Social Identity and Preferences: A Replication

Nathan Fiala and Victor Volkman

May 14, 2024

Abstract

In a highly cited paper, Benjamin et al. (2010) tested the main contention of self-categorization theory: That increasing the salience of an individual's racial identity will cause them to align their preferences closer to what is expected from them as a member of that racial group. This was done through an experiment that assessed the effect of ethnically charged survey questions on the time and risk preferences of Asian American respondents. Primed Asian respondents were more patient than those in the control group with no effects observed for primed White respondents. However, this experiment was done on a sample of students from very prestigious and even Ivy League universities. We attempt to replicate this study due to concerns that this sample is not properly representative of the general public. Since this was conducted during the COVID-19 pandemic, these experiments were conducted remotely. While we do find the effects of variables such as larger stakes to be similar to what was reported in the original study, we fail to replicate the main result of the paper that Asian Americans chose more patient options and had lower Log-interest rates after treatment. On a sample of the general population, treatment actually increased Log-interest rates for Asian respondents. When we restricted this sample to only students, the sign of the effect was in line with BCS. However, the effect was not significant. This may suggest the results of BCS were indeed due to it using a non-representative sample of convenience.

Keywords: replication, priming, self-categorization theory, risk aversion, communal norms, race, patience, intertemporal choice, online

JEL Codes: D81, J15, J16, and Z13

1 Introduction

The differences in the way social groups set economic priorities has been a curiosity for many economists and social psychologists. Empirical evidence suggests Asian Americans have "high educational mobility" and a greater willingness to invest into the development of human capital when compared to their White peers (Sue & Okazaki (1990), Carroll et al. (1994)). The way racial groups approach financial decisions can also be very different. Asian and White Americans are more likely to invest in tax-deferred savings accounts and participate in 401(k) retirement plans than Black Americans (Springstead & Wilson (2000), Chatterjee & Zahirovich-Herbert (2009)). Sowell (1975) and Chiswick (1983) argue that heterogeneous trends in economic behavior stem from the norms established by communities founded on key social identities, such as race. However, these norms are often malleable to a group's environment and constantly evolving over time (Penner & Saperstein (2009), Helliwell et al. (2015)). It is thus difficult to distinguish the effect of racial norms from those developed from associated factors such as poverty or exploit differences in those norms directly to test this hypothesis(Mood & Jonsson (2015), Boonmanunt et al. (2020)). Benjamin et al. (2010) leveraged the existing literature on "self-categorization theory" and behavioral priming to implement an experiment in which this hypothesis could be tested. This article details our attempt to replicate the results of that paper, and how our results ultimately fail to line up with its findings.

Behavioral priming—the practice of introducing a subject to a subtle stimuli or situational cue and noting the changes in their behavior—is a common tool used in psychology and behavioral economics. It is argued that this allows social scientists to capture effects of environmental factors that are otherwise hard to quantify (Bargh & Chartrand (2000), Cohn et al. (2015), Khadjavi (2014)). However, the robustness of several priming studies have come under

question due to a large number of replications failing to show the same or similar results to the original paper. This "replication crisis"(Cesario (2014), Wiggins & Christopherson (2019)) has been a notable concern in the field of psychology with the Open Science Collaboration's 2015 paper showing that out of 100 tested psychological studies only roughly 39% were deemed replicable (Aarts et al. (2015)). Following this, the Center for Open Science conducted their own wide scale test of replicability in 2018 and found that even charitable estimates of the robustness of psychology studies showed only about half of them could be reproduced(Klein et al. (2018)).

While this "crisis" is most publicized in psychology, replicability is a concern for the field of Economics as well. Economics journals generally tend to de-prioritize replication studies and the rate of original publications in economics massively outpaces the rate of published replications (Camerer et al. (2016), Mueller-Langer et al. (2018)). The replicability of lab experiments in top journals may be particularly suspect. Ioannidis et al. (2017) levies criticism against the power of studies published in 159 of the top empirical economics literatures, finding the median statistical power across these studies is only 18% .Ioannidis (2008) further contends that the bias towards creating publishable results additionally leads a major portion of researchers to overstate their findings across several different disciplines. Explanations for this range from small sample sizes leading to Type S and Type M errors (Gelman & Carlin (2014)) to generally non-representative samples of tested populations. It is also possible that behaviors observed through priming are very specific to a time and place. Methods such as manipulation checks have been introduced to judge the priming efficacy of an intervention (Benjamin et al. (2016)) , but these are often imperfect measures of a treatments impression on a subject.

On the other hand, the variability found in the results of these priming ex-

periments has suggested the need for higher standards in regards to replications.

At the request of the Center for Open Science, we revisited the experiment conducted by Benjamin et al. (2010) (hereafter BCS). The orignal paper found that Asian American students change decisions in several choice games based on being primed on their heritage, while White and Black students do not. The basis for this claim is an experiment conducted with students from Harvard, Temple and the University of Michigan. In the first round the sample included 71 Asian and 66 White respondents, and in the second the sample included 222 White and 124 Black respondents made up of both immigrants and native-born American citizens. The Center for Open Science and the Department of Justice supplied funding for a replication of this paper as a part of their ongoing research into the replicability of studies.

We attempt to replicate these results using a sample of the general US population through an online survey platform, Prodege. Differentiating it from platforms such as mTurk, Prodege charges pre-specified rates per person surveyed and conducts checks of each respondent to reduce inattention issues. Our experiment was conducted in two rounds and included 452 Asian, 460 White and 238 Black individuals.[1] The sample size was chosen based on power calculations, allowing us to report a significant effect given an effect size of 1.81 standard deviations; less than half that of the the main result of the original first experiment. [2] Our protocols and methods, described in the next section, mirror those of BCS as much as possible. We do replicate the findings of BCS in regards to decisions with larger stakes. However, we are unable to replicate any of the original results in regards to the effect of treatment on the Log-interest rates or risk premia of respondents or the percentage of safe or impatient choices

[1] In keeping with the original paper, only the second round incorporated a sample of Black individuals, hence the lack of parity between the three groups. The Asian and White samples are further unequal due to respondents that chose not to finish the experiment.

[2] Our sample size would also allow us to report a significant effect given an effect size half of that of the main result of second experiment

by respondents The only significant effect found on time or risk preferences are an increase in risk seeking in Black men, which conflicts with the findings of the original paper.

We were required to implement this experiment remotely due to the onset of the COVID 19 pandemic. However, the liberties with the design of the original experiment were not taken lightly. Decisions made a remote context can naturally differ considerably from those made in person, as has been noted in past experimental research. In addition, taking a sample of the general population meant that the composition of the experiment and control groups were different in key factors such as education and income level. This meant that the questions used in our survey also needed to be adapted as the original question sets were tailored to students. It is further essential that the conception of "Asian American" used for our sample draws in this replication lines up with that used in BCS. These points underpin the summary statistics we report in later sections. Moreover, these factors taken together contribute to the reading of this as a "close" rather than exact replication of BCS.

Yet, our results still provide a substantial challenge to the conclusions of BCS. While our sample differs from that of BCS in demographic categories such as education, the conclusions of BCS are not tailored to the sample of students used in the experiment. Instead, the original paper tries to make a general claim about the effect of race salience on Asian Americans in America based on a in-person sample of convenience. The results of experiments like this are usually generalizable only to other samples of college students, and while there advantages to using homogeneous samples of convenience, this means the scope of claims made by studies that use them should be narrower. Peterson & Merunka (2014), Jager et al. (2017).

This is of particular concern given racial salience may be relevant to the

differences in human capital accumulation between groups. The sample in BCS not only made up of undergraduate students, who likely have a higher level of human capital than the average US citizen, but students from prestigious and even Ivy League universities. If communal norms cause Asian Americans to prioritize human capital accumulation as described in BCS, they should have an effect on those with low human capital to push them to develop more. Thus, an experiment like that used in BCS should have a sample that includes those with low human capital The sampling method used in our replication, we argue, is a more representative slice of the general population that the conclusions of the original paper attempt to speak to. In this way, a "close" replication may be a more appropriate course than an "exact" one.

BCS is a well cited paper in the fields of social identity, preferences and priming. It has been cited over 650 times (Google Scholar), with nearly a third of these citations in the last 3 years alone, and when it is cited it is presented specifically as experimental proof of the importance of racial salience to the decision of Asian Americans. With this influence, it is essential to test that this is a robust result and not limited to a segment of the population. The results of our replication run counter to this notion, and its paramount that this is factored in to our understanding of the results of BCS, even if only to limit the scope of the claim it is trying to make. As such, we will attempt to prove that the true nature of the reaction witnessed in BCS cannot be fully understood without additional context. Context that our replication provides a small piece of.

This paper proceeds as follows. In Section 2 we detail our procedures. In Section 3 we will discuss our framework and analysis plan. Section 4 will detail the results of the replication experiment. Section 5 will conclude.

2 Procedure

The experiment was conducted in two rounds, and followed closely to the design of the first experiment presented in the BCS paper. The first round was conducted in 2020 from December 1st to December 13th with a sample of 217 White and 205 Asian Americans. The second was conducted in 2021 from March 9th to March 29th taken from a sample of 242 White, 238 Black, and 254 Asian Americans. In each round, half of the respondents from each racial group were randomly selected to be given the priming questionnaire, with questions pertaining to their ethnic heritage. A necessary component of conducting this experiment on the general public was to update the wording of the control and priming questionnaires to address non-student respondents. The spirit of each question from the 2010 paper was maintained as closely as possible to make sure that the vector of priming was consistent between the two studies.[3] A list of changes made to these questions are included in the supplementary materials along with justifications for each.

Following the questionnaire, each respondent was then asked to complete a number of choice exercises consisting of 46 patience questions and 18 risk preference questions. These were taken directly from the 2010 study and respondents were explicitly told that this would effect their ultimate payouts from the game. We randomly chose one of the 62 questions to determine what individuals would be paid. These responses were used to recover the Log-Interest Rates and Risk Premia of respondents. The procedure for this is outlined in the next section. Finally, respondents were given a debriefing survey based on that given in the second experiment in the 2010 study. This survey was slightly altered from the original to assure that respondents were non-identified, meaning that a detailed

[3]One question about number of meals eaten weekly in dining halls was omitted from the control questionnaire as this may have primed respondents towards financial concerns such as meal plans and contaminated the effect of treatment. This was replaced with a question about proximity to relatives to ensure that both questionnaires had the same number of questions.

account of the immigration status of a respondents parents was not recorded.

We have additionally leveraged the design of subsequent experiments conducted by the original authors (Benjamin et al. (2016)) by implementing a manipulation check. Whilst not directly included in the initial analysis plan for this replication due to its absence in the original paper, the survey responses used to test for manipulation were part of the original debriefing survey in the second experimental round in BCS and included as a relevant co-variate in the determination of risk and time preferences in our analyis plan. Thus, this consideration was included, per the advice of the original authors, to be consistent with modern standards for behavioral experiments. An OLS regression of treatment on responses to 4 questions pertaining to racial identity was used to judge the impression left by the priming survey, which is included in the appendix.[4] Whilst treatment appears to have varying effectiveness based on racial group and the effect on Asian respondents appears to be less significant than other groups, treatment does appear to affect the preferences of respondents on the vector of racial identity/ racial salience, suggesting that the intervention caused its intended effect. The sign of reported coefficients also suggests this effect is consistent across multiple questions and not an isolated case caused by statistical noise. The effect on total racial identification score is significant at the 10% level for the full sample, which is higher than that used in Experiment 2 of BCS. These results further suggest that, even with the necessary changes made to the wording of the treatment and control surveys, treatment likely affected racial salience for respondents in a way not borne out in the responses of the control group. The fact that this was part of the exit survey at the end of the experiment also ensures norms for economics experiments are adhered to

[4]The main outcome observed in this table was a score determined by the same method as the second experiment in BCS. Though this was not originally applied to the Asian sample in the first experiment in BCS, it serves as a decent manipulation check without the need for changes in wording in our replication.

as these questions could not contaminate the effect of treatment on economic decision making.

3 Framework

As with BCS, the structure of our framework borrows from Akerlof & Kranton (2000). With this, we assume that adherence to norms, based on a social category C, is a good that an individual can invest in to gain a higher utility depending on its salience to the current maximization problem. The best response for the individual involves choosing x to maximize

$$U = -(1 - w(s))(x - x_0)^2 - w(s)(x - x_C)^2, \tag{1}$$

with s representing the strength of association with the category being observed, $0 \leq w(s) \leq 1$ being the weight placed on the category C, and x_0 and x_C representing decisions made in the absence of identities influence and fully influenced by identity respectively.

With this underlying framework in mind, we recover the log-interest rates and relative risk premia of individuals. In each we find the approximate compensating variation necessary for said individual to be indifferent between an immediate payment versus a delayed but higher payment in the former case and a safe option versus a risky option in the latter case.

3.1 Calculating Log-Interest Rates

To recover log-interest rates, we first separated the questions into two groups, dependent on whether the safe option was \$3 or \$7 to see if higher stakes had an effect on decision making. Then we took the lowest delayed payment that an individual preferred to the immediate payment and recovered the interest

rate by dividing the former value by the latter and subtracting 1. Conceptually, if respondent preferences are consistent there should be a singular cutoff point where an individual will prefer the delayed payoff to the immediate payoff. However, there were several instances where this was not the case. To account for these instances, we used interval regressions based on upper and lower bounds of the interest rate recovered via the method shown above in the same vein as the 2010 study. The lower bound is the lowest possible candidate for the cutoff where the individual chooses the delayed payment, where the upper bound is the highest possible interest rate.

Finally, we note that in cases where the individual never chooses the delayed payment we consider an upper bound for log interest rates across all respondents. As in the original paper, we use interval regression to account for the possibility of choices that contradict motonicity(i.e. choosing \$2 when you would reject \$4). Whilst, interval regression allows for variables to be unbounded, setting a logical range for outcomes assists in ensuring convergence even when the range is extreme. It may also be relevant to consider how ranges affect measures of effect size and significance. We set the upper bound to be 4, which represents a compensating variation of roughly \$54.60. This upper bound was selected as it is rounded up from the maximum absolute value of all other recorded log interest rates. In considering alternative upper bounds, changes did not affect the significance of the treatment on any category.[5] Results of regressions using these alternative bounds are included in the appendix and supplementary materials for reference.

[5]Whilst the significance of the effects observed were not affected, attributes such as the R-squared value were. In addition while there it is not shown to be significant, there was a change in the direction of the effect for the White sample for the higher IR bound. However, it is difficult to assess the importance of these changes. The same is true of the alternative bounds considered for calculating risk premia.

3.2 Calculating Risk Premia

The compensating variation required to make the respondent indifferent between the risky and safe option was recovered in much the same way as the log-interest rate. The premia cut off was found by subtracting the value of the safe option from the risky option where the individual first picked the risky option. Interval regression was used to account for cases where choices were not consistent. The upper bound for the risk premia was set at $5, which was 1 greater than the highest possible profit from picking the risky option rather than the safe option and represents a level of risk aversion where individuals are only willing to engage with risk after more than doubling the payout on offer. Alternative upper bounds did not show significant changes in the results in any category.

4 Results

Table 1 shows the results of a basic t-test comparing the difference between the choices made by respondents given the control questionnaire and the priming questionnaire within our replication and compares them to those of the original experiment.[6] The left column group details the average proportion of impatient to patient choices chosen by race, while the right group compares safe to unsafe choices. [7] Though we do find significant effects, the results of BCS are not reproduced. In contrast to the 2010 study, White and Asian respondents in the treatment group chose a higher proportion of impatient options compared to their peers in the control group. The effect of treatment was also only significant for White respondents in our replication, where it was only significant for Asian

[6] As Black respondents were only introduced in the second experiment and generalizing an effect size for all Black respondents from the effect on native Black and immigrant Blacks may be improper given the different type of questionnaire they were given, an effect size for Black respondents are omitted from the BCS portion of this table. This is additionally true of Table 2.

[7] These averages represent percentages and should be read as being out of one-hundred.

	Percent impatient choices			Percent safe choices		
	Asian	White	Black	Asian	White	Black
Panel A: 2020/2021 Replication						
Control	38.15	37.30	39.76	64.94	64.59	64.65
	(31.29)	(31.06)	(30.64)	(31.66)	(30.21)	(31.82)
Treatment	39.69618	41.39	37.03	64.72685	65.46547	61.00629
	(29.99)	(31.06)	(29.14)	(31.63)	(30.48)	(31.32)
p-value of diff.	0.298	0.004	0.163	0.417	0.535	0.078
N	452	460	238	452	460	238
Panel B: 2010 Experiment						
Control	26.37	20.90	-	66.67	57.96	-
	(17.49)	(17.94)		(21.54)	(25.00)	
Treatment	12.63	27.14	-	64.41	57.28	-
	(16.28)	(17.78)		(25.07)	(16.34)	
p-value of diff.	0.001	0.164	-	0.687	0.900	-
N	71	66	-	71	66	-

Table 1: Percent of Impatient or Safe Choices, Replication vs. Original

respondents in the original study. Black respondents in the treatment group chose safe options less frequently, but this is only significant at the 10% level. Black respondents in the treatment group also made fewer impatient choices, but this effect was not significant.

Table 2 displays a direct regression of the treatment onto log-interest rates and risk premia of of respondents recovered through the method discussed in section 3. As with Table 1, these are compared to the results found for the first experiment in BCS. The regressions on log interest rate include dummies for "larger stakes" and the time of the "impatient" choice listed in the table as "1 week versus 2 weeks". The former has a value of 1 if the impatient choice is \$7 rather than \$3, while the latter has a value of 1 if the question asks the respondent to choose between payment in a week or two weeks rather than an immediate payment versus payment in a week. We do find, like in BCS, that the Log-interest rates of all respondents decreased significantly when making decisions with larger stakes. However, these regressions fail to show that the priming survey had any significant effect on the risk premia nor log-interest rates for any race when controlling for the factors listed above.

	Log interest rate			Risk Premia		
	Asian	White	Black	Asian	White	Black
Panel A: 2020/2021 Replication						
Treatment	.0028	.0740	-.1769	-.036	.0803	-.2355
	(.106)	(.106)	(.139)	(.105)	(.101)	(.147)
Larger stakes	-.7513***	-.7649***	-.8941***	-	-	-
	(.127)	(.127)	(.166)			
1 week versus 2 weeks	.2173	.2420	.3437*	-	-	-
	(.130)	(.130)	(.170)			
Larger stakes ×	15.62	15.37	14.60	-	-	-
(1 week versus 2 weeks)	(239.0)	(189.1)	(153.9)			
Constant	-1.023***	-1.100***	-1.014***	1.363***	1.369***	1.413***
	(.106)	(.103)	(.132)	(.076)	(.070)	(.098)
$\hat{\sigma}$	.6406***	.6389***	.5859***	.7899***	.7557***	.7996***
	(.020)	(.020)	(.027)	(.017)	(.017)	(.023)
N	1808	1840	952	452	460	238
Panel A: 2010 Experiment						
Treatment	-1.417***	.4220	-	-.0336	-.0210	-
	(.378)	(.371)		(.070)	(.066)	
Larger stakes	-.3909***	-.5592***	-	-	-	-
	(0.101)	(0.127)				
1 week versus 2 weeks	-.0605	-.3272	-	-	-	-
	(.156)	(.180)				
Larger stakes ×	-.0584	.0887	-	-	-	-
(1 week versus 2 weeks)	(.151)	(.177)				
Constant	-2.432***	-2.784***	-	-.206***	.0887**	-
	(.245)	(.311)		(.0509)	(0.0440)	
$\hat{\sigma}$	1.636	1.646	-	.2918	.2652	-
	(.135)	(.146)		(.028)	(.025)	
N	284	264	-	71	66	-

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Priming Treatment Effect on Log Interest Rate and Risk Premium, Replication vs. Original

	Men				Women			
	Total	Asian	White	Black	Total	Asian	White	Black
Panel A: Percent impatient choices								
Control	39.74**	38.26**	40.02	42.43	36.33	37.62	34.71**	37.38
	(31.5)	(31.1)	(31.8)	(31.7)	(30.6)	(31.4)	(30.2)	(29.8)
Given Priming Survey	42.62**	42.94**	42.88	42.05	37.13	36.51	39.85**	33.47
	(31.0)	(33.1)	(29.5)	(28.2)	(30.2)	(30.0)	(30.46)	(29.31)
p-value of diff.	.03	.03	.16	.90	.52	.59	.01	.13
Panel B: Percent safe choices								
Control	63.66	66.77	61.02	63.98	65.45	62.87	68.17	64.71
	(29.6)	(29.2)	(28.7)	(32.1)	(32.5)	(34.0)	(31.4)	(31.6)
Given Priming Survey	62.71	63.51	63.57	59.09	65.28	64.22	67.43	62.37
	(31.2)	(31.3)	(30.3)	(32.8)	(31.2)	(32.1)	(30.6)	(30.3)
p-value of diff.	.45	.10	.19	.13	.90	.54	.72	.38
N	575	239	232	104	587	220	227	133

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Percent of Impatient or Safe Choices, Both Rounds by Racial and Gender Groups

Table 3 shows the results of another t-test, focusing now solely on the replication data, separating out the respondents into categories based on race and gender. This reveals that men in general and Asian men in particular chose the impatient option more in the treatment group. This effect was also observed in White women. In the former case, this again conflicts with the result reported in BCS that Asian respondents make more patient choices when exposed to treatment. Asian men were also less likely to describe their race as being important to them when compared to Asian women. [8] Yet, the effect of treatment is much less significant for Asian women.

5 Robustness

5.1 Round 1 vs. Round 2

One potential concern of this study is establishing parity between the results gathered in the first and second round of the experiment. For all intents and purposes, the choice game, control questionnaire, and priming questionnaire were virtually identical between rounds.[9] However, the time elapsed between the two rounds may present an issue as the second round was conducted in a period where hate crimes against Asian Americans spiked. (Gover et al. (2020)) A "Round 1 vs. Round 2" variable was specified in the pre-analysis plan in place of the original study's school-based dummy variable. Including this variable in our regressions, along with an interaction term with treatment, will allow us to separate out the possible effect of this spike via our regressions in Table 4.

Table 4 shows that the round of the study was significant for Asian American

[8] We consider the effect of gender on group identity through a simple linear regression similar to what we did with our manipulation checks. These can be found in the appendix.

[9] The only change was 2 additional questions added to the debriefing survey at the end of the experiment and that Black respondents were only included in the second round, meaning the associated coefficients were omitted. Conceptually, this should have no effect on results regarding a respondents recovered log-interest rate and risk premium as these are additions after the questions pertaining to these are asked.

	Asian	White	Black	Men	Women
Panel A: Log interest rate					
Treatment	.0739	.0623	-.1976	-.1011	.0586
	(.147)	(.146)	(.143)	(.124)	(.112)
1 week versus 2 weeks	1.718***	1.620***	1.794***	1.667***	1.783***
	(.151)	(.150)	(.147)	(.128)	(.116)
Round 1 versus 2	.5046**	-.0091	.1687	.0256	.3985*
	(.181)	(.172)	(1.46)	(.164)	(.155)
Treatment ×	-.1896	.0451	-	.3138	-.2693
(Round 1 versus 2)	(.218)	(.216)		(.201)	(.194)
1 week versus 2 weeks ×	-.1896	.0452	-	.3139	-.2693
(Round 1 versus 2)	(.218)	(.216)		(.201)	(.194)
Constant	-1.619***	-1.478***	-1.453***	-1.356***	-1.688***
	(.123)	(.116)	(.111)	(.099)	(.091)
$\hat{\sigma}$	.7304***	.7173***	.6779***	.7372***	.6914***
	(.019)	(.019)	(.027)	(.017)	(.017)
N	904	920	476	1150	1174
Panel B: Risk premium					
Treatment	-.2100	-.2067	-.2443	-.2716*	-.1157
	(.141)	(.136)	(.147)	(.114)	(.114)
Larger stakes	5.2×10^{-16}	9.5×10^{-16}	-1.9×10^{-15}	1.1×10^{-15}	-2.1×10^{-16}
	(.141)	(.136)	(.146)	(.114)	(.114)
Round 1 versus 2	.0098	-.0173	.9111	-.1544	.2101
	(.185)	(.173)	(1.58)	(.160)	(.168)
Treatment ×	.3972	.6196**	-	.4519*	.5746**
(Round 1 versus 2)	(.210)	(.202)		(.183)	(.196)
Larger stakes ×	-1.0×10^{-15}	-8.2×10^{-16}	-	-9.5×10^{-16}	3.1×10^{-16}
(Round 1 versus 2)	(.209)	(.202)		(.183)	(.196)
Constant	1.359***	1.377***	1.413***	1.354***	1.358***
	(.126)	(.116)	(.122)	(.097)	(.098)
$\hat{\sigma}$	.7877***	.7510***	.7993***	.7448***	.7996***
	(.017)	(.017)	(.023)	(.015)	(.015)
N	904	920	476	1150	1174

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Interactions between Effects of Questionnaire, Stakes, and Round of Survey

	Men				Women			
	Total	Asian	White	Black	Total	Asian	White	Black
Panel A: Log-Interest Rates								
Treatment	-.0949	.0523	.0124	-.3404	.0772	.1914	.1133	-.0219
	(.121)	(.205)	(.210)	(.221)	(.109)	(.195)	(.195)	(.175)
Larger stakes	-.6667***	-.6411***	-.5891**	-.8814***	-.8920***	-.8517***	-.9279***	-.9050***
	(.115)	(.181)	(.182)	(.263)	(.107)	(.175)	(.178)	(.211)
1 week versus 2 weeks	.2131	.1566	.3293	.04558	.3132**	.3023	.1538	.5781**
	(.118)	(.185)	(.187)	(.271)	(.109)	(.178)	(.182)	(.216)
Larger stakes ×	15.63	15.19	15.22	15.16	15.27	14.73	15.14	14.05
(1 week versus 2 weeks)	(224.2)	(178.3)	(265.4)	(263.6)	(192.3)	(157.1)	(170.2)	(179.7)
Round 1 versus Round 2	.0030	.3109	-.1770	-	.3234*	.5839**	.1634	-
	(.140)	(.221)	(.211)		(.131)	(.211)	(.204)	
Treatment ×	.2957	.3150	.0509	-	-.2714	-.7613**	.0049	-
(Round 1 versus 2)	(.196)	(.302)	(.303)		(.188)	(.292)	(.298)	
Constant	-1.017***	-1.191***	-1.059***	-.6791**	-1.227***	-1.313***	-1.127***	-1.298***
	(.106)	(.183)	(.176)	(.208)	(.098)	(.176)	(.170)	(.170)
Panel B: Risk Premia								
Treatment	-.2655*	-.2684	-.0859	-.5687*	-.1157	-.1179	-.3415	.0580
	(.114)	(.191)	(.194)	(.223)	(.114)	(.209)	(.191)	(.194)
Round 1 versus round 2	-.1544	-.4344*	.1508	-	.2101	.4730*	-.1650	-
	(.131)	(.205)	(.195)		(.137)	(.225)	(.202)	
Treatment ×	.4396*	.3838	.3298	-	.5746**	.4057	.9845***	-
(Round 1 versus 2)	(.184)	(.281)	(.279)		(.196)	(.311)	(.293)	
Constant	1.354***	1.571***	1.119***	1.462***	1.358***	1.136***	1.626***	1.321***
	(.079)	(.141)	(.130)	(.145)	(.080)	(.154)	(.134)	(.132)
N	575	239	232	104	587	220	227	133

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Log-Interest rate and Risk Premia Interval Regressions, Both Rounds by Racial and Gender Groups

respondents in determining effective log-interest rates with a less significant effect shown for women. However, the treatment coefficient shows that the effect of racial priming on log-interest rates was not significant for White or Asian respondents in round 1, and the coefficient of the interaction term reveals that this effect did not change significantly between rounds. This is similarly true for men and women. It is notable that the effect of treatment for Asian Americans does change sign from positive to negative. The size of our reported coefficients compared to the constant term may suggest that the omission of round number in Table 2 may have masked the true effect of treatment. Regardless, this suggests the effect is likely small and, at the very least, is not robust to alternate specifications. The effect of treatment on risk premium was shown to change significantly between rounds for White respondents and women, but BCS did not find the effect of treatment on risk preferences to be significant.

Table 5 shows the results of the interval regressions done in Table 2, now separating respondents into groups based on gender and race and accounting for the round of the study. Interestingly, while both Asian respondents and women did not show a significant change in the effect of treatment between rounds, Asian women did display this change. This is particularly notable as results shown in the appendix suggest that Asian men may identify less with their race more than Asian women, especially as the sign of the coefficient of the interaction term in Table 5 is consistent with Asian women being more patient. While this may suggest that racial salience may have been affected by the Atlanta spa shooting and this in turn affected how Asian women responded to racial priming, this effect could also be due to a number of factors during the COVID-19 pandemic.

The sign of the interaction coefficient does suggest the significance of treatment decreased. Conceptually, this fits with the idea that race became more

salient in every day life for Asian Americans, and thus our treatment may not have as much of a priming effect. However, even when considering the effect of treatment in round 1 alone, before these possible shocks to racial salience occurred, the findings of BCS still do not replicate.

5.2 Students

Whilst the results above show that the effect of treatment reported in BCS does not replicate on a sample of the general population, it may be worthwhile to consider if the same would be true for a sample restricted to only students like in the original experiment. Table 6 reports the percent of impatient and safe choices for respondents in the control and treatment groups that reported that they were students at the time of the experiment. Note that the sample size is comparable to that of the original experiment. Like in Table 1, White respondents in the treatment group are significantly more likely to pick the impatient choice than in the control. Asian and Black respondents in the treatment group were also significantly less likely to pick the safe option than their peers in the control group. Neither of these results match those of BCS. Asian respondents in the treatment group are now less likely to pick the impatient choice, which was the main finding of BCS. However, here this effect is not significant where it was significant at the 1% level in BCS. Table 7 shows the results of regressions conducted on the same sample. This shows that Black respondents in the treatment group had a significantly higher risk premia. When decisions had larger stakes, the Log-interest rates of respondents also decreased significantly only if they were Asian. Treatment did not show a significant effect on the Log-interest rates of Asian respondents like in BCS, but the sign of the effect is now more in line with original compared to our regressions on the general population. Given the 2010 experiment was conducted on students that were members of

	Percent impatient choices			Percent safe choices		
	Asian	White	Black	Asian	White	Black
Students in 2020/2021 Replication						
Control	41.94	37.41	37.61	68.28	53.01	46.20
	(34.53)	(29.09)	(25.98)	(24.43)	(26.88)	(29.87)
Treatment	39.64	49.68	35.94	60.55	56.94	62.5
	(30.63)	(29.81)	(34.56)	(33.78)	(20.98)	(27.75)
p-value of diff.	0.553	0.006	0.745	0.039	0.288	0.001
N	124	96	76	124	96	76

Table 6: Percent of Impatient or Safe Choices for Students

	Log interest rate			Risk Premia		
	Asian	White	Black	Asian	White	Black
Students in 2020/2021 Replication						
Treatment	-.2701	.4813	-.1257	-.0345	.3040	.9779***
	(.259)	(.314)	(.398)	(.239)	(.261)	(.266)
Larger stakes	-.8456**	-.6633	-.3941			
	(.307)	(.368)	(.477)			
1 week versus 2 weeks	.4105	.4713	.2354			
	(.317)	(.390)	(.490)			
Larger stakes ×	14.63	13.60	14.94			
(1 week versus 2 weeks)	(253.8)	(293.2)	(226.5)			
Constant	-.9864***	-1.458***	-1.218**	1.235***	.5423**	.1259
	(.262)	(.295)	(.383)	(.179)	(.173)	(.177)
$\hat{\sigma}$	.5949***	.5098***	.6825***	.6854***	.4968***	.4199***
	(.050)	(.065)	(.071)	(.042)	(.055)	(.061)
N	284	176	140	142	88	70

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Priming Treatment Effects For Students

prestigious Ivy league schools and we did not ask respondents what schools they attended, the human capital accumulation between the sample in BCS and this group may not be exactly comparable. However, given these results are closer to that of BCS than those shown in previous sections, it is possible that this is due to the original experiment being limited to those with already high human capital.

5.3 Respondents With Doubts

Finally, the debriefing survey included a question about whether or not respondents believed that the payment they received would actually be dependent on

	Men				Women			
	Total	Asian	White	Black	Total	Asian	White	Black
Panel A: Log-Interest Rates								
Treatment	-.0834	-.0623	.0196	-.3030	.0772	.1948	-.0116	.0575
	(.154)	(.260)	(.253)	(.290)	(.133)	(.241)	(.235)	(.213)
Larger stakes	-.6721***	-.6703**	-.6191**	-.7797*	-.9557***	-.8321***	-1.062***	-.9987***
	(.143)	(.222)	(.223)	(.346)	(.130)	(.215)	(.214)	(.256)
1 week versus 2 weeks	.2196	.2828	.2938	-.0917	.3696**	.3931	.1999	.5976*
	(.147)	(.228)	(.228)	(.353)	(.133)	(.219)	(.218)	(.261)
Larger stakes ×	14.89	14.65	14.76	14.18	15.00	14.54	14.68	13.66
(1 week versus 2 weeks)	(138.1)	(214.3)	(162.7)	(154.6)	(229.9)	(183.9)	(170.9)	(233.4)
Round 1 versus Round 2	.0052	.3229	-.3019	-	.1941	.4160	.0649	-
	(.169)	(.261)	(.260)		(.160)	(.266)	(.249)	
Treatment ×	.2171	.2098	.1326	-	-.1243	-.5666	.2096	-
(Round 1 versus 2)	(.242)	(.369)	(.372)		(.227)	(.359)	(.359)	
Constant	-1.015***	-1.226***	-.9391***	-.7663**	-1.264***	-1.407***	-1.086***	-1.332***
	(.134)	(.226)	(.215)	(.273)	(.120)	(.217)	(.210)	(.205)
Panel B: Risk Premia								
Treatment	-.1776	.0043	-.1676	-.6106*	-.1585	-.1064	-.4773	.0432
	(.139)	(.252)	(.213)	(.258)	(.142)	(.262)	(.246)	(.224)
Round 1 versus round 2	-.1935	-.6325*	.0699	-	-.1003	.0288	-.5937*	-
	(.152)	(.252)	(.221)		(.171)	(.288)	(.263)	
Treatment ×	.5785**	.5551	.3760	-	1.085***	.9027*	1.578***	-
(Round 1 versus 2)		(.217)	(.357)	(.313)		(.242)	(.390)	(.376)
Constant	1.095***	1.393***	.9907***	.9603***	1.346***	1.221***	1.715***	1.164***
	(.096)	(.181)	(.146)	(.167)	(.101)	(.194)	(.180)	(.151)
N	359	141	160	54	378	141	148	84

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Interactions Excluding Non-Believing Respondents

their choices in the 62 choice games. Table 8 shows the results of regressions where all respondents who reported that they did not believe their choices mattered are omitted. This attempts to correct for the possibility that the decisions of the respondents did not feel "real" enough for them to appropriately be extrapolated to a real economic context given this may be a concern for the studies conducted online. The change in the effect of treatment between rounds is now insignificant. In addition, the interaction between treatment and round number is no longer significant at any level for Asian women. Of note is that the observation numbers show that White women were less likely than White men to believe their choices matter, but Black and Asian women were more likely than Black and Asian men to believe so.

6 Conclusion

We fail to replicate the main result of BCS that increasing the salience of the Asian American identity makes Asian respondents more patient. Restricting our samples to only students does make the sign of the effect of treatment on the Log-interest rates of Asian respondents align with those of the original study, but this effect is not significant. This could mean the effect reported in BCS was due to the sample being made up of students, and extrapolating this to the general population is improper. The original experiment may alternatively simply have been under-powered.

Still, it is important to consider how an experiment conducted remotely may differ from one that is in person. It is difficult to gauge if decisions "feel" the same to respondents. The Asian group does appear to react to the manipulation in a disparate way to all other groups given treatment, but by making Asian respondents identify less with their racial group. It is difficult to assess if this was similar to the effect of treatment in the original experiment as the results

to these questions were not available for BCS. We did find that the effect of larger stakes was similar to that of the original experiment, but we find very different results from BCS in several other outcomes. The results of BCS are at minimum not robust to these alternate specifications.

The lack of reproducability of some research on environmental priming in psychology and behavioral economics is a growing concern. We believe that more replications are needed, especially of highly cited articles. It is very likely that environmental priming is exceptionally multifaceted and insufficiently understood by the current literature to make sweeping conclusions about results from this type of experimental setup. BCS does show an awareness of the limitations of this methodology, but we believe that caution when interpreting the results of future studies is very important and should be further stressed.

The main experiment featured in BCS was motivated by the differences in human capital accumulation between racial groups. It is for this reason that drawing from a sample of only college students—and Ivy league students in particular—is potentially problematic. The results featured in BCS appear to be, at minimum, biased towards the sensibilities of those with higher human capital, and this could explain why we did not find the same results.

Still, we are cognizant of the limitations of our own data. Given events such as the spike in Asian hate crimes during 2020, there may have been a political shift that had unique interactions with the primes introduced in our experiment. However, our robustness checks show that the effect of treatment did not line up with the original results of BCS even before these events occurred. It is also difficult to quantify how political factors such as the SARS pandemic or the Iraq War may have effected racial salience at the time of the original study. We contend that 2004 was no more "neutral" a time period for this experiment than 2020. Given the complex nature of environmental primes, any future research

wishing to utilize the results of the original paper or even our replication should proceed with caution.

Ultimately, we hope that by highlighting this example, more attention will be paid to the replicability of economics experiments as the field continues to grow.

References

Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., Babel, M., Bahník, , Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F. & Penuliar, M. (2015), 'Estimating the reproducibility of psychological science', *Science* **349**.

Akerlof, G. & Kranton, R. (2000), 'Economics and identity', *The Quarterly Journal of Economics* **115**, 715–753.

Bargh, J. & Chartrand, T. (2000), 'The mind in the middle: A practical guide to priming and automaticity research', *Handbook of research methods in social and personality psychology* pp. 253–285.

Benjamin, D., Choi, J. & Strickland, A. (2010), 'Social identity and preferences', *The American economic review* **100**, 1913–1928.

Benjamin, D. J., Choi, J. J. & Fisher, G. (2016), 'Religious Identity and Economic Behavior', *The Review of Economics and Statistics* **98**(4), 617–637.

Boonmanunt, S., Kajackaite, A. & Meier, S. (2020), 'Does poverty negate the impact of social norms on cheating?', *Games and Economic Behavior* **124**, 569–578.

Camerer, C., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. & Wu, H. (2016), 'Evaluating replicability of laboratory experiments in economics', *Science* **351**.

Carroll, C. D., Rhee, B.-K. & Rhee, C. (1994), 'Are There Cultural Effects on Saving? Some Cross-Sectional Evidence*', *The Quarterly Journal of Economics* **109**(3), 685–699.
URL: *https://doi.org/10.2307/2118418*

Cesario, J. (2014), 'Priming, replication, and the hardest science', *Perspectives on Psychological Science* **9**.

Chatterjee, S. & Zahirovich-Herbert, V. (2009), 'Retirement plan participation in the united states: Do public sector employees save more?', *European Journal of Social Sciences* **7**.

Chiswick, B. (1983), 'The earnings and human capital of american jews', *Journal of Human Resources* **18**, 313–336.

Cohn, A., Engelmann, J., Fehr, E. & Maréchal, M. (2015), 'Evidence for countercyclical risk aversion: An experiment with financial professionals', *American Economic Review* **105**, 860–885.

Gelman, A. & Carlin, J. (2014), 'Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors', *Perspectives on psychological science : a journal of the Association for Psychological Science* **9**, 641–651.

Gover, A., Harper, S. & Langton, L. (2020), 'Anti-asian hate crime during the covid-19 pandemic: Exploring the reproduction of inequality', *American Journal of Criminal Justice* **45**.

Helliwell, J., Wang, S. & Xu, J. (2015), 'How durable are social norms? immigrant trust and generosity in 132 countries', *Social Indicators Research* **128**.

Ioannidis, J. (2008), 'Why most discovered true associations are inflated', *Epidemiology (Cambridge, Mass.)* **19**, 640–8.

Ioannidis, J., Stanley, T. & Doucouliagos, H. (2017), 'The power of bias in economics research', *The Economic Journal* **127**, F236–F265.

Jager, J., Putnick, D. L. & Bornstein, M. H. (2017), 'Ii. more than just convenient: The scientific merits of homogeneous convenience samples.', *Monogr Soc Res Child Dev* **82**(2), 13–30.

Khadjavi, M. (2014), 'On the interaction of deterrence and emotions', *Journal of Law, Economics, and Organization* **31**.

Klein, R., Vianello, M., Hasselman, F., Adams, B., Adams, R., Alper, S., Aveyard, M., Axt, J., Babalola, M., Bahník, , Barlow, F., Berkics, M., Bernstein, M., Berry, D., Bialobrzeska, O., Bocian, K., Brandt, M., Busching, R., Cai, H. & Nosek, B. (2018), 'Many labs 2: Investigating variation in replicability across sample and setting'.

Mood, C. & Jonsson, J. (2015), 'The social consequences of poverty: An empirical test on longitudinal data', *Social Indicators Research* **127**.

Mueller-Langer, F., Fecher, B., Harhoff, D. & Wagner, G. (2018), 'Replication studies in economics—how many and which papers are chosen for replication, and why?', *Research Policy* **48**.

Penner, A. & Saperstein, A. (2009), 'How social status shapes race', *Proceedings of the National Academy of Sciences of the United States of America* **105**, 19628–30.

Peterson, R. A. & Merunka, D. R. (2014), 'Convenience samples of college students and research reproducibility', *Journal of Business Research* **67**(5), 1035–1041.
URL: *https://www.sciencedirect.com/science/article/pii/S014829631300307X*

Sowell, T. (1975), *Race and Economics*, New York: David McKay Company, Inc.

Springstead, G. & Wilson, T. (2000), 'Participation in voluntary individual savings accounts: An analysis of iras, 401 (k)s, and the tsp', *Social security bulletin* **63**, 34–9.

Sue, S. & Okazaki, S. (1990), 'Asian-american educational achievements: A phenomenon in search of an explanation', *The American psychologist* **45**, 913–20.

Wiggins, B. & Christopherson, C. (2019), 'The replication crisis in psychology: An overview for theoretical and philosophical psychology.', *Journal of Theoretical and Philosophical Psychology* **39**, 202–217.

A Appendix

	Total	Asian	White	Black	Male	Female	As. Male	As. Female
Panel A: Simple Regression on Racial Identification Score (Replication)								
Treatment	.2419*	-.1116	.4748**	.1846	.3910**	.0906	-.0136	-.2095
	(.1020)	(.1661)	(.1603)	(.1949)	(.1449)	(.1437)	(.2440)	(.2239)
N	1166	452	460	238	575	578	230	220
Panel B: Simple Regression on Racial Identification Score (BCS)								
Treatment	1.3177	-	1.5716	.5523	.8022	1.9622	-	-
	(.6848)		(1.0460)	(.8780)	(1.1535)	(1.0266)		
N	217	-	90	82	78	103	-	-

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: The Dependent variable in this table (Racial Identification Score) is generated by respondents degree of agreement with the statements "My race is an important part of my self-image" and "My race is an important reflection of who I am" minus their degree of agreement with the statements " My race has very little to with how I feel about myself." and "My race is unimportant to my sense of what kind of person I am." in accordance with how it was computed in BCS.

Table 9: Simple Linear Regression Manipulation Check

Questions Asked	The Effect of Male = 1			
	Total Sample	Only White	Only Asian	Only Black
"My gender is an important part of my self-image."	-.0217 (.0291)	.0961* (.0494)	.0022 (.0496)	-.1264** (.0588)
"My gender has very little to do with how I feel about myself."	.1565*** (.0494)	.3484*** (0754)	-.0422 (.0772)	.2728** (.1223)
"My gender is unimportant to my sense of what kind of person I am."	.1567*** (.0479)	.1813** (.0733)	.1787** (.0765)	.2285** (.1124)
"My race is an important part of my self-image."	-.1362*** (.0356)	.0936 (.0646)	-.2837*** (.0490)	-.1092 (.0570)
"My race has very little to do with how I feel about myself."	.2432*** (.0479)	.2633*** (.0669)	.2263*** (.0785)	.2773** (.1202)
"My race is unimportant to my sense of what kind of person I am."	.0874* (.0472)	.1300** (.0633)	.0643 (.0796)	.2072* (.1165)
N	1166	460	452	238

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Simple Linear Regression of Gender on Followup Survey Responses

lndiscratelo lndis∼i	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
Round 1: Low LIR Bound (Asian Sample)							
primingQ	-0.075	0.244	-0.31	0.759	-0.553	0.403	
largestakes	-0.65	0.089	-7.29	0	-0.824	-0.475	***
longterm	0.163	0.129	1.26	0.207	-0.09	0.415	
largelong	15.549	0.537	28.98	0	14.498	16.601	***
Constant	-0.813	0.191	-4.25	0	-1.187	-0.438	***
Constant	0.693	0.052	13.38	0	0.592	0.795	***
Mean Dep. var	-0.571	SD	2.385				
Obs	820.000	χ^2	1281.387				
Prob $>\chi^2$	0.000	AIC	2432.178				
Round 1: Low LIR Bound (White Sample)							
primingQ	0.105	0.211	0.50	0.618	-0.309	0.519	
largestakes	-0.725	0.106	-6.87	0	-0.932	-0.518	***
longterm	0.25	0.116	2.16	0.031	0.023	0.476	**
largelong	14.52	0.542	26.81	0	13.458	15.581	***
Constant	-1.126	0.189	-5.95	0	-1.496	-0.755	***
Constant	0.591	0.056	10.47	0	0.48	0.702	***
Mean Dep. var.	-0.611	SD	2.328				
Obs.	868.000	χ^2	1163.370				
Prob $>\chi^2$	0.000	AIC	2323.167				

Table 11: Regressions w/Lower Upper Bound for Log Interest Rate (LIR = 2)

lndiscratelo lndis∼i	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
Round 1: High IR Bound (Asian Sample)							
primingQ	-0.173	0.431	-0.40	0.688	-1.018	0.672	
largestakes	-0.661	0.157	-4.20	0	-0.969	-0.353	***
longterm	-0.184	0.229	-0.80	0.422	-0.633	0.265	
largelong	29.5	1.264	23.34	0	27.022	31.978	***
Constant	-0.024	0.353	-0.07	0.946	-0.716	0.667	
Constant	1.266	0.069	18.24	0	1.13	1.402	***
Mean Dep Var.	0.424	SD dependent var	4.449				
Obs.	820.000	χ^2	1092.562				
Prob $>\chi^2$	0.000	AIC	3076.162				
Round 1: High LIR Bound (White Sample)							
primingQ	-0.044	0.345	-0.13	0.899	-0.721	0.633	
largestakes	-0.798	0.184	-4.33	0	-1.159	-0.437	***
longterm	-0.05	0.194	-0.26	0.796	-0.431	0.33	
largelong	26.979	1.309	20.62	0	24.414	29.544	***
Constant	-0.501	0.332	-1.51	0.132	-1.152	0.151	
Constant	1.109	0.082	13.52	0	0.949	1.27	***
Mean Dep. var.	0.283	SD	4.272				
Obs.	868.000	χ^2	974.102				
Prob $>\chi^2$	0.000	AIC	2934.985				

Table 12: Regressions w/Higher Upper Bound for Log Interest Rate (LIR = 6)