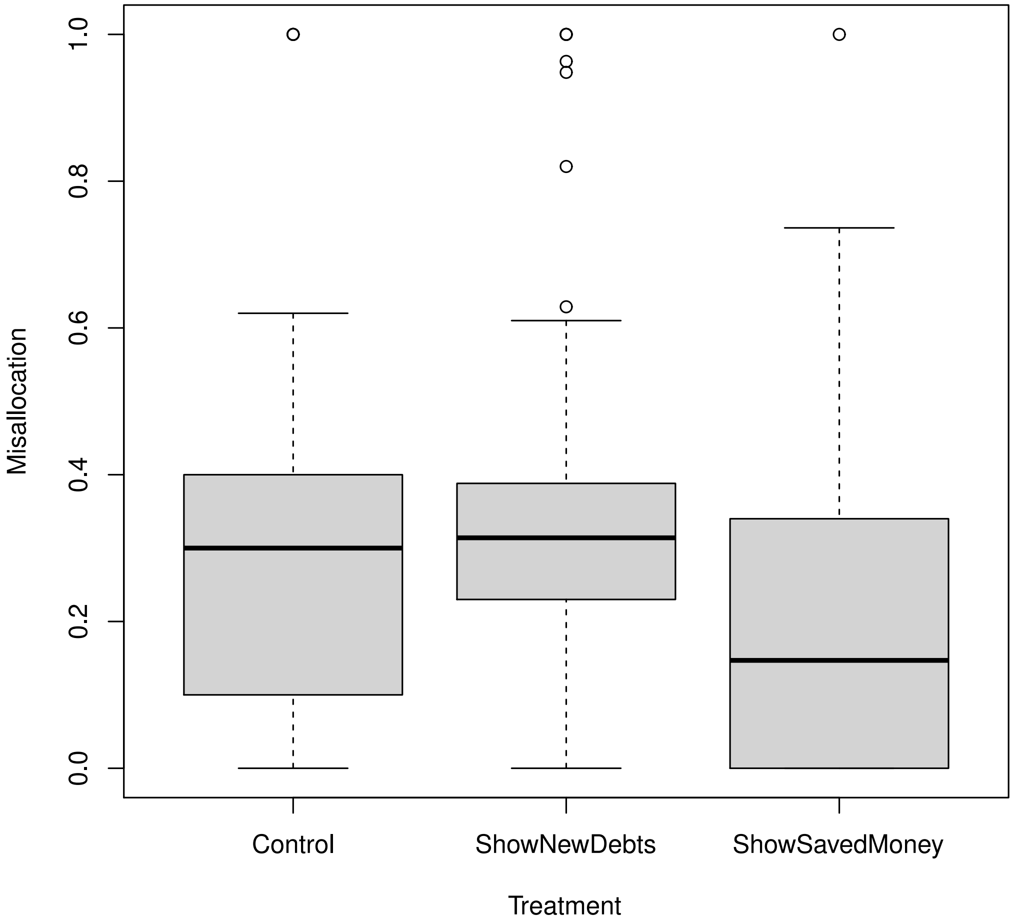

Figure 5: Boxplots of misallocation split by treatment. ShowNewDebts is the sludge and ShowSavedMoney is the nudge treatment.

Table 14: Misallocation split by round class, random effects regression[a]

	Dependent variable: Misallocation		
	Minimal	Akaike-optimal	Full model
	(1)	(2)	(3)
High_int_class	−0.224***	−0.224***	−0.219***
Standard error	(0.037)	(0.037)	(0.038)
p-value	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.000]	[0.000]	[0.000]
ShowNewDebts	0.112	0.102	0.106
Standard error	(0.059)	(0.059)	(0.059)
p-value	[0.057]	[0.080]	[0.075]
Holm adjusted	[0.114]	[0.221]	[0.224]
ShowSavedMoney	−0.181***	−0.188***	−0.173***
Standard error	(0.046)	(0.046)	(0.046)
p-value	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.000]	[0.000]	[0.000]
High_int_class · ShowNewDebts	−0.098	−0.098	−0.098
Standard error	(0.055)	(0.055)	(0.057)
p-value	[0.073]	[0.074]	[0.088]
Holm adjusted	[0.073]	[0.221]	[0.224]
High_int_class · ShowSavedMoney	0.150**	0.150**	0.139**
Standard error	(0.042)	(0.042)	(0.044)
p-value	[0.000]	[0.000]	[0.001]
Holm adjusted	[0.001]	[0.002]	[0.006]
Financial literacy		−0.015	−0.023
Standard error		(0.010)	(0.017)
p-value		[0.135]	[0.183]
Holm adjusted		[0.221]	[0.224]
Years of education (yoe)		−0.011*	−0.012*
Standard error		(0.005)	(0.005)
Credit card order (desc.)			−0.024
Standard error			(0.024)
Dummy: Male			−0.003
Standard error			(0.024)
Age			−0.001
Standard error			(0.001)
# Credit cards			−0.000
Standard error			(0.006)
# Additionally accessible credit cards			−0.007
Standard error			(0.010)
Dummy: Use credit card at work			0.007
Standard error			(0.032)
Dummy: At work, but usually don't use			−0.035
Standard error			(0.071)
Dummy: Usually do not use credit cards			0.007
Standard error			(0.034)
ShowNewDebts · Financial literacy			0.011
Standard error			(0.026)
ShowSavedMoney · Financial literacy			0.001
Standard error			(0.023)
Constant	0.328***	0.508***	0.551***
Standard error	(0.040)	(0.089)	(0.103)
Observations	522	522	498
Overall R^2	0.230	0.245	0.246
Akaike Inf. Crit.	5.17	−0.83	17.95

Note: $^{*}p<0.05$;$^{**}p<0.01$;$^{***}p<0.001$ for the Holm-adjusted p-values
Financial literacy is centralized at a value of 4.

[a] High_int_class = 1 if the high interest rate credit card produces more debt in the observed round, High_int_class = 0 otherwise. ShowNewDebts is the sludge and ShowSavedMoney is the nudge treatment. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. P-values adjusted for High_int_class, ShowNewDebts, ShowSavedMoney, High_int_class · ShowNewDebts, High_int_class · ShowSavedMoney and Financial literacy. Asterisks indicate significance after adjustment.

Table 15: Comparison of the three treatments via OLS regression, with misallocation as dependent variable.[a]

	Dependent variable: Misallocation		
	Minimal	Akaike-optimal	Full model
	(1)	(2)	(3)
ShowNewDebts	0.038	0.030	0.031
Standard error	(0.025)	(0.024)	(0.030)
p-value	[0.131]	[0.212]	[0.303]
Holm adjusted	[0.131]	[0.212]	[0.303]
ShowSavedMoney	-0.087^{***}	-0.091^{***}	-0.080^{*}
Standard error	(0.023)	(0.022)	(0.029)
p-value	[0.000]	[0.000]	[0.006]
Holm adjusted	[0.000]	[0.000]	[0.011]
Financial literacy		-0.046^{***}	-0.044^{**}
Standard error		(0.008)	(0.014)
p-value		[0.000]	[0.002]
Holm adjusted		[0.000]	[0.006]
Constant	0.274^{***}	0.421^{***}	0.409^{***}
Standard error	(0.018)	(0.067)	(0.077)
Observations	404	404	379
Interact. Fin.lit._treatments	No	No	Yes
Further control variables	No	only YOE[b]	Yes
R^2	0.068	0.183	0.189
Akaike Inf. Crit.	−168.51	−217.46	−213.4

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$ for the Holm-adjusted p-values

[a] Model 1 includes dummy variables for the respective treatments, model 3 includes all control variables and interactions, model 2 is the AIC-optimal model. ShowNewDebts is the sludge and ShowSavedMoney is the nudge treatment. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for all the reported coefficients, but not the control variables. Asterisks indicate significance after adjustment.
All models show that the misallocation smaller in the ShowSavedMoney treatment. An increase of one unit in the financial literacy sum index leads to an average decrease in the misallocation by more than 4% in every treatment, so pre-knowledge seems to have an effect on the overall misallocation.

[b] Years of education

Table 16: Random effects regression of misallocation for each round[a]

	Dependent variable: Misallocation		
	Minimal	Akaike-optimal	Full model
	(1)	(2)	(3)
Round	0.020^{***}	0.020^{***}	0.020^{***}
Standard error	(0.002)	(0.002)	(0.002)
p-value	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.000]	[0.000]	[0.000]
ShowNewDebts	0.038	0.030	0.036
Standard error	(0.025)	(0.024)	(0.025)
p-value	[0.130]	[0.210]	[0.151]
Holm adjusted	[0.130]	[0.210]	[0.151]
ShowSavedMoney	-0.087^{***}	-0.091^{***}	-0.088^{***}
Standard error	(0.023)	(0.022)	(0.022)
p-value	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.000]	[0.000]	[0.000]
Financial literacy		-0.046^{***}	-0.044^{**}
Standard error		(0.008)	(0.014)
p-value		[0.000]	[0.002]
Holm adjusted		[0.000]	[0.003]
Constant	0.166^{***}	0.267^{***}	0.255^{**}
Standard error	(0.020)	(0.070)	(0.078)
Observations	4,040	4,040	3,790
Interact. Fin.lit._treatments	No	No	Yes
Further control variables	No	only YOE[b]	Yes

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$ for the Holm-adjusted p-values
Financial literacy was centralized at a value of 4.

[a] Model (1) is without further control variables, model (3) contains all control variables and model (2) contains only the variables that are optimal according to Akaike's information criterion from the main analysis in table **??**. ShowNewDebts is the sludge and ShowSavedMoney is the nudge treatment. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for all the reported coefficients, but not the control variables. Asterisks indicate significance after adjustment.

[b] Years of education

Table 17: Misallocation split by round class (including screened out participants), random effects regression[a]

	Dependent variable: Misallocation		
	Minimal	Akaike-optimal	Full model
	(1)	(2)	(3)
High_int_class	−0.226***	−0.226***	−0.222***
Standard error	(0.037)	(0.037)	(0.038)
p-value	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.000]	[0.000]	[0.000]
ShowNewDebts	0.110	0.101	0.104
Standard error	(0.058)	(0.058)	(0.059)
p-value	[0.060]	[0.084]	[0.080]
Holm adjusted	[0.120]	[0.238]	[0.240]
ShowSavedMoney	−0.185***	−0.193***	−0.178***
Standard error	(0.045)	(0.045)	(0.046)
p-value	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.000]	[0.000]	[0.001]
High_int_class · ShowNewDebts	−0.096	−0.096	−0.095
Standard error	(0.055)	(0.055)	(0.057)
p-value	[0.079]	[0.079]	[0.095]
Holm adjusted	[0.120]	[0.238]	[0.240]
High_int_class · ShowSavedMoney	0.156***	0.156***	0.145 $* *$
Standard error	(0.042)	(0.042)	(0.043)
p-value	[0.000]	[0.000]	[0.001]
Holm adjusted	[0.001]	[0.001]	[0.003]
Financial literacy		−0.015	−0.023
Standard error		(0.010)	(0.017)
p-value		[0.135]	[0.187]
Holm adjusted		[0.238]	[0.240]
Years of education (yoe)		−0.011*	−0.011*
Standard error		(0.005)	(0.005)
Credit card order (desc.)			−0.024
Standard error			(0.024)
Dummy: Male			−0.005
Standard error			(0.024)
Age			−0.001
Standard error			(0.001)
# Credit cards			−0.001
Standard error			(0.005)
# Additionally accessible credit cards			−0.006
Standard error			(0.010)
Dummy: Use credit card at work			0.008
Standard error			(0.031)
Dummy: At work, but usually don't use			−0.035
Standard error			(0.071)
Dummy: Usually do not use credit cards			0.005
Standard error			(0.034)
ShowNewDebts · Financial literacy			0.011
Standard error			(0.026)
ShowSavedMoney · Financial literacy			0.002
Standard error			(0.022)
Constant	0.329***	0.506***	0.549***
Standard error	(0.040)	(0.087)	(0.099)
Observations	528	528	504
Overall R^2	0.232	0.246	0.247
Akaike Inf. Crit.	0.4	−5.32	13.71

Note: $^{*}p$<0.05;** p<0.01;*** p<0.001 for the Holm-adjusted p-values
Financial literacy is centralized at a value of 4.

[a] High_int_class = 1 if the high interest rate credit card produces more debt in the observed round, High_int_class = 0 otherwise. ShowNewDebts is the sludge and ShowSavedMoney is the nudge treatment. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values for the variables we interpret in brackets. The p-values are adjusted for High_int_class, ShowNewDebts, ShowSavedMoney, High_int_class · ShowNewDebts, High_int_class · ShowSavedMoney and Financial literacy, but not the control variables. Asterisks indicate significance after adjustment.

A.3.3 An additional experiment with independent rounds as a robustness check

A potentially important limiting factor of experiment #2 might be that subjects are fully comparable only at the beginning of the experiment, because repayment decisions in the earlier rounds determine whether the Cuckoo Fallacy becomes possible at all. Participants who repay non-optimally from the very beginning, i.e. who always repay the cheaper card, often do not even get the chance to succumb to the Cuckoo Fallacy. This endogeneity problem casts doubts on the internal validity of our treatments in later rounds. However, real credit card repayments *are* endogenous. Since credit card debt has a revolving character, credit card repayments are usually not one shot decisions, but a series of decision, where the later decisions depend on the earlier ones. If we ignore this dependency, we lose external validity. In particular, it could be the case that the dependency itself influences our participants' reactions to the nudge and the sludge. This is the reason why we focus heavily on the dependent rounds design, but to solve this dilemma, we run an additional experiment which uses ten independent rounds as a robustness check.

The setup of this experiment #2.2 is as identical to the main experiment #2 as possible. We keep the three treatments (*ShowNewDebts* as a sludge, *ShowSavedMoney* as a nudge, and the control group), and design their framing identical to experiment #2. The values for income and interest rates stay at $250, 3% and 5%, respectively. The first round also starts with $2200 of debts on each card. The major change is that while in experiment #2 the rounds were dependent, here they are not. From round 2 on, participants go through a series of 9 independent decision problems where we change the values for the balances. This design is identical to the idea of a "scenario" in experiment #1, thus we refer to each round as a scenario.

We set the balances in the nine scenarios such that we get three different possible "scenario types", with three scenarios per type. Table 18 shows the scenarios. In the first scenario type (*CuckooPossible*) the 3% credit card accumulates more new debts when interests are charged than the 5% credit card, so it is possible to succumb to the Cuckoo Fallacy. In the second scenario type (5%, *impossible*) the 5% card has a higher debt balance than the 3% card, so the Cuckoo fallacy is not possible. The third scenario type, (3%, *impossible*), stands right between the other two types. It takes into account that the 3% card might have a higher balance than the 5% card, but not so much that it accumulates more debts. Thus, the Cuckoo Fallacy is also not possible. We implement this additional type to distinguish the effects of the Cuckoo Fallacy from a simple "repay the higher balance card" heuristic.

We randomize the order of these 9 scenarios. We also randomize the credit card order in round 1, but

Table 18: Description of the scenarios[a]

Scenario No.	Scenario type	Checking account	Credit card 1	Credit card 2	Interest rate 1	Interest rate 2
1	First scenario	$250	$-2200	$-2200	3 %	5 %
2	CuckooPossible	$250	$-2550	$-1360	3 %	5 %
3	CuckooPossible	$250	$-2440	$-1230	3 %	5 %
4	CuckooPossible	$250	$-2260	$-1300	3 %	5 %
5	3%, impossible	$250	$-2270	$-1950	3 %	5 %
6	3%, impossible	$250	$-2330	$-1890	3 %	5 %
7	3%, impossible	$250	$-2400	$-1720	3 %	5 %
8	5%, impossible	$250	$-2020	$-2300	3 %	5 %
9	5%, impossible	$250	$-1990	$-2430	3 %	5 %
10	5%, impossible	$250	$-1300	$-2170	3 %	5 %

[a] This table shows the values for the income on the checking account, for the credit card balances, and for the interest rates. Each block contains the scenarios for one scenario type. The scenarios were presented in randomized order except for the *First scenario* which always was presented first to recreate the starting point from Experiment # 2.

then keep it the same for all other rounds to stay close to experiment #2. Additionally, while in the main experiment #2 participants can see the results of their actions in the changes of the balances at the start of the next round, in experiment #2.2 we do not give them any information after the rounds. Instead we show them the total results after all ten rounds are finished. We implement this change to minimize any dependencies between the rounds, such as learning, as much as possible.

The dependent variable is misallocation and the regression formula is as follows:

$$Y_{i,j} = \beta_0 + \beta_1 \cdot treatment_i + \beta_2 \cdot scenario\_type_{i,j} + \beta_3 \cdot treatment_i \times scenario\_type_{i,j} +$$
$$\beta \cdot controls_i + u_i + \epsilon_{i,j},$$

where j is one scenario and u_i is the random intercept term for participant i, $controls_i$ stands for the vector of additional control variables, depending on the model, and β for the corresponding vector of estimated coefficients. While the treatments are between-subject and identical to experiment #2, the scenario types are within-subject. In the Akaike-optimized model, these are financial literacy and years of education. In the full model, they also include the interaction between treatments and financial literacy, age, # credit cards, # additionally accessible credit cards, gender, credit card order, and dummy variables that indicate if credit cards are used at work, and if the participant usually does not use credit cards, but generally knows how they work.

Outside of the experimental stage, we only make minor changes to the wording in the instructions and adapt the second comprehension task. We keep the three treatments and the post experimental questionnaire, and pay a $1 show-up fee and up to $2 as bonus, which is again calculated via the repayment efficiency.

805 MTurkers started our experiment, of which 660 finished it. Out of the 145 who did not finish the experiment, 37 dropped out before the basic numeracy question, 55 did not pass the basic numeracy question, and 53 dropped out within the experiment or the post experimental questionnaire - 17 in each the Basic and the ShowSavedMoney treatment, 19 in the ShowNewDebts treatment. Out of the remaining 660 participants, 496 passed both attention tests. This number already indicates that the data quality of this sample, which we collected in December 2021 and January 2022, is lower than the sample from experiment #2, which we collected more than three years earlier and where we only lost 10 participants to the attention checks. Other authors find such drops within the same time frame as well (Chmielewski and Kucker, 2020). The open anti-bot question in which we asked for the repayment strategies shows additional problems, which it did not in experiment #2, because there the few suspicious participants were already screened out due to the other data quality measures. In experiment #2.2 however, a large number of participants who passed the automatic screening process gave answers that did not fit the question ("i choose with my own perfection"), that were generic answers such as "good survey", "no" or "very interesting" or that clearly showed that the respective participant is not fluent in English. Others described not how they repaid in the experiment but how they generally think one uses credit cards (e.g. "Pay off your balance every month"), and some even copied the first sentences of some Google search results ("Two of the most popular strategies for paying off debt on your own are the snowball method and the avalanche method. Both methods require making the minimum monthly payments on all but one debt, which you put extra money towards" [remark: comment ends here]). Two raters went over the answers independently to mark them as "suspicious" based on these problems. For our main analysis we only use the data from participants who none of the raters marked as suspicious. These 291 participants form our main sample. We report summary statistics in Table 19. As a robustness check we add the participants which only one rater found suspicious. We dropped everyone who both raters marked as suspicious.

Because the data quality is such an obvious confounder, we refrain from am lengthy comparison of the results from experiment #2 with experiment #2.2 and mainly focus on investigating our hypotheses. Nevertheless, we consider it interesting to investigate whether participants in experiment #2.2 generally perform differently compared to participants in experiment #2, in particular because experiment #2.2 was

Table 19: Summary statistics of participants (additional experiment)

Overall statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Financial Literacy	291	3.54	1.38	0	3	4	5	6
Age	291	36.86	10.33	20	30	35	42	78
# Credit cards	268	2.27	1.82	0	1	2	3	12
# Additionally accessible credit cards	255	0.69	1.30	0	0	0	1	10
# Years of education	291	15.47	2.17	9	14	16	16	21
Experiment duration (min:sec)	291	19:13	13:08	4:11	10:59	14:37	22:22	100:58
Payoff (USD)	291	1.72	0.33	0.00	1.62	1.76	2.00	2.00
Gender info	Males: 168		Females: 121		Third gender: 2			
Basic treatment	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Financial Literacy	95	3.46	1.48	0	2	4	5	6
Age	95	35.62	8.12	23	30	34	39	62
# Credit cards	85	1.96	1.73	0	1	2	2	12
# Additionally accessible credit cards	77	0.52	1.01	0	0	0	1	6
# Years of education	95	15.55	1.95	10	14	16	16	21
Experiment duration (min:sec)	95	18:44	11:40	4:11	10:58	14:32	22:16	66:46
Payoff (USD)	95	1.66	0.36	0.00	1.56	1.73	1.86	2.00
Gender info	Males: 57		Females: 37		Third gender: 1			
ShowNewDebts-treatment	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Financial Literacy	94	3.66	1.29	1	3	4	5	6
Age	94	36.01	10.56	20	28.2	34	40.8	78
# Credit cards	87	2.56	2.02	0	1	2	3.5	10
# Additionally accessible credit cards	86	0.64	1.18	0	0	0	1	6
# Years of education	94	15.60	2.22	9	14	16	16	21
Experiment duration (min:sec)	94	17:50	13:36	4:59	10:42	13:47	20:37	100:58
Payoff (USD)	94	1.75	0.32	0.02	1.65	1.79	2.00	2.00
Gender info	Males: 55		Females: 39		Third gender: 0			
ShowSavedMoney-treatment	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Financial Literacy	102	3.51	1.36	0	3	4	4.8	6
Age	102	38.78	11.69	21	30	37	45.5	77
# Credit cards	96	2.27	1.68	0	1	2	3	8
# Additionally accessible credit cards	92	0.89	1.57	0	0	0	1	10
# Years of education	102	15.28	2.32	10	14	16	16	21
Experiment duration (min:sec)	102	20:54	13:53	6:44	11:53	16:27	24:20	88:29
Payoff (USD)	102	1.75	0.31	0.02	1.64	1.77	2.00	2.00
Gender info	Males: 56		Females: 45		Third gender: 1			

conducted in the COVID-19 pandemic, while experiment #2 was conducted before the pandemic. Some studies find that economic preferences and behavior might change during the COVID-19 pandemic. Harrison et al. (2022), for example, find greater risk aversion in the pandemic, while De Pue et al. (2021) report a severe impact on the mental health of older adults, a group who self-reports reduced cognitive functioning. Alsharawy et al. (2021) show that various economic preferences such as risk and time preferences depend on the fear of the pandemic. We might find different repayment behavior in our study as well.

We only compare the first rounds of each experiment as only there the situation is completely identical for participants. Linear regressions show that the average participant in experiment #2.2 exhibits less misallocation than the average participant in experiment #2. However, this difference is only significant when we exclude the participants which at least one rater marked as suspicious, and include all control variables. There is no significant interaction of experiment and treatment. We take this as evidence that the lower data quality as well as the pandemic time frame of experiment #2.2 are only reflected in the necessity to screen out more participants, but not in a potentially worse performance of the remaining participants.

We now evaluate the results of experiment #2.2. Unlike in the other experiments, we interpret both adjusted and unadjusted p-values, because we want to highlight that the combination of problematic data and the testpower-draining adjustment procedure increases the probability for a type II error considerably.

We present the results in Table 20. We repeat our approach from experiment #2 and report the minimal model employing only the treatments, scenario types and their interactions, the full model with all control variables we also used in experiment #2, and the model with the control variables according to Akaike's information criterion. However, since the data quality is lower than expected, we run each model on two different datasets and report the results for both sets, for a total of 6 regressions. On one hand we use a strict data set where we require both raters to agree that an observations is not suspicious and on the other hand - as robustness check - a tolerant data set where only one rater has to identify an observation as non-suspicious.

The first main results for this additional experiment can be found in the "scenario type"-block of the table. Recall that in both the *CuckooPossible* - which in this table is the reference category - as well as the $3\%, impossible$ scenario type, the 3% accumulates more new debts. If participants would only follow a "repay the higher balance"-heuristic, we might see a significant $5\%, impossible$, but not a significant coefficient for $3\%, impossible$. However, in all regressions, the $3\%, impossible$ and the $5\%, impossible$ type both show significantly lower misallocation than the *CuckooPossible* type. Thus we can verify additional misallocation specifically in Cuckoo Fallacy situations beyond a simple "repay the higher balance" heuristic.

The second block in the table shows the treatment effects, our second main results. As in experiment #2, we do not see any significant difference for the ShowNewDebts treatment. For the ShowSavedMoney treatment however, there is evidence that the Cuckoo Fallacy is less of a problem than in the control group. All models show significant unadjusted p-values, and in models (2), (5) and (6) the coefficients survive the multiple-hypothesis adjustment.

The third block of the table presents the interactions between treatments and scenario type. In general, we do not find any significant interactions, no matter if with or without adjustment, except for the very last interaction. If we ignore p-value adjusting, we find significantly positive effects of the ShowSavedMoney treatment when the 5% card accumulates more debts and thus the Cuckoo Fallacy is impossible. This means that in this situation, the nudge does not work as well as when the Cuckoo Fallacy is possible. Since the nudge is supposed to suppress the fallacy, this is in line with our expectations. However, these effects only survive multiple-hypothesis adjustment in model (5), which is why we interpret this as very weak evidence.

We present a visualization of misallocation in treatments and scenario types in Figure 6. Both the ShowNewDebts and the ShowSavedMoney treatment show lower misallocation than the control group. While the average misallocation ranges between 16.7% and 23.9% when the Cuckoo fallacy is possible, the misallocation declines in the other scenario types and ranges only between 7.8% and 11.3% when the 5% credit card has a higher debt balance.

Interestingly, in this version of the experiment the sludge treatment also has a lower misallocation, even if it is not significantly lower than in the control treatment. This might indicate that the potential for problems or abuse is not that high. To conclude, the results of experiment #2.2 show that the Cuckoo Fallacy is an important driver of misallocation, and there are clues that we can manipulate the presentation to make it less likely. However, the evidence for the latter claim is weaker than in experiment #2.

Table 20: Misallocation in additional experiment, random effects regression[a]

	Dependent variable: Misallocation					
Model	Minimal	Minimal	Akaike-optimal	Akaike-optimal	Full model	Full model
Outscreening	Strict	Tolerant	Strict	Tolerant	Strict	Tolerant
	(1)	(2)	(3)	(4)	(5)	(6)
Scenario types						
3%, impossible	−0.059***	−0.056**	−0.059***	−0.056**	−0.072***	−0.066**
Standard error	(0.015)	(0.015)	(0.015)	(0.015)	(0.017)	(0.018)
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.000]	[0.001]	[0.000]	[0.001]	[0.000]	[0.002]
5%, impossible	−0.125***	−0.115***	−0.125***	−0.115***	−0.140***	−0.133***
Standard error	(0.018)	(0.017)	(0.018)	(0.017)	(0.020)	(0.018)
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Holm adjusted	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Treatments						
ShowNewDebts	−0.045	−0.035	−0.037	−0.025	−0.029	−0.035
Standard error	(0.030)	(0.030)	(0.028)	(0.028)	(0.029)	(0.027)
p-value	[0.129]	[0.242]	[0.187]	[0.363]	[0.323]	[0.198]
Holm adjusted	[0.517]	[0.968]	[0.747]	[1.000]	[0.970]	[0.793]
ShowSavedMoney	−0.071	−0.076*	−0.069	−0.067	−0.072*	−0.070*
Standard error	(0.029)	(0.029)	(0.028)	(0.028)	(0.025)	(0.025)
p-value	[0.013]	[0.008]	[0.014]	[0.015]	[0.004]	[0.004]
Holm adjusted	[0.080]	[0.046]	[0.086]	[0.091]	[0.024]	[0.026]
Interactions						
ShowNewDebts × 3%, impossible	−0.029	−0.019	−0.029	−0.019	−0.007	−0.009
Standard error	(0.022)	(0.022)	(0.023)	(0.022)	(0.024)	(0.024)
p-value	[0.202]	[0.393]	[0.202]	[0.393]	[0.765]	[0.696]
Holm adjusted	[0.605]	[1.000]	[0.747]	[1.000]	[0.970]	[1.000]
ShowNewDebts × 5%, impossible	0.010	0.013	0.010	0.013	0.024	0.022
Standard error	(0.026)	(0.024)	(0.026)	(0.024)	(0.027)	(0.025)
p-value	[0.699]	[0.591]	[0.699]	[0.591]	[0.368]	[0.385]
Holm adjusted	[1.000]	[1.000]	[1.000]	[1.000]	[0.970]	[1.000]
ShowSavedMoney × 3%, impossible	0.008	0.007	0.008	0.007	0.031	0.021
Standard error	(0.020)	(0.019)	(0.020)	(0.019)	(0.021)	(0.022)
p-value	[0.679]	[0.722]	[0.679]	[0.722]	[0.134]	[0.342]
Holm adjusted	[1.000]	[1.000]	[1.000]	[1.000]	[0.535]	[1.000]
ShowSavedMoney × 5%, impossible	0.053	0.047	0.053	0.047	0.070*	0.054
Standard error	(0.025)	(0.024)	(0.025)	(0.024)	(0.026)	(0.025)
p-value	[0.034]	[0.050]	[0.034]	[0.050]	[0.007]	[0.030]
Holm adjusted	[0.170]	[0.251]	[0.170]	[0.251]	[0.036]	[0.150]
Financial literacy			−0.040***	−0.048***	−0.053**	−0.060**
Standard error			(0.007)	(0.007)	(0.015)	(0.015)
p-value			[0.000]	[0.000]	[0.000]	[0.000]
Holm adjusted			[0.000]	[0.000]	[0.002]	[0.001]
Constant	0.239***	0.255***	0.157*	0.198**	0.213*	0.264**
Standard error	(0.021)	(0.021)	(0.063)	(0.073)	(0.088)	(0.090)
Observations	2619	3015	2619	3015	2277	2493
Subjects	291	335	291	335	253	277
Interact. Fin.lit._treatments	No	No	No	No	Yes	Yes
Further control variables	No	No	only YOE^{b}	only YOE^{b}	Yes	Yes
R^2 overall	0.056	0.040	0.123	0.131	0.151	0.199

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Financial literacy is centralized at a value of 4.

[a] This table shows the misallocation when the Cuckoo Fallacy is possible (baseline) vs. when it is not, but the 3%-card still has the higher debt balance (card3p>card5p) vs. when it is not and the 5% card has the higher debt balance (card5p>card3p). Additionally the table shows the treatments and the interaction with these scenario types. For each set of control variables there is a more strict out-screening for subjects (at least one rater screens them out) and - for robustness - a more tolerant one (both raters have to screen them out). The Akaike-optimal models (3) and (4) have the same control variables as in the main analysis. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for all the reported coefficients, but not the control variables. Asterisks indicate significance after adjustment.

[b] Years of education

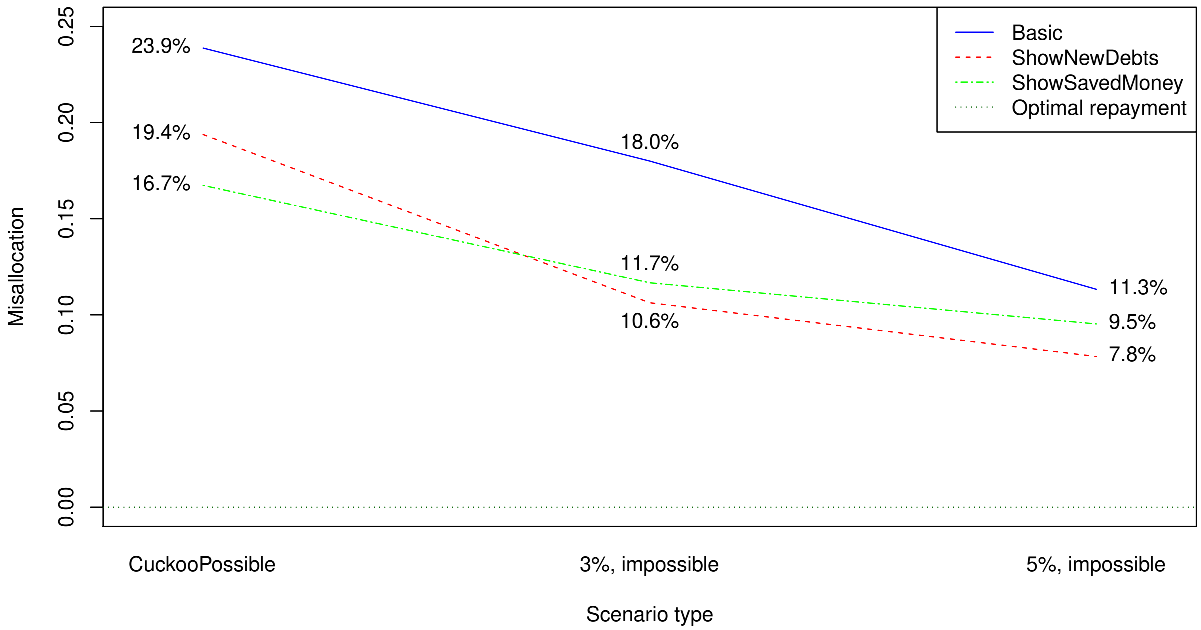

Figure 6: This Figure shows the interaction plot between treatment and scenario type. The lines represent the treatments. ShowNewDebts (in red) is the sludge and ShowSavedMoney (in green) is the nudge treatment. The x-axis shows the scenario types. We differ between scenarios whether the Cuckoo Fallacy was possible (*CuckooPossible*) vs. whether the 3% credit card has a higher debt balance than the 5% card but not enough that the Cuckoo Fallacy was possible (3%, *impossible*), or whether the 5% card has a higher debt balance than the 3% card (5%, *impossible*).

Appendix A.4 - Lab replication

We replicate the Basic treatment of experiment #2 in the experimental econ laboratory of the University of Heidelberg in July 2019 (n=96). The experiment was translated in German and adapted to the lab. Overall we find more misallocation than in the MTurk experiment, although we had higher financial incentives (participation fee: Euro 4, bonus: up to Euro 10). See the results in Table 21 and Figure 7. Although the lab does not differ significantly in all regression models, we can conclude that participants on MTurk at least do not exhibit higher misallocation than lab participants, despite of clearly lower stakes. This legitimizes the usage of MTurk for this experiment. Furthermore, we also detect a slight increase of misallocation over the experiment rounds, such that - like on MTurk - we do not see any learning effects in the lab.

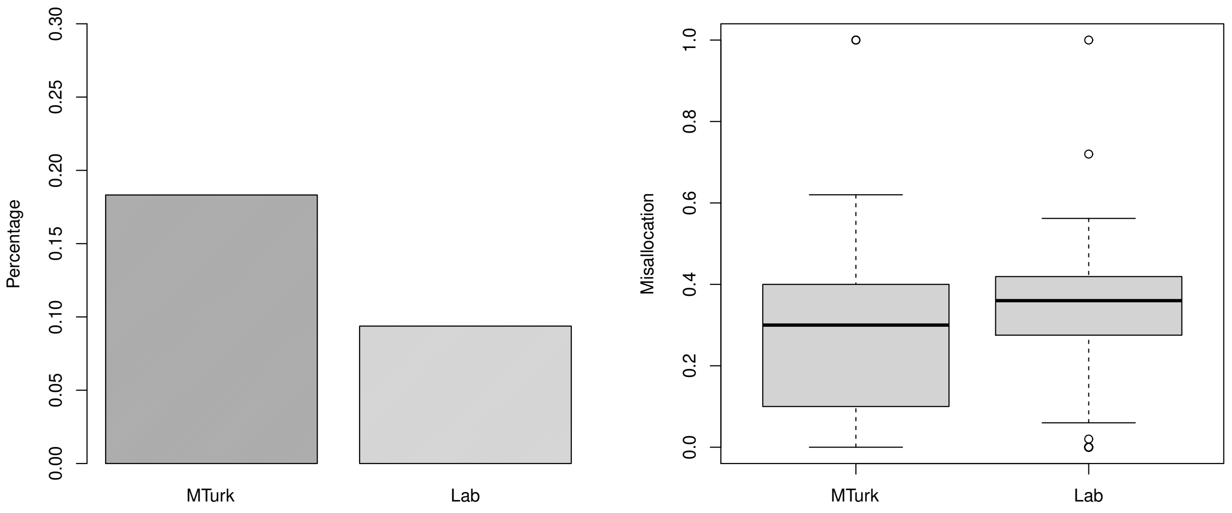

Figure 7: The bars in the left Figure show the proportion of subjects without any misallocation in all the experiment rounds. 18.3% of the subjects in MTurk and 9.4% of the subjects in the lab do not exhibit misallocation. The difference is not significant (p=0.0526). The right Figure shows the boxplots of misallocation of all participants on average. The average misallocation in the lab is significantly higher than on MTurk (p = 0.0242).

Table 21: Comparison Lab and MTurk via OLS regression, with misallocation as dependent variable.[a]

	Dependent variable: Misallocation		
	Minimal	Akaike-optimal	Full model
	(1)	(2)	(3)
Dummy Laboratory	0.057^{*}	0.060^{**}	0.049
Standard error	(0.024)	(0.022)	(0.027)
p-value	[0.018]	[0.007]	[0.065]
Holm adjusted	[0.018]	[0.007]	[0.065]
Credit card order (desc.)		0.067^{**}	0.067^{**}
Standard error		(0.023)	(0.023)
p-value		[0.004]	[0.004]
Holm adjusted		[0.008]	[0.007]
Financial literacy		-0.048^{***}	-0.049^{***}
Standard error		(0.010)	(0.010)
p-value		[0.000]	[0.000]
Holm adjusted		[0.000]	[0.000]
Constant	0.274^{***}	0.421^{***}	0.451^{***}
Standard error	(0.018)	(0.041)	(0.069)
Observations	227	227	227
Further control variables	No	No	Yes
R^2	0.022	0.167	0.170
Akaike Inf. Crit.	−115.79	−148.06	−143.03

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$ for the Holm-adjusted p-values

[a] Model (1) includes only a dummy for the lab, model (3) includes all control variables, model (2) is the AIC-optimal model. Robust standard errors in parentheses, unadjusted p-values and Bonferroni-Holm adjusted p-values in brackets. The p-values are adjusted for all the reported coefficients, but not the control variables. Asterisks indicate significance

Appendix references

Agarwal, S., S. Chomsisengphet, C. Liu, and N. S. Souleles (2015). Do Consumers Choose the Right Credit Contracts? *Review of Corporate Finance Studies 4*(2), 239–257.

Akaike, H. (1974, December). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Alsharawy, A., S. Ball, A. Smith, and R. Spoon (2021). Fear of COVID-19 changes economic preferences: evidence from a repeated cross-sectional MTurk survey. *Journal of the Economic Science Association 7*(2), 103–119.

Amar, M., D. Ariely, S. Ayal, C. E. Cryder, and S. I. Rick (2011). Winning the Battle but Losing the War: The Psychology of Debt Management. *Journal of Marketing Research 48*, 38–50.

Camerer, C. F. (2015). The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List. In G. Fréchette and A. Schotter (Eds.), *Handbook of Experimental Economic Methodology*, Chapter 14, pp. 249–296. Oxford University Press.

Camerer, C. F. and R. M. Hogarth (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework, journal = Journal of Risk and Uncertainty. *19*(1-3), 7–42.

Chmielewski, M. and S. C. Kucker (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science 11*(4), 1–10.

De Pue, S., C. Gillebert, E. Dierckx, M.-A. Vanderhasselt, R. De Raedt, and E. Van den Bussche (2021). The impact of the COVID-19 pandemic on wellbeing and cognitive functioning of older adults. *Scientific Reports 11*(1).

Dhami, S. (2016). *The Foundations of Behavioral Economic Analysis*. Oxford University Press.

Gathergood, J., N. Mahoney, N. Stewart, and J. Weber (2019). How Do Individuals Repay Their Debt? The Balance-Matching Heuristic. *American Economic Review 109*(3), 844–875.

Harrison, G., A. Hofmeyr, H. Kincaid, B. Monroe, D. Ross, M. Schneider, and J. Swarthout (2022). Subjective beliefs and economic preferences during the COVID-19 pandemic. *Experimental Economics 25*, 795–823.

Keys, B. J., D. G. Pope, and J. C. Pope (2016). Failure to Refinance. *Journal of Financial Economics 122*(3), 482–499.

Keys, B. J. and J. Wang (2019). Minimum Payments and Debt Paydown in Consumer Credit Cards. *Journal of Financial Economics 131*(3), 528–548.

Peer, E., J. Vosgerau, and A. Acquisti (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods 46*(4), 1023–1031.

Ponce, A., E. Seira, and G. Zamarripa (2017). Borrowing on the Wrong Credit Card? Evidence from Mexico. *American Economic Review 107*(4), 1335–1361.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika 18*, 267–276.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics 13*, 75–98.