Some Biases are WEIRDer than Others: Testing Canonical Behavioral Biases in Two Non-WEIRD Samples

Nicholas Owsley[*], Narges Hajimoladarvish[†]

March 24, 2024

Abstract

While efforts to replicate behavioral biases with non-Western populations have expanded, little research tests these biases with the world's poorest people. At the same time, behavioral economics insights increasingly inform policies targeting the world's poor. This paper presents results from an experiment testing 10 of the canonical biases from the behavioral economics literature with two distinct non-WEIRD populations in India: low-income people targeted by a government campaign applying behavioral insights, and university students from an elite Indian university. We find evidence of the tested bias in 9 of 11 contrasts for the pooled sample, and in 8 of the 11 contrasts for both sub-samples separately, with effect sizes broadly comparable to meta-analytic benchmarks from Western samples. For all but two biases, effect sizes are no different between the student and low-income samples. For these exceptions — mental accounting and the common ratio effect — the student sample (plausibly more similar to educated Western samples) showed more bias than the low-income sample. Thus, while our results support the replication of most behavioral biases with Non-WEIRD samples, they also suggest that some biases might be WEIRDer than others.[1]

[*]University of Chicago, Booth School of Business, Illinois, USA; Busara Center for Behavioral Economics, Nairobi, Kenya, nowsley@chicagobooth.edu; orcid: 0000-0001-7898-6947

[†]Centre for Social and Behaviour Change, Ashoka University, India, narges.hajimoladarvish@ashoka.edu.in

[1]This research was funded by the Centre for Social and Behaviour Change at Ashoka University.

1 Introduction

Most research subjects in behavioral economics are culturally and socioeconomically homogeneous (Henrich et al., 2010b). Western populations account for 96% of research subjects in top psychology journal publications (Arnett, 2016) and the bulk of subjects in foundational behavioral economics research (Thaler, 2018; Kahneman, 2011). Even within this peculiar group, the significant majority of research participants are university students, and even more specifically, psychology and economics undergraduates. While this sampling bias does not pose a challenge to the internal validity of individual effects, the accumulation of these findings has produced a skewed body of evidence that is far too often taken as universal and which provides the basis for decision-making models and the direction of research in applied settings (Fehr and Fischbacher, 2002; Cohen et al., 2020; Thaler, 2018; Camerer et al., 2016; Rad et al., 2018). This is particularly problematic given that several studies find considerable heterogeneity in decision patterns across different populations (Henrich et al., 2010b,c; Na et al., 2010; Barrett et al., 2016). Social preferences, risk preferences, and several other areas of decision-making appear to differ sharply across different geographies and socio-economic groups (Falk et al., 2018; Henrich et al., 2010a), and there is some evidence that more fundamental differences in cognition might exist (Nisbett et al., 2001; Tomasello et al., 2005; Talhelm et al., 2014). This highlights the potential risk that behavioral biases described in the behavioral economics literature are specific to Western samples. That is, they might be WEIRD (Western, Educated, Industrialized, Rich, and Democratic). Importantly, insights from the behavioral sciences are increasingly applied to policy in non-WEIRD contexts, often targeting some of the poorest groups (Manning et al., 2020; OECD, 2017). If these insights do not generalize to these groups, this could lead to substantial wasted resources in some of the world's most resource-scarce environments.

Recent attempts to replicate core behavioral findings and expand subject pools to less typical samples represent an important shift in the field. Several studies have demonstrated, with large cross-country samples, which effects replicate and which ones do not (Klein et al., 2018; Camerer et al., 2016; Collaboration, 2015; Ruggeri et al., 2020). However, samples from these studies remain predominantly Western and homogeneous. While several studies have made important efforts to sample from non-Western populations (Priolo et al., 2023; Bago et al., 2022; Klein et al., 2018), these are typically concentrated amongst elite university students or convenience samples. There thus remains a dearth of evidence of the replicability of behavioral science insights amongst "harder to reach" populations, specifically low-income people in developing countries (Rad et al., 2018; Bryan et al., 2021). Most replication studies also focus on large samples split across many geographic sites. While this is efficient for identifying the overall reproducibility of an average treatment effect, it is not ideal for testing heterogeneous effects across important sub-samples (McShane et al., 2019). However, identifying heterogeneous effects is both crucial for enriching models of decision-making and determining whether applications of behavioral economics will succeed across different contexts (Bryan et al., 2021). This paper attempts to expand the evidence for behavioral biases amongst non-Western and non-typical samples. We also specifically attempt to map heterogeneity in effects across important sub-samples within a non-WEIRD context (Bryan et al., 2021): specifically, we test these biases with "hard to reach", low-income participants, similar to those targeted by "behavioral development policy"[2].

[2]In the same region, government and development aid projects use similar sampling criteria as those in this study. See

We present the results of a randomized controlled trial in an experimental laboratory to test core behavioral economics measures with samples drawn from two non-Western populations of interest: 457 low-income people in Haryana State, India (typical targets of behavioral policy), and 457 students from an elite Indian University based in the same state (more similar to typical Western samples and the target of newer replication studies expanding to non-WEIRD regions). We elicited responses for each of 10 behavioral biases through 11 individual survey-based measures modeled on original measures[3]. We adjusted the exact wording of several measures through a structured contextualization process to make them more appropriate to the local context [4]. This included working with Global South behavioral scientists, "prototyping", and piloting with local populations. Pilot and qualitative data show that contextualization reduced several forms of potential measurement error (observed with similar measures in another low-income context (Coman et al., 2017)). We randomly assigned treatment for each final measure and tested for both the existence of the behavioral bias for the "pooled sample", and for effect heterogeneity between each sub-sample. The majority of the measures produced significant effects for our pooled sample: 9 of the 11 contrasts are highly significant. Of the 9 effects that replicate in the pooled sample, 6 are not statistically distinguishable from effects found in meta-analyses of the same or similar measures, while 2 are actually larger than meta-analysis effects. This suggests that our "non-WEIRD" sample largely displays these "behavioral biases" to a similar extent as Western samples. For most measures, the effects were not statistically nor qualitatively different between our "hard to reach" low-income and student sub-samples, suggesting that these populations do not systematically differ regarding these biases. However, the student population alone showed the presence of "Mental Accounting" (Thaler, 1985), and showed higher rates of the "Common Ratio effect" (Kahneman and Tversky, 1979). We also investigate effect heterogeneity across several covariates and find that the effects are remarkably stable, further supporting the generalizability of these findings.

The results from this study show that many of the core findings from behavioral economics are robustly observed across two highly distinct non-WEIRD populations. This gives cause for optimism regarding the replicability of these and similar findings across geographic and socio-economic contexts. The presence of heterogeneity in several important effects (eg., in risk preferences and Mental Accounting), however, provides an impetus for future theorizing and empirical work with more diverse samples. This study was also only conducted in one state in India, in order to give robust estimates for these specific sub-groups. This limits this study's ability to generalize to other settings and suggests that more work testing the reproducibility of similar effects amongst low-income populations is needed.

This study adds to the behavioral economics literature by testing core measures with non-typical samples, providing replication evidence for 9 behavioral biases and building the body of behavioral economics evidence, specifically in controlled laboratory environments, in the Global South. Furthermore, this study provides details on a process for contextualizing measures which shows promise for generating more valid data in non-WEIRD contexts. Finally, this study tests specifically for heterogeneity in effects between national sub-populations of scientific and policy interest and finds evidence generally supporting

here and here for examples.

[3]See the Procedures and tasks section for the full list of measures, and the Appendix for the exact wording of each

[4]The title alludes to this change: we adapt the famous "Linda Problem" (Tversky and Kahneman, 1974), measuring an instance of the Representativeness Heuristic , to a similarly constructed "Preeti Problem".

the application of behavioral models across both groups.

Methods

Sample

Summary Statistics by Sample Group

	Pooled Sample Mean	Low-income Sample Mean	Student Sample Mean
Age	23.52	27.35	19.69
	(6.84)	(7.74)	(2.07)
Female	0.62	0.68	0.56
	(0.49)	(0.47)	(0.5)
Years of Education	12.43	11.65	13.23
	(1.8)	(1.75)	(1.49)
Finished School	0.81	0.61	1.00
	(0.39)	(0.49)	(0)
Completed Some College	0.45	0.24	0.65
	(0.5)	(0.43)	(0.48)
Taken Psychology Course	0.24	0.14	0.33
	(0.43)	(0.35)	(0.47)
Married	0.31	0.61	0.00
	(0.46)	(0.49)	(0.05)
Home Language, English	0.41	0.02	0.80
	(0.49)	(0.15)	(0.4)
Home Language, Haryanvi	0.15	0.29	0.00
	(0.35)	(0.46)	(0)
Home Language, Hindi	0.43	0.68	0.19
	(0.5)	(0.47)	(0.39)
Home Language, Other	0.01	0.00	0.02
	(0.1)	(0.07)	(0.12)
English Proficiency (0-5)	3.57	2.82	4.31
	(1.39)	(1.47)	(0.77)
Political conservatism (1-9)	3.62	3.78	3.46
	(2.28)	(2.78)	(1.63)
Observations	914	457	457

Three primary criteria guided the selection of target populations. i) We sought populations from a country in the Global South in order to determine if core biases established through studies on Western samples are observed in a Global South context. ii) We sought two highly distinct groups within a Global South country to test for replication and heterogeneity across socio-economic dimensions. iii) We sought two relevant groups regarding the status of the behavioral economics literature: a group of elite university students, a population that replications and theoretical studies in the behavioral economics and judgement and decision-making are now focusing on to expand their reach to 'non-WEIRD' populations; and a group of low-income Indians, similar to those typically targeted by behaviorally informed development policy

and field research, but which continue to be excluded from more theoretical decision laboratory studies.

The final sample for this study thus consists of two groups of participants, for a full sample of 914 people:

- A sample of 457 low-income people from peri-urban villages in the vicinity of Sonepat, Haryana State, India.
- A sample of 457 students from an elite liberal arts university in Haryana State, India.

To recruit the low-income sample, we identified all distinct villages within a 35km radius of Sonepat, Haryana State, India. We then randomly selected 14 villages from this list for inclusion. We first contacted local village leaders and with their permission and help, actively recruited participants in focal public areas within each village, such as local schools. In several villages, we recruited large proportions of the adult population. We first acquired consent and then recorded baseline data for each potential participant. Participants were only included in the study if they had completed eight years of education, and were above the age of 18 and below the age of 60. A total of 457 participants from this pool completed the full experiment.

To recruit the university sample, participants were recruited from the student administration of an elite liberal arts university in Haryana State; the primary eligibility requirement was being a registered student at the university. We excluded students who were in a psychology program to avoid possible reporting bias from exposure to similar materials, and included a question and control for whether or not students had taken any behavioral economics or psychology courses. A total of 457 participants from this pool completed the full experiment.

The study is powered to detect an Minimum Detectable Effect Size (MDES) of approximately 0.2 standard deviations for within-sub-sample treatment effects, and 0.41 for between-sub-sample differences in treatment effects (described below in the Econometric strategy).

Procedures and Tasks

Task Selection

The tasks were selected to cover a range of the most well-known and influential biases in the behavioral economics literature (Thaler, 2015). We focus largely on measures derived from the early 'Heuristics and Biases' and 'Prospect Theory' work by Kahneman and Tversky; these measures and the papers in which they were originally published account for some of the most highly cited social science papers of all time (Tversky and Kahneman, 1974; Kahneman and Tversky, 1979), and Prospect Theory has specifically been described as the most influential social science theory since the beginning of the 20th century (Ruggeri et al., 2020; Barberis, 2013). In each case, we derived our final measures from original or commonly used versions of the measure to better compare to the findings with Western samples. In some cases, we underwent a rigorous iterative contextualization process (described below) to make measures culturally appropriate and preserve their ability to identify underlying biases. We finally selected 11 measures to identify 10 different biases. The full list of measures is described below, with details and wording of each measure in the Appendix.

List of Biases and Measures:

1. **Representativeness Heuristic**: Specifically the 'conjunction fallacy', a bias in probabilistic reasoning owing to the 'representativeness' of certain judgement targets leading to overestimating probability judgements (Tversky and Kahneman, 1974).
2. **Cognitive Reflection Test**: A test of the capacity for reflective thinking over wrong intuitive responses. (Frederick, 2005)
3. **Mood Heuristic**: The effect of 'mood' on general well-being assessments. (Strack et al., 1988)
4. **Anchoring**: The effect of arbitrary numerical primes, that can serve as 'anchors', on evaluations and willingness to pay (Tversky and Kahneman, 1974).
5. **Loss Framing**: A test for differences in choices in identical gambles depending on loss and gain framing (Tversky and Kahneman, 1974).
6. **Default - Assessment**: Whether a default selection for an assessment on a numerical scale affects the final assessment.
7. **Defaults - Charity Contribution**: Whether a pre-selected default option affects decisions on which charity to contribute to (Johnson and Goldstein, 2003).
8. **Present-bias**: Whether participants are more 'impatient' over choices that could yield benefits now compared with in the future.
9. **Common Ratio Effect**: Whether participants show inconsistent risk preferences across different elicitation mechanisms (specifically different probabilities) (Kahneman and Tversky, 1979).
10. **Decoy (Atrraction) Effect**: Whether the inclusion an irrelevant 'decoy' option in a menu affects choices (Huber et al., 1982).
11. **Mental Accounting**: Whether different framing of identical monetary endowments changes purchasing decisions (Thaler, 1985).

Task Contextualization

Most tasks (9 out of 11) were closely drawn from the original or most common measures in the literature [5]. Several of the these nine measures contained culturally-specific reference points that were potentially not known to the Indian subject pools. In a related study, Coman et al. (2017) find a high degree of measurement error in responses to several identical original measures in a laboratory study with student and low-income Kenyan samples [6]. As such, in these cases, there was a high risk of measurement error due to potential low comprehension of questions and choice sets, which could attenuate the estimates of biases (and as such, would not allow us to replicate these effects even if they exist with our populations of

[5] The two exceptions were the two measures of "default" effects which were straightforward applications of the principle of default selections in economic decisions.

[6] They find that several measures generate data that is equivalent to "randomly generated" data, significant evidence of "straight-lined" answers, and frequent selection of strictly and weakly dominated options.

interest). The measures therefore underwent a review and multi-step contextualization process in order to preserve the core components of original measures [7].

First, a team of India- and Kenya-based behavioral science researchers reviewed each measure for contextual appropriateness for the target subject pools; they specifically identified all measures and features that included references or concepts that were distinctly 'Western', or that might appear 'alien' to the local target populations. After determining the inappropriate features, the authors and these researchers identified the fundamental structure and key elements of each measure. This imposed a clear structure for identifying and substituting local cultural references for foreign ones; there was then a thorough review to ensure that the format and features of each original measure was preserved. All contextualized measures were then piloted with Indian field staff, which included a combination of cognitive debriefing, prototyping, and response collection. This was in order to check for comprehension, cultural appropriateness, and for markers of measurement error. Measures were then iterated following several rounds of qualitative feedback and pilot data. This process was then repeated with participants from the target subject pools, and measures were again adapted and finalized until low comprehension and clear patterns of measurement error (such as random responding, "straight-lining", or selecting strictly dominated options) were minimized. Ultimately, we finalized nine India-specific measures that preserved the structure of the originals.

Data Collection

Data collection took place between July and September 2019 and was conducted in a decision laboratory for all participants, divided into individual sessions. Each session included 20 participants, who conducted tasks and surveys on individual Windows tablet computers using the oTree software. The full set of instructions were translated and back-translated into Hindi; each screen presented instructions in both English and Hindi, and research assistants read instructions out loud for all tasks for the entire set of participants in both languages. Each participant received 400 Indian Rupees for their participation, provided in cash directly after the session. We determined that for all participants, but particularly for the low-income sample, the controlled laboratory setting with clear, translated, and dictated instructions was important for identifying biases and comparing them across sub-samples.

Each of the 11 tasks were administered sequentially after a basic demographic survey and the order of each task was randomized.

Treatment Assignment

Out of the 11 measures, 7 include between-subject treatment and control conditions - in these 7 cases, participants were randomly assigned with equal probability to each condition, and assignment occurred on the session-task level, such that individuals within the same session were randomly assigned to different conditions 'across tasks' but the same condition 'across individuals'. For each respective task, individuals within the same session are in the same treatment group (See the Appendix for details of treatment

[7] See the Appendix for both originals and contextualized versions, and for a more detailed example of the process described below

conditions for each task). For 4 of the 11 measures, we can identify the bias within-subject (The Appendix describes how each bias is identified).

Empirical Strategy

There are two types of tasks we use to identify biases: between subjects' measures, where participants are randomly assigned to a treatment or control condition, and within-subjects' measures. For the within subjects' measures, we identify individual bias for each participant as per each measure's description in the Appendix (in most cases, this is a simple transformation of their responses). The identification of the mean incidence of the bias is equivalent to an intercept-only model for each bias. To identify differences in the bias between each sub-sample for the within-subject measures, we estimate the following model:

$$Y_i = \beta_0 + \beta_1 Student_i + \epsilon_i \tag{1}$$

For biases where the identification relies on a between-subjects' design, and thus assignment to a treatment condition, we use the following set of models:

$$Y_i = \beta_0 + \beta_1 Treatment_i + \epsilon_i \tag{2}$$

$$Y_i = \beta_0 + \beta_1 Student_i + \beta_2 Treatment_i + \beta_3 Student_i * Treatment_i + \epsilon_i \tag{3}$$

Equation 2 simply allows us to identify the overall bias across both sub-samples (what we call the 'pooled sample') through β_1. For all models, Y_i is the outcome of interest for participant i (the response for each measure), $Student_i$ is a dummy variable equal to 1 if the participant is in the student sub-sample, $Treatment_i$ is a dummy variable equal to 1 if the participant was randomly assigned to a treatment group, and $Student_i * Treatment_i$ is the interaction term between student sub-sample and treatment group assignment and is used to identify the differential effect of the treatment for students. β_1 in Equation 2 provides the raw treatment effect (or bias estimate) for the full sample. β_2 in Equation 3 provides the treatment effect for the low-income group alone, while β_3 in Equation 3 provides the differential treatment effect for students compared to the low-income sub-sample. For robustness checks, each of the above models is also estimated with a vector of control variables X_i included in an additional specification.

These models apply to all measures with the exception of the Mood Heuristic, which follows a slightly different analytical procedure owing to the treatment effect being the product of an additional interaction [8].

[8] This is described under the 'Mood Heuristic' in the Appendix.

2 Results

2.1 All Biases - Overall Results

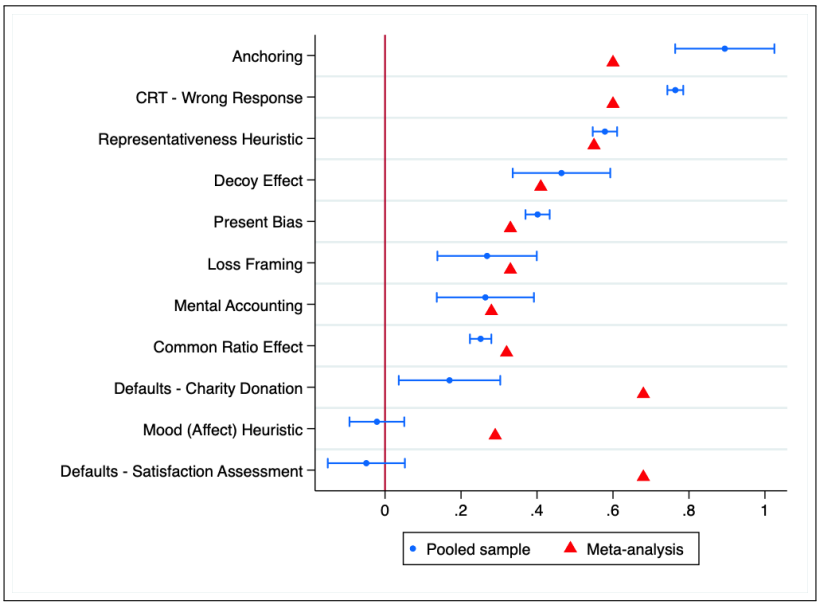
Table 1: Full Regression Results - All Measures

	Panel A - Within Subjects' Measures					
	(1)	(2)	(3)	(4)	(5)	(6)
	Overall Mean		Low-income Mean	Student Mean Difference		
Representativeness	0.579***		0.584***	-0.011		
	(0.022)		(0.037)	(0.044)		
CRT - Wrong Response	0.764***		0.929***	-0.329***		
	(0.026)		(0.008)	(0.019)		
Common Ratio	0.252***		0.140***	0.223***		
	(0.024)		(0.019)	(0.034)		
Present Bias	0.402***		0.403***	-0.002		
	(0.019)		(0.026)	(0.038)		
	Panel B - Between Subjects' Measures					
	(1)	(2)	(3)	(4)	(5)	(6)
	Overall Treatment Effect	Control Group Mean	Low-income Treatment Effect	Student Effect Difference	Low-income Control Group Mean	Student Control Group Mean Difference
Anchoring	0.894***	1.024	0.895***	-0.001	1.024	0.182**
	(0.058)	(0.051)	(0.088)	(0.116)	(0.059)	(0.075)
Loss Framing	0.269***	0.894	0.185	0.164	0.929	-0.168
	(0.086)	(0.065)	(0.137)	(0.173)	(0.070)	(0.102)
Mental Accounting	0.264***	0.579	0.055	0.417**	0.680	0.153*
	(0.083)	(0.061)	(0.116)	(0.156)	(0.058)	(0.087)
Defaults - Charity	0.170**	1.400	0.173	-0.006	1.398	-0.030
	(0.082)	(0.091)	(0.146)	(0.169)	(0.128)	(0.141)
Defaults - Assessment	-0.049	3.646	-0.109	0.114	3.667	-1.379***
	(0.062)	(0.035)	(0.074)	(0.121)	(0.032)	(0.089)
Decoy Effect	0.464***	1.627	0.472***	-0.014	1.624	-0.339***
	(0.062)	(0.058)	(0.087)	(0.125)	(0.068)	(0.085)
Mood Heuristic	-0.021	3.495	-0.104	0.104	3.460	-0.036
	(0.048)	(0.063)	(0.069)	(0.096)	(0.086)	(0.140)

Panel A shows the sample means, and mean differences between the Low-income and Student sub-sample for the within-subject bias measures, described in each row. Column 1 represents the mean for the full sample, with standard errors of the means in parentheses, while column 3 represents the same for the Low-income sample specifically, and column 4 for the mean difference between the Low-Income and Student samples (alternatively, columns 3 and 4 can be measured using β_0 and β_1 from Equation 1). Panel B represents output from regressions of the outcomes for all between-subjects biases from Equations 2 and 3. All regressions include a dummy for 'Student', but no other control variables. This table is reproduced in the Appendix but including a set of control variables. Column 1 represents the treatment effect from Equation 2; column 3:6 represents results from Equation 3. Specifically, column 3 represents β_2, column 4 represents β_3, while columns 5 and 6 represent β_0 and β_1 respectively. Standard errors of coefficients are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Standard errors are clustered at the session level (the level of randomization).

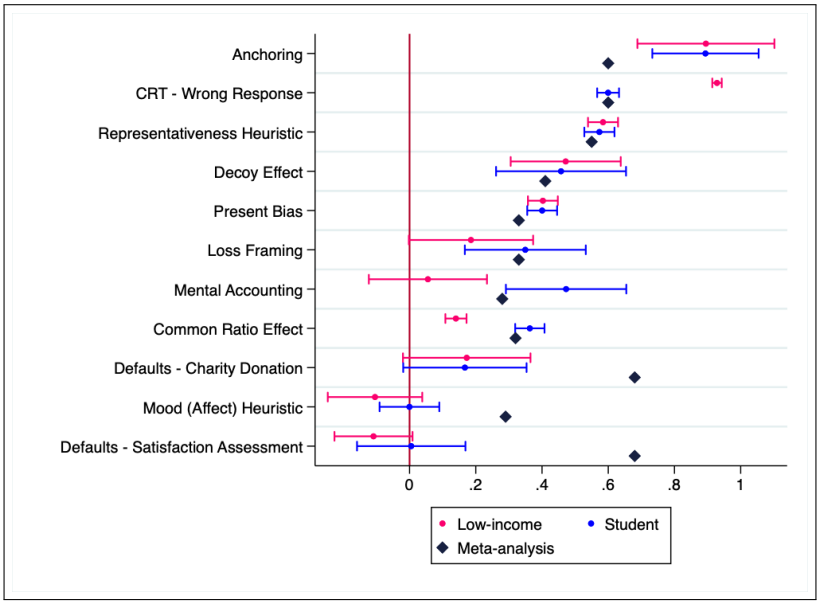
Overall, there is clear evidence for the existence of most of the behavioral biases in the pooled Indian sample - 9 out of the 11 measures show significant effects in the expected direction (Figure 2; Table 1, Column 1, Panels A and B), 8 of which have a p-value less than 0.01, and 6 of which have a p-value less than 0.001. The exceptions are the "Mood Heuristic" and the default selection for assessment. Figure 1 illustrates the treatment effect sizes, and provides a comparison with effect sizes from meta-analyses for

Figure 1: Treatment Effect Sizes: Comparison with Meta-analyses Effect Sizes



Note: Blue points represent effect sizes from Equation 2 for the main effect for each measure, and blue lines represent the 95% confidence intervals for these estimates. Notably, all effects except for Anchoring are measured in proportional terms (ie. the proportion of the sample showing the bias, or the average proportional bias for the 'Defaults - Assessment'). The Anchoring Effect represents the coefficient of the anchor itself on the willingness to pay (also reasonably scalable between 0 and 1; the interpretation allows a value of 1, meaning the anchor perfectly explains the willingness to pay, and 0, that the anchor does not explain any of the willingness to pay). In all cases, the red triangle represents the effect sizes in the same units from meta-analyses conducted with these measures.

Figure 2: Treatment Effect Sizes by Sub-sample: Comparison with Meta-analyses



Note: Blue and pink points represent effect sizes from Equation 2, run separately for each sub-sample (Student and low-income samples respectively), for the main effect for each measure, and blue and pink lines represent the 95%a confidence intervals for these estimates. Notably, all effects except for Anchoring are measured in proportional terms (ie. the proportion of the sample showing the bias, or the average proportional bias for the 'Defaults - Assessment'). The Anchoring Effect represents the coefficient of the anchor itself on the willingness to pay (also reasonably scalable between 0 and 1; the interpretation allows a value of 1, meaning the anchor perfectly explains the willingness to pay, and 0, that the anchor does not explain any of the willingness to pay). In all cases, the black diamond represents the effect sizes in the same units from meta-analyses conducted with these measures. These are the same meta-analysis effect sizes as in Figure 1.

each bias amongst almost entirely Western samples[9]. For 4 of the 11 measures, the meta-analytic effect size is within the 95% confidence interval of the pooled effect size in this study, while for 6 measures we cannot reject that the pooled effect size and the meta-analytic effect size are the same. For those that differ significantly, 2 measures in the present study have a larger effect size than the meta-analysis effects and 3 are significantly lower. Thus, the effects in the pooled sample both largely exist and are mostly consistent with existing evidence.

Second, the patterns of bias are predominantly the same across both the student and low-income sub-samples. Figure 2 depicts treatment effect sizes by sub-sample and compares them with effect sizes from meta-analyses. There is a significant difference in the effect estimate for only 3 of the 11 tests between students and the low-income sample (Table 1, Column 5, Panels A and B), and in 2 of these cases the bias is still significant in both sub-samples in the same direction, but simply more pronounced in one than the other (for the Cognitive Reflection Test and for the Common Ratio Effect). Moreover, there are no measures where the effects are differently signed, and in 8 of the 11 measures, both sub-samples show an effect that is significant at least at the 10% level.

Third, there are notable differences in biases between each sub-sample for 3 of the measures. The low-income sample scored significantly lower for the Cognitive Reflection Test (CRT); this is not unexpected given the correlation between years of education and CRT scores (Frederick et al., 2005). More interestingly, the student sample displayed significantly larger biases in both the Common Ratio Effect and in the measure of Mental Accounting. For Mental Accounting, specifically, the low-income sample showed no evidence of applying mental accounts ($p > 0.5$) while the student-sample showed a highly significant effect ($p < 0.001$), more similar to that found in previous studies.

2.2 Effect Heterogeneity

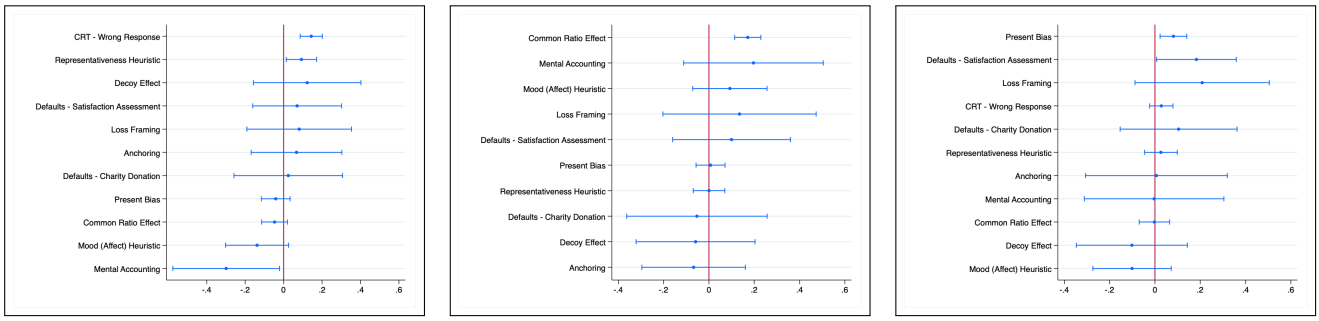
The main analysis shows that biases are largely the same across the student and low-income sub-samples, but differ in notable ways for some measures. To further test the stability of these effects, and to explore possible explanations for their variation between sub-samples, we test effect heterogeneity across additional covariates. For the pooled sample, we specifically explore differences in effects by gender (male and female participants), CRT score (low and high performers, median split), and "Cognitive Style" (those defined as having "relational" compared with "categorical" thinking styles (Nisbett et al., 2001), commonly used to distinguish between Western and non-Western cognition[10]).[11].

[9]This is intended to benchmark the effects for each sub-sample. Meta-analyses include almost entirely Western samples except for the CRT, Present Bias, Common Ratio Effect and Mental Accounting. Brañas-Garza et al. (2019) conducted a meta-study of the CRT, which includes studies from Argentina, Brazil, and Colombia. Ashraf et al. (2006) use a field experiment in the Philippines to elicit time preferences and the review of Common Ratio Effects by Blavatskyy et al. (2023) includes a study in South Africa. Priolo et al. (2023) conducted a series of Mental Accounting tests across 21 countries, including Brazil, Egypt, India and Indonesia. We remove these observations in this final study to observe the meta-analytic effect amongst Western samples, alone. Overall, non-Western samples account for less than 1% of the observations across meta-analyses. See Appendix 1 for more details regarding meta-analyses.

[10]Participants are classified using their responses to the "triad task" from Ji et al. (2004), an established measure in cultural psychology used to distinguish between "Western" compared with "Eastern" styles of cognition (ie. people who classify objects as being conceptually linked based on their relationships in context as compared to classifying them according to their categorical or taxonomic similarities).

[11]These dimensions are plausibly strong moderators of responses (Niederle and Vesterlund, 2007; Talhelm et al., 2014; Toplak et al., 2013) and are highly relevant to behavioral biases in this context. Specifically, the CRT measures automatic

Figure 3: Heterogeneity in Biases for the Pooled Sample



(a) Female (Gender) Effect Difference

(b) CRT High Scorer Effect Difference

(c) Relational Cognitive Style Effect Difference

Note: The above graphs show the effect difference across each binary category. In panel a, we show the effect difference for female participants (compared with male), in panel b) the effect difference for those scoring median or above for the CRT (compared to those below median), and in panel c) the effect difference for those with the more "relational" analytic style (associated with non-Western cognition) compared with more categorical. For within subject biases, this is estimated using model 1) and including a dummy variable for each category. For between-subjects' biases we include an interaction term and a level term of each dummy variable with the treatment variable. Lines represent the 95%a confidence intervals for these estimates.

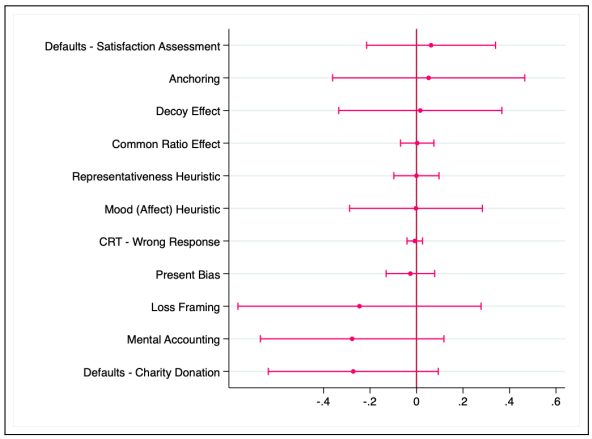
Overall, we find limited variation in effects across these dimensions. This suggests that these biases are relatively stable across different types of participants. For the CRT, high and low scorers do not differ significantly for 9 of the 10 biases (excluding the CRT itself). This is contrary to evidence that suggests that the CRT is highly predictive of several decision heuristics (Juanchich et al., 2016; Toplak et al., 2013). The Common Ratio Effect is an exception: participants in this study with a higher CRT score are significantly *more likely* to display the Common Ratio Effect. This is consistent with the finding in this paper that students are significantly more likely to display this effect; however, the CRT moderates the effect even within each sub-sample (Appendix Table A2). Participants with a more "relational" (non-Western) cognitive style do not show a different effect for 10 of the 11 biases, but are slightly more present-biased. This suggests that a Western style of thinking largely does not moderate these biases. Finally, gender predicts differences in the CRT score and the Representativeness Heuristic, such that female participants perform worse on the CRT (consistent with previous work: Brañas-Garza et al. (2019)) and apply representativeness more often. However, male participants are more likely to apply Mental Accounting. For the remaining 8 biases, there is no significant difference by gender.

In addition to the above dimensions for the pooled sample, we test for heterogeneity within the low-income sample across both age (whether participants are older than 25 or not, a median split) and education (whether the participant completed high school or not, a median split). As discussed, most original effects and replications are established with Western student samples: age and education therefore vary little in the existing literature, but could moderate effects (Henrich et al., 2010c). This might also help explain the few differences we observe between students and the low-income sample in this study.

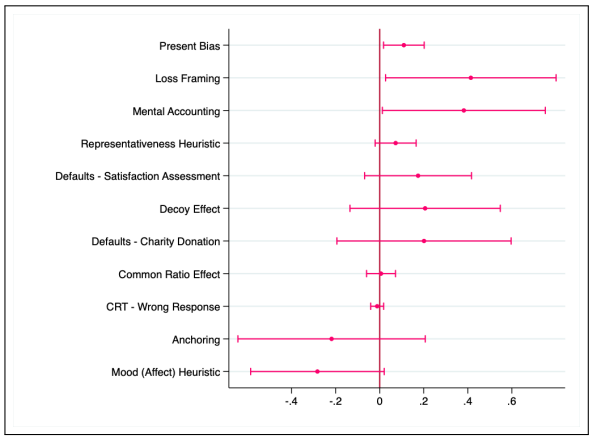
Low-income participants over the age of 25 do not display any significant differences in effects across any

compared to reflective thinking, which is thought to underly a large range of biases (Toplak et al., 2013; Kahneman, 2011). For cognitive style, given large differences between "WEIRD" and non-WEIRD populations, this could moderate effects that are established on Western preferences and beliefs. Finally, gender continues to be a highly relevant social dimension of behavior, particularly in India (Lowe and McKelway, 2021; McKelway, 2018). We also have a reasonable amount of variation within both student and low-income sub-samples on each dimension, allowing this to be a useful for of heterogeneity to measure.

Figure 4: Heterogeneity in Biases for the Low-income Sample



(a) Older than 25
Effect Difference



(b) Finished High School
Effect Difference

Note: The above graphs show the effect difference across each binary category for the low-income sample only. In panel a, we show the effect difference for participants over the age of 25 years (compared with those 25 years and younger), in panel b) the effect difference for those who have finished high school (compared to those who have not) For within subject biases, this is estimated using model 1) and including a dummy variable for each category (exclusively on the low-income sample). For between-subjects' biases we include an interaction term and a level term of each dummy variable with the treatment variable. Lines represent the 95%a confidence intervals for these estimates.

measure compared to those younger than 25 years. While these comparisons suffer from power limitations, as they test differences in effects within a smaller sample, the point estimates for the interaction are very close to zero for most estimates (Figure 4a). Age thus does not appear to moderate these effects. Educational attainment, however, shows a slightly different pattern. While there are only significant differences by education for 3 of the 11 biases, more educated participants display a higher bias in each case, and have point estimates that are higher for 8 of the 11 effects. For a composite measure of the effects, more educated participants in the low-income sample are significantly more biased overall ($p < 0.01$). Notably, they are significantly more biased for Mental Accounting, consistent with the finding that the student sample applies Mental Accounting more than the low-income sample.

2.3 Relationships Between Effects

Given these biases are often explained as the result of the application of decision-making heuristics, it follows that some people might be more prone to applying heuristics across a range of decisions than others (some people may be systematically more 'biased' than others). Moreover, certain biases might be the result of similar heuristic processes, while others might be the result of more distinct ones. We therefore test for relationships between biases. For biases from within-subjects' measures, we simply analyze correlations between each measure. For biases drawn from between-subjects' measures, we use an interaction between individual within-subject responses (biases measured within subject) and treatment status in between subject measures to test whether within-subjects' bias predicts between subjects' bias. We do not measure the relationship between biases that use between-subjects' measures. The results are presented in Table 2. Overall, there is surprisingly little relationship between different biases in our samples. Across 45 tests of pairwise relationships between biases, only 5 are significant at the 5%

level. This suggests that these biases are not necessarily capturing subjects' general cognitive tendencies or general inattentiveness, nor can they be explained by some underlying heuristic process. To further explore more general relationships between biases, we create an overall bias measure by summing the Representativeness Heuristic, Common Ratio Effect, Present Bias, and CRT Intuitive scores for each participant, and then interacting it with treatment for each between-subjects bias measure in the same way. This measure, too, does not meaningfully predict any other individual bias (Table 3, Column 6).

Perhaps somewhat surprising is that the CRT does not predict many biases, though it is often considered as a proxy for more general 'automatic' thinking. However, some biases are correlated. There is a significant negative association between the Common Ratio Effect and both incorrect CRT responding and Intuitive CRT responding (choosing the intuitive but wrong answer in questions in the CRT). Hence, the identified bias in the Common Ratio Effect, which is to display inconsistent risk preferences across question framings, is driven by individuals with *higher* CRT scores (consistent with the previous section).

Another exception is a strong significant relationship between the Representativeness Heuristic and Mental Accounting. This indicates that individuals who make wrong probability judgements due to how 'representative' some description feels, are more likely to use 'mental accounts' when making purchasing decisions (and therefore to treat money as non-fungible).

Finally, we found that those who respond more intuitively to the CRT are more prone to defaulting to the charity donation option. This relationship alone corresponds with the literature arguing that the CRT measures a proclivity towards "automatic" behaviors.

Table 2: Relationships between Biases

	CRT Wrong	CRT Intuitive	Representativeness Heuristic	Common Ratio Effect	Present Bias	Total Bias
CRT Wrong	1					
CRTIntuitive	0.660***	1				
Representativeness Heuristic	-0.013	-0.031	1			
Common Ratio Effect	-0.202***	-0.143***	0.045	1		
Present Bias	0.009	0.056*	0.043	0.044	1	
Anchoring	-0.002	0.229	0.045	0.104	0.015	0.059
	(0.150)	(0.175)	(0.130)	(0.161)	(0.112)	(0.047)
Loss framing	0.015	0.388*	0.190	0.148	-0.024	0.126
	(0.253)	(0.213)	(0.155)	(0.161)	(0.123)	(0.079)
Mental Accounting	-0.163	-0.217	0.286***	0.177	-0.107	0.074
	(0.236)	(0.219)	(0.103)	(0.140)	(0.121)	(0.047)
Defaults - Satisfaction Assessment	-0.187	-0.116	0.073	-0.091	0.015	-0.025
	(0.195)	(0.141)	(0.114)	(0.118)	(0.090)	(0.054)
Defaults - Charity Donation	0.189	0.712***	0.035	-0.109	0.166	0.112
	(0.229)	(0.208)	(0.149)	(0.128)	(0.177)	(0.088)
Decoy Effect	0.115	0.231	0.201	0.052	-0.204	0.013
	(0.220)	(0.202)	(0.132)	(0.188)	(0.127)	(0.073)
Mood Heuristic	-0.218*	-0.173*	-0.078	0.006	-0.074	-0.046
	(0.091)	(0.065)	(0.064)	(0.189)	(0.077)	(0.031)

The top panel reports Pearson's correlations between within subjects' measures of each bias. For each of these measures, we are able to identify individual bias measures. These estimates thus provide measures of the relationships between within-subjects' biases. The second panel reports coefficients from regressions of each primary outcome for each between-subject measure, using Equation 2, but with the addition of an interaction term between the individual within-subject bias measure (ie. the bias measure for CRT Wrong, CRT Intuitive, Representativeness Heuristic, Common Ratio Effect, Present Bias) and the treatment variable for the relevant between-subject measure. Each regression is run separately, dealing with the interaction of each individual bias measure (shown in the columns) and each between-subject treatment variable (represented by each row) separately. Specifically, we report the coefficient on the interaction term. This coefficient allows us to identify whether those displaying the bias in the within-subjects' measures are more or less likely to display the between subjects' biases. Standard errors are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Standard errors are clustered at the session level (the level of randomization).

3 Discussion and Conclusion

Overall, our findings suggest that our samples of interest - low-income Indians likely to be targeted by behavioral development policy, and elite Indian university students - largely demonstrate the selected behavioral biases from the behavioral economics literature. Of the 11 contrasts tested in this paper, 9 were significant for the pooled sample in the predicted direction. For 6 of the 9 contrasts where we observe an effect, we fail to reject that this effect is different from those from meta-analyses or large-scale replications (with Western samples)[12]. This suggests that these biases are not exclusively the outcome of decision-making patterns of "WEIRD" participants, but likely apply to non-Western populations, too. Moreover, we find that the pattern of results is largely the same for elite university students and low-income people living in the same Indian state. For 8 of the 11 measures, we cannot reject that the effect is the same for the low-income and student samples.

To discuss some of our specific results, for instance, the proportion of our pooled sample showing present-biased time preferences over money (40%) is consistent with the seminal literature on this phenomenon, that finds that between 35% and 45% of typical samples are present-biased (Cohen et al., 2020)[13]. Both students and low-income samples were identical with respect to the prevalence of Present Bias. While the effect of "Defaults" on choice were identical across our sub-samples (Cohen's D=0.2), they were substantially lower than Default Effects in seminal default studies (Cohen's D=0.9) and in meta-analyses (Cohen's D=0.68; (Johnson and Goldstein, 2003; Jachimowicz and Johnson, 2019)). However, in contrast to the other biases tested in this paper (which use relatively uniform survey measures tested in a consistent manner in decision laboratory environments), defaults in the meta-analysis and in seminal studies include a diverse range of default applications across both field and lab studies (Blumenstock et al., 2018; Jachimowicz and Johnson, 2019). In our application[14], the alternative option was highly visible and nearly costless to select (participants merely needed to select the non-default option on their computer tablet). Given this low cost to reverse the default choice, we hesitate to claim that our sample is specifically less susceptible to Default Effects than others[15]. Finally, the Anchoring Effect in our study is large and identical across sub-samples (Cohen's D= 0.89), and larger than those found in other studies. The remainder of the pooled effects in this study are very similar to meta-analytic results, with the exception of the "Mood Heuristic", which fails to replicate in this study[16].

In spite of the lack of systematic differences between our elite student and low-income samples, there are significant and relatively large differences in three effects: the cognitive reflection test (CRT), Mental Accounting, and the Common Ratio Effect. For the CRT, a test of whether participants apply reflective thinking when engaging seemingly simple math problems, the low-income sample performed significantly worse than the elite student sample, completing 19% of questions correctly on average compared with

[12]This included the Representativeness Heuristic, Decoy Effect, Present-Bias, Loss Framing, Mental Accounting, and the Common Ration Effect

[13]That is, participants who are present-biased discount future monetary payoffs more strongly when the choice is between payoffs *now* and the future compared with payoffs between two time periods in the future

[14]A default selection of a charity to which the experimenters would donate money on participants' behalf

[15]This is in contrast to some defaults in the meta-analysis and seminal papers that, for example, include default enrollment in savings and organ donation programs that require considerable energy and cost to change, such as completing multiple forms and travelling to and queuing in public administration offices

[16]Please see the Appendix for each individual measure's result in the original units.

43% for the elite student sample ($p < 0.001$). For context, our Indian student sample displayed nearly identical results to that for student samples in both the original study (41% correct; Frederick (2005)) and in meta-analytic results (40% correct; Brañas-Garza et al. (2019)). Given well-documented correlations between cognitive ability and the CRT (Obrecht et al., 2009; Frederick, 2005), and between cognitive ability and socioeconomic status (Capron and Duyme, 1989), the difference between our sub-samples on this measure is largely unsurprising. However, the differences in effects for the Common Ratio Effect and Mental Accounting are arguably more interesting. For both measures, elite student sample effects are similar to estimates from Western samples. Low-income participants, however, are significantly less likely to show Mental Accounting and the Common Ratio Effect compared to both the Indian student sample and to meta-analysis effects. For Mental Accounting, specifically, the point estimate for the low-income group is not statistically different from zero[17]. This is consistent with a recent large-scale replication of Mental Accounting effects, that generally replicated effects across contexts (including with Global South participants) but found that the effect was significantly *lower* for low-income participants, less educated participants, and participants with lower financial literacy (Priolo et al., 2023). Olsen et al. (2019) and Muehlbacher and Kirchler (2019) also find a positive correlation between Mental Accounting and income, while Millroth et al. (2019) observe that those with lower financial literacy are less likely to show several Prospect Theory biases (including the Common Ratio Effect).

How can we reconcile these findings? Recent evidence suggests that people with lower incomes are plausibly less susceptible to framing effects for financial decisions (Mullainathan and Shafir, 2013; Shah et al., 2015). Across a range of studies, Shah et al. (2015) show that people under financial scarcity, compared with those with more wealth, are less likely to change their purchasing decisions when the frame of the choice changes. Similarly, poor consumers pay more attention to variations in hidden taxes and adjust consumption more consistently with a rational benchmark than higher income consumers (Kroft et al., 2020). Taken together, it appears that people with lower-incomes and less education, both in our sample and the literature, are less prone to several biases in financial decision-making. Shah et al. (2015) propose that financial scarcity focuses attention towards meeting immediate financial needs, which reduces the effect of framing and generates financial decisions that align more closely with a rational benchmark. Though we find heterogeneity consistent with this theory in our data, we cannot provide support for a causal interpretation. Moreover, we do not find the same pattern of heterogeneity across all measures that apply framing differences to financial decisions: we see no difference in effects for our measure of the Decoy Effect. Future work should examine these relationships further, for example by testing whether exogenous shocks to financial scarcity affect these biases, or by testing the effect of financial literacy training on these measures.

In addition to exploring important heterogeneity in several biases, this study adds to recent replication efforts in the behavioral sciences. For instance, Camerer et al. (2016) successfully replicate 11 out of 18 (61%) behavioral economics experimental effects published in prominent economics journals between

[17]The result for Mental Accounting for the student sample is in contrast to a recent study that uses the exact same measure amongst an elite Indian student sample (Banerjee et al., 2019) and finds no Mental Accounting. However, the sample in Banerjee et al. (2019) is 67, far less than our student sample of 457, and their use of inference to confirm a null hypothesis is not appropriate. By contrast, the effect for the student sample in the present study is highly significant ($p < 0.001$) and robust.

2011 and 2014. Klein et al. (2018) run a large replication project across multiple countries, including a number in the Global South, to try reproduce several seminal effects in social psychology. They find a lower replication rate (approximately 50%) than our study, but similarly find little heterogeneity between WEIRD and "less-WEIRD" (including middle-income countries) sub-samples. More similar to our work, Ruggeri et al. (2020) and Priolo et al. (2023) attempt to replicate the core contrasts of Prospect Theory and forms of Mental Accounting, respectively, across several countries (using similar measures to those used in this paper). Consistent with the findings in the present paper, both studies find that effects replicate at a very high rate [18] with effect sizes similar to those documented in the literature. Priolo et al. (2023), specifically, find evidence of Mental Accounting amongst their subset of countries from the Global South. These studies, like ours, also find limited heterogeneity in these effects. Our paper builds on these by specifically demonstrating that these biases replicate with a "harder to reach" low-income sample. Our study uses a controlled decision laboratory, contextualized measures, and several safeguards (eg., translation and back-translation of instructions into multiple languages, literacy tests, comprehension checks, and audibly read-out and repeated instructions) to elicit high fidelity responses from this underrepresented group. Our paper also builds on this work by testing intra-country heterogeneity in effects with large per-site samples; this also allows us to test intra-group heterogeneity across other covariates. Consistent with this previous work, we ultimately find limited effect heterogeneity.

Overall, this study shows that several behavioral economics biases exist amongst two highly distinct non-WEIRD populations. This suggests that these biases might generalize across socioeconomic groups. However, this study is not sufficient to show that these effects generalize to poor and socioeconomic elites across all contexts. More work is needed to expand findings to other "hard to reach" groups, particularly low-income groups in different parts of the world. We provide a possible methodology to conduct more theoretical "laboratory-style" research in such low-income contexts. Finally, although the effects were remarkably stable across different covariates, the heterogeneity in both risk preference-related measures (the Common Ratio Effect) and Mental Accounting suggest space for future theorizing. Given that these effects are also relevant to important household behaviors (Dupas and Robinson, 2013; Mahapatra et al., 2022), this and related work should also inform future applications of behavioral insights to policy.

[18] Ruggeri et al. (2020) find a 94% replication rate, while (Priolo et al., 2023) find a 100% replication rate

References

Jeffrey J Arnett. The neglected 95%: why american psychology needs to become less american. 2016.

Nava Ashraf, Dean Karlan, and Wesley Yin. Tying odysseus to the mast: Evidence from a commitment savings product in the philippines. *The Quarterly Journal of Economics*, 121(2):635–672, 2006.

Bence Bago, Marton Kovacs, John Protzko, Tamas Nagy, Zoltan Kekecs, Bence Palfi, Matus Adamkovic, Sylwia Adamus, Sumaya Albalooshi, Nihan Albayrak-Aydemir, et al. Situational factors shape moral judgements in the trolley dilemma in eastern, southern and western countries in a culturally diverse sample. *Nature human behaviour*, 6(6):880–895, 2022.

Pronobesh Banerjee, Promothesh Chatterjee, Sanjay Mishra, and Anubhav A Mishra. Loss is a loss, why categorize it? mental accounting across cultures. *Journal of Consumer Behaviour*, 18(2):77–88, 2019.

Nicholas C Barberis. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–196, 2013.

H Clark Barrett, Alexander Bolyanatz, Alyssa N Crittenden, Daniel MT Fessler, Simon Fitzpatrick, Michael Gurven, Joseph Henrich, Martin Kanovsky, Geoff Kushnick, Anne Pisor, et al. Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, 113(17):4688–4693, 2016.

Pavlo Blavatskyy, Valentyn Panchenko, and Andreas Ortmann. How common is the common-ratio effect? *Experimental Economics*, 26(2):253–272, 2023.

Joshua Blumenstock, Michael Callen, and Tarek Ghani. Why do defaults affect behavior? experimental evidence from afghanistan. *American Economic Review*, 108(10):2868–2901, 2018.

Pablo Brañas-Garza, Praveen Kujal, and Balint Lenkei. Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics*, 82:101455, 2019.

Christopher J Bryan, Elizabeth Tipton, and David S Yeager. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, 5(8):980–989, 2021.

Colin F Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 2016.

Christiane Capron and Michel Duyme. Assessment of effects of socio-economic status on iq in a full cross-fostering study. *Nature*, 340(6234):552–554, 1989.

Jonathan Cohen, Keith Marzilli Ericson, David Laibson, and John Myles White. Measuring time preferences. *Journal of Economic Literature*, 58(2):299–347, 2020.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251): aac4716, 2015.

Alan Coman, Johannes Haushofer, Chaning Jang, Elizabeth Levy Paluck, and Oliver Schweickart. Behavioral biases across cultures: Evidence from kenya and the usa. *Unpublished manuscript*, 2017.

Pascaline Dupas and Jonathan Robinson. Why don't the poor save more? evidence from health savings experiments. *American Economic Review*, 103(4):1138–1171, 2013.

Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692, 2018.

Ernst Fehr and Urs Fischbacher. Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. *The economic journal*, 112(478):C1–C33, 2002.

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

Joseph Henrich, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, et al. Markets, religion, community size, and the evolution of fairness and punishment. *science*, 327(5972):1480–1484, 2010a.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. Most people are not weird. *Nature*, 466(7302): 29–29, 2010b.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010c.

Joel Huber, John W Payne, and Christopher Puto. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research*, 9(1):90–98, 1982.

Shannon Duncan Elke U. Weber Jachimowicz, Jon M. and Eric J. Johnson. When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, 3(2):159–186, 2019.

Li-Jun Ji, Zhiyong Zhang, and Richard E Nisbett. Is it culture or is it language? examination of language effects in cross-cultural research on categorization. *Journal of personality and social psychology*, 87(1): 57, 2004.

Eric J Johnson and Daniel Goldstein. Do defaults save lives?, 2003.

Marie Juanchich, Chris Dewberry, Miroslav Sirota, and Sunitha Narendran. Cognitive reflection predicts real-life decision outcomes, but not over and above personality and decision-making styles. *Journal of Behavioral Decision Making*, 29(1):52–59, 2016.

Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

Kory Kroft, Jean-William P Laliberté, René Leal-Vizcaíno, and Matthew J Notowidigdo. Salience and taxation with imperfect competition. Technical report, National Bureau of Economic Research, 2020.

Matt Lowe and Madeline McKelway. Coupling labor supply decisions: An experiment in india. 2021.

Mousumi Singha Mahapatra, Jayasree Raveendran, and Ram Kumar Mishra. Role of mental accounting in personal financial planning: A study among indian households. *Psychological Studies*, 67(4):568–582, 2022.

Lauren Manning, Abigail Goodnow Dalton, Zeina Afif, Renos Vakos, and Faisal Naru. Behavioral science around the world volume ii: Profiles of 17 international organizations. *eMBeD report. Washington, D.C.: World Bank Group*, 2020.

Madeline McKelway. Women's self-efficacy and women's employment: Experimental evidence from india. *Unpublished Working Paper*, 2018.

Blakeley B McShane, Jennifer L Tackett, Ulf Böckenholt, and Andrew Gelman. Large-scale replication projects in contemporary psychological research. *The American Statistician*, 73(sup1):99–105, 2019.

Philip Millroth, Håkan Nilsson, and Peter Juslin. The decision paradoxes motivating prospect theory: The prevalence of the paradoxes increases with numerical ability. *Judgment and decision making*, 14 (4):513–533, 2019.

Stephan Muehlbacher and Erich Kirchler. Individual differences in mental accounting. *Frontiers in Psychology*, 10, 2019.

Sendhil Mullainathan and Eldar Shafir. *Scarcity: Why having too little means so much*. Macmillan, 2013.

Jinkyung Na, Igor Grossmann, Michael EW Varnum, Shinobu Kitayama, Richard Gonzalez, and Richard E Nisbett. Cultural differences are not always reducible to individual differences. *Proceedings of the National Academy of Sciences*, 107(14):6192–6197, 2010.

Muriel Niederle and Lise Vesterlund. Do women shy away from competition? do men compete too much? *The quarterly journal of economics*, 122(3):1067–1101, 2007.

Richard E Nisbett, Kaiping Peng, Incheol Choi, and Ara Norenzayan. Culture and systems of thought: holistic versus analytic cognition. *Psychological review*, 108(2):291, 2001.

Natalie A Obrecht, Gretchen B Chapman, and Rochel Gelman. An encounter frequency account of how experience affects likelihood estimation. *Memory & cognition*, 37(5):632–643, 2009.

OECD. *Behavioural Insights and Public Policy*. 2017. doi: https://doi.org/https://doi.org/10.1787/9789264270480-en. URL https://www.oecd-ilibrary.org/content/publication/9789264270480-en.

Jerome Olsen, Matthias Kasper, Christoph Kogler, Stephan Muehlbacher, and Erich Kirchler. Mental accounting of income tax and value added tax among self-employed business owners. *Journal of Economic Psychology*, 70:125–139, 2019.

Giulia Priolo, Stablum Federica, Vacondio Martina, D'Ambrogio Simone, Caserotti Marta, Conte Beatrice, and De Roni Prisca. The robustness of mental accounting: a global perspective. *Science advances*, 2023.

Mostafa Salari Rad, Alison Jane Martingano, and Jeremy Ginges. Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, 115(45):11401–11405, 2018.

Kai Ruggeri, Sonia Alí, Mari Louise Berge, Giulia Bertoldo, Ludvig D Bjørndal, Anna Cortijos-Bernabeu, Clair Davison, Emir Demić, Celia Esteban-Serna, Maja Friedemann, et al. Replicating patterns of prospect theory for decision under risk. *Nature human behaviour*, 4(6):622–633, 2020.

Anuj K Shah, Eldar Shafir, and Sendhil Mullainathan. Scarcity frames value. *Psychological science*, 26 (4):402–412, 2015.

Fritz Strack, Leonard L Martin, and Norbert Schwarz. Priming and communication: Social determinants of information use in judgments of life satisfaction. *European journal of social psychology*, 18(5):429–442, 1988.

Thomas Talhelm, Xuemin Zhang, Shigehiro Oishi, Chen Shimin, Dongyuan Duan, Xuezhao Lan, and Shinobu Kitayama. Large-scale psychological differences within china explained by rice versus wheat agriculture. *Science*, 344(6184):603–608, 2014.

Richard Thaler. Mental accounting and consumer choice. *Marketing science*, 4(3):199–214, 1985.

Richard H Thaler. *Misbehaving: The making of behavioral economics*. WW Norton Company, 2015.

Richard H Thaler. From cashews to nudges: The evolution of behavioral economics. *American Economic Review*, 108(6):1265–87, 2018.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691, 2005.

Maggie E Toplak, Richard F West, and Keith E Stanovich. Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of child psychology and psychiatry*, 54(2):131–143, 2013.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185 (4157):1124–1131, 1974.

Appendix

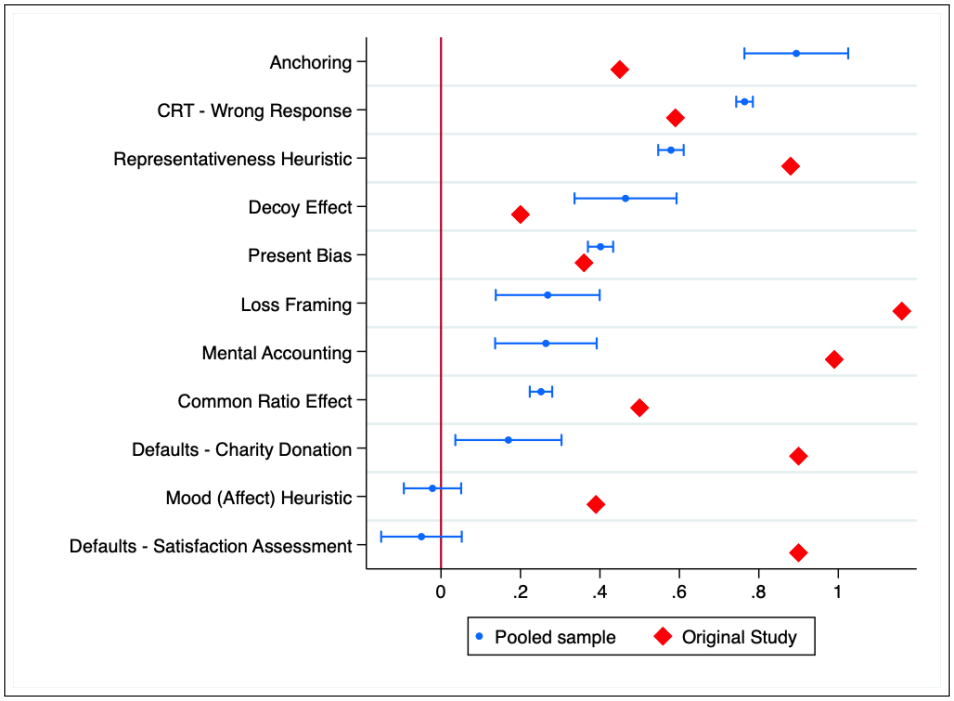
Replication package available from authors upon request.

1.1 Contextualization example

To illustrate the contextualization process, the 'Linda Problem', an original measure of the 'Representativeness Heuristic' (Tversky & Kahneman, 1974), was identified and contextualized (see Appendix 1 for both versions). The original measure relies on research subjects understanding culturally-embedded characteristics of a young woman's profile such that certain selection options, even though statistically less likely due to the conjunction fallacy, appear more likely because they are 'representative' of the target. Thus, subjects should be cued that Linda is a certain 'type' of person that they are familiar with and should therefore have somewhat predictable expectations of her likelihood of engaging in different occupations on an accompanying list. Given this, the reference points in the original 'Linda Problem' are distinctly Western. They require understanding of the type of Western woman Linda might be and the likelihood of her engaging in each of a range of typically Western occupations. If the key occupation options do not seem representative due to a lack of knowledge of the context surrounding Linda, then the measure will not identify the conjunction fallacy. In pilot data in Kenya and India, responses to this measure showed signatures of a high degree of noise. Notably, there was a largely uniform distribution of selections, suggesting that subjects chose at random even when ex ante there should be options that clearly dominate others. In debriefing, participants suggested that many of the occupations did not seem 'representative' of Linda even when they were supposed to be. While a conceptual replication would allow the development of a separate measure that included the core elements of representativeness, we attempted to more closely recreate the measure to compare effect sizes to other similar measures. The core elements of the measure that we identified for preservation were: i) a short profile of a recognizable 'type' of person in the Indian context (for comparability, we decided to also use a socially engaged young woman); ii) a list of plausible occupations that could associate with that person; iii) variation in the likelihood of each occupation; iv) two occupations on the list — one that is likely and one that is unlikely - that seem incongruous, but the combination of which increase the 'representativeness'. Within this structure, a profile for 'Preeti' and a list of occupations that an Indian woman 'of this type' could reasonably undertake were developed. See Appendix 1 for the exact measure.

Alternative Effect Size Figures

Figure A1: Treatment Effect Sizes: Comparison with Original Effect Sizes



Note: Blue points represent effect sizes from Equation 2 for the main effect for each measure, and blue lines represent the 95% confidence intervals for these estimates. Notably, all effects except for Anchoring are measured in proportional terms (ie. the proportion of the sample showing the bias, or the average proportional bias for the 'Defaults - Assessment'). The Anchoring Effect represents the coefficient of the anchor itself on the willingness to pay (also reasonably scalable between 0 and 1; the interpretation allows a value of 1, meaning the anchor perfectly explains the willingness to pay, and 0, that the anchor does not explain any of the willingness to pay). In all cases, the red diamond represents the effect sizes in the same units from the original studies (see Appendix 1 for the relevant references). This is intended to benchmark these effects; however, we expect that small sample, originating studies are likely to have higher effect sizes.

Main regression with controls

Table A1: Full Regression Results with controls

	Panel A - Within Subjects' Measures					
	(1)	(2)	(3)	(4)	(5)	(6)
	Overall Mean		Low-income Mean	Student Mean Difference		
Representativeness	0.704**		0.703**	0.000		
	(0.277)		(0.311)	(0.073)		
CRT - Wrong Response	0.375**		0.814***	-0.264***		
	(0.157)		(0.128)	(0.042)		
Common Ratio	0.941***		0.667**	0.164**		
	(0.298)		(0.315)	(0.069)		
Present Bias	0.403		0.403	0.000		
	(0.315)		(0.330)	(0.076)		
	Panel B - Between Subjects' Measures					
	(1)	(2)	(3)	(4)	(5)	(6)
	Overall Treatment Effect	Control Group Mean	Low-income Treatment Effect	Student Effect Difference	Low-income Control Group Mean	Student Control Group Mean Difference
Anchoring	0.906***	-0.564	0.919***	-0.026	-0.577	0.153
	(0.060)	(0.612)	(0.096)	(0.120)	(0.604)	(0.113)
Loss Framing	0.284***	1.644	0.201	0.161	1.700	-0.197
	(0.088)	(0.693)	(0.143)	(0.176)	(0.691)	(0.197)
Mental Accounting	0.262***	0.656	0.039	0.443***	0.760	0.015
	(0.084)	(0.672)	(0.113)	(0.156)	(0.661)	(0.144)
Defaults - Charity	0.176**	2.028	0.195	-0.035	2.010	-0.081
	(0.084)	(0.583)	(0.151)	(0.175)	(0.588)	(0.200)
Defaults - Assessment	-0.054	3.449	-0.107	0.102	3.494	-1.266***
	(0.064)	(0.559)	(0.077)	(0.121)	(0.557)	(0.155)
Decoy Effect	0.471***	1.357	0.490***	-0.038	1.349	-0.485***
	(0.060)	(0.840)	(0.086)	(0.118)	(0.843)	(0.128)
Mood Heuristic	-0.025	4.309	-0.110	0.109	4.234	-0.542***
	(0.046)	(0.820)	(0.066)	(0.093)	(0.838)	(0.130)

Table reproduces Table 1 with a set of control variables including Age, Age square, Female, Finished-School-Only, Finished-Some-College, Finished-Some-University, Took-Psych-Course, Married and Home-tongue. Panel A shows the sample means, and mean differences between the Low-income and Student sub-sample for the within-subject bias measures, described in each row. Column 1 represents the mean for the full sample, while column 4 represents the same for the Low-income sample specifically, and column 5 for the mean difference between the Low-Income and Student samples. Panel B represents output from regressions of the outcomes for all between-subjects biases from Equations 2 and 3 with control variables and a dummy for 'Student'. Standard errors of coefficients in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Standard errors are clustered at the session level (the level of randomization).

Heterogeneity Tables

Table A2: Effect Heterogeneity: Pooled Sample

	Gender			CRT Score			Cognitive Style		
	Pooled Sample	Low-Income Only	Student Only	Pooled Sample	Low-Income Only	Student Only	Pooled Sample	Low-Income Only	Student Only
CRT - Wrong Response	0.144***	0.053**	0.145***				0.028	-0.017	0.067*
	(0.029)	(0.023)	(0.031)				(0.026)	(0.017)	(0.036)
Representativeness Heuristic	0.093**	0.134**	0.057	0.001	-0.042	0.042	0.026	0.13**	-0.076*
	(0.039)	(0.064)	(0.049)	(0.035)	(0.052)	(0.047)	(0.036)	(0.051)	(0.041)
Present Bias	-0.041	-0.042	-0.042	0.007	0.072	-0.033	0.082***	0.071*	0.092**
	(0.037)	(0.058)	(0.053)	(0.032)	(0.064)	(0.055)	(0.029)	(0.037)	(0.046)
Common Ratio Effect	-0.047	0.055	-0.082*	0.171***	0.077	0.088***	-0.002	-0.006	0.005
	(0.033)	(0.034)	(0.047)	(0.029)	(0.053)	(0.033)	(0.033)	(0.041)	(0.05)
Anchoring	0.067	0.175	-0.029	-0.067	-0.073	-0.101	0.006	-0.094	0.111
	(0.117)	(0.184)	(0.145)	(0.114)	(0.221)	(0.133)	(0.156)	(0.265)	(0.174)
Decoy Effect	0.123	-0.046	0.231	-0.059	0.119	-0.208	-0.102	0.051	-0.237
	(0.139)	(0.161)	(0.192)	(0.131)	(0.185)	(0.186)	(0.122)	(0.134)	(0.211)
Mental Accounting	-0.299**	-0.476**	-0.056	0.196	-0.05	0.03	-0.004	0.2	-0.197
	(0.138)	(0.191)	(0.17)	(0.153)	(0.219)	(0.23)	(0.153)	(0.197)	(0.221)
Loss Framing	0.081	0.304**	-0.043	0.135	0.413	-0.134	0.208	0.43**	0.015
	(0.135)	(0.121)	(0.193)	(0.168)	(0.289)	(0.225)	(0.147)	(0.202)	(0.197)
Defaults - Charity	0.024	-0.341*	0.341*	-0.053	0.094	-0.243	0.104	-0.046	0.193
	(0.141)	(0.189)	(0.182)	(0.154)	(0.277)	(0.193)	(0.128)	(0.198)	(0.175)
Defaults - Assessment	0.07	0.056	0.187	0.1	0.083	0.069	0.183**	0.207	0.128
	(0.115)	(0.136)	(0.162)	(0.129)	(0.142)	(0.196)	(0.088)	(0.147)	(0.101)
Mood (Affect) Heuristic	-0.138	-0.014	-0.232**	0.093	0.172	0.107	-0.101	-0.242	-0.042
	(0.084)	(0.142)	(0.104)	(0.083)	(0.152)	(0.113)	(0.088)	(0.169)	(0.107)

Table A3: Effect Heterogeneity: Low-income Sample

	Finished School	Older than 25
CRT - Wrong Response	-0.012	-0.007
	(0.016)	(0.016)
Representativeness Heuristic	0.072	0
	(0.051)	(0.047)
Present Bias	0.11**	-0.026
	(0.043)	(0.051)
Common Ratio Effect	0.006	0.003
	(0.03)	(0.035)
Anchoring	-0.219	0.052
	(0.209)	(0.201)
Decoy Effect	0.206*	0.016
	(0.124)	(0.17)
Mental Accounting	0.382**	-0.277
	(0.191)	(0.192)
Loss Framing	0.414**	-0.245
	(0.176)	(0.254)
Defaults - Charity	0.201	-0.272
	(0.209)	(0.177)
Defaults - Assessment	0.174	0.063
	(0.163)	(0.135)
Mood (Affect) Heuristic	-0.283*	-0.002
	(0.154)	(0.145)