Online Supplement

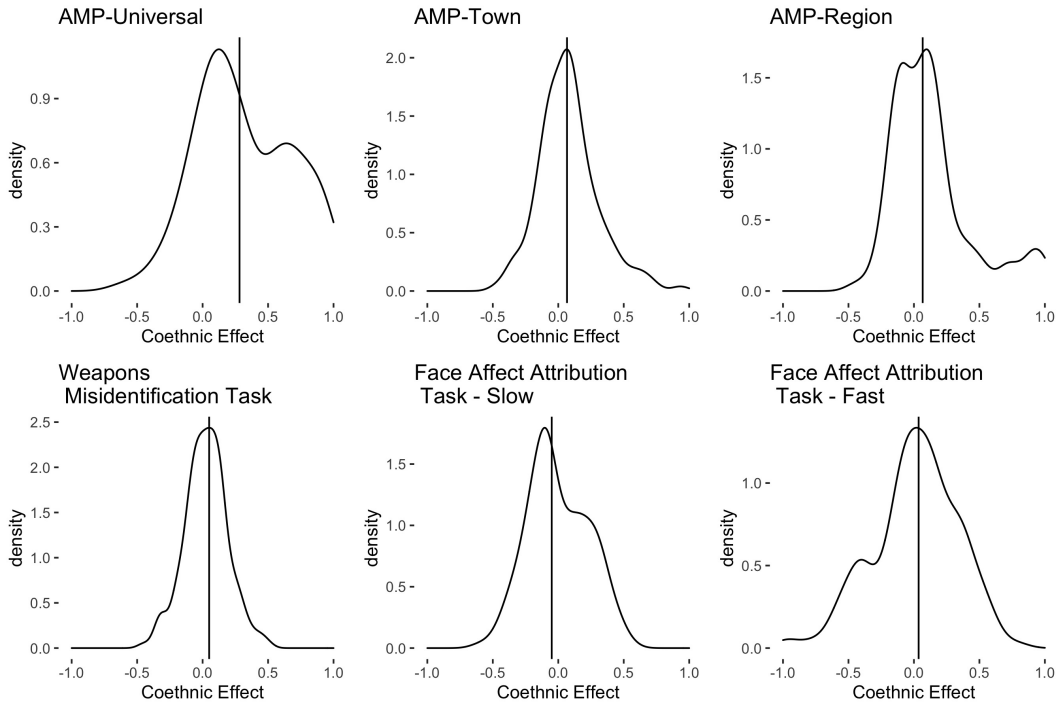
Measuring Ethnic Biases: Can Misattribution-Based Tools from Social Psychology Reveal Group Biases that Economics Games Cannot?

Ashley Blum, Chad Hazlett, & Daniel N. Posner

A Distribution of individual level results

Due to previous research showing bimodal distributions in individual ethnic effects on the AMP (Hazlett and Berinsky 2018) and because the AMP, WMT, and FAA tasks provide numerous trials per individual with which to estimate individual level ethnic effects, we also examine the entire distribution of ethnic effect estimates across individuals in these tasks.

Figure A1. Distribution of Individual Effects, Misattribution Tasks



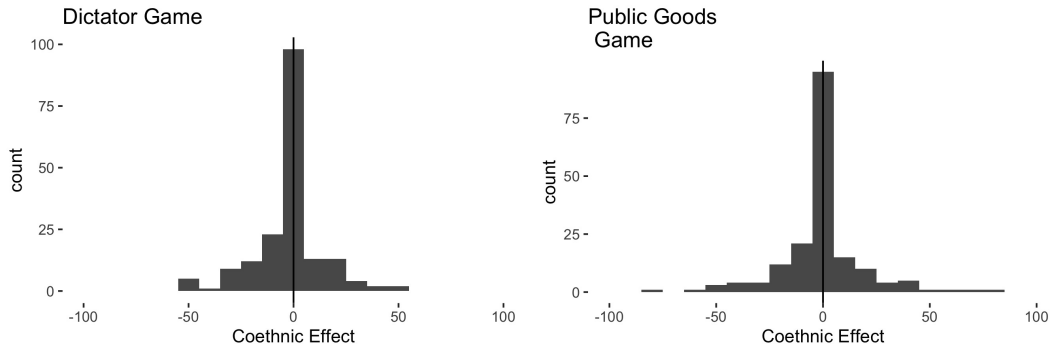
Note: Distributions of person-level ethnic effect estimates for each task.

In the *AMP-universal*, we see a large mode with apparently no effect, but a second strong mode with a large effect of over 50 percentage points. The *AMP-Town* shows a less certain mode above 50 percentage points, whereas the *AMP-region* has a pronounced group with effects larger than 50 percentage points. The Face Affect Attribution tasks show what appear to be two separate groups (modes), particularly for the slow face task. However because one of the modes is below zero, there is no detectable average effect. Finally, in the *WMT*, even though the mean proves significantly distinguishable from zero, the distribution of effects does not show clear "low" and "high" effect

sub-groups.

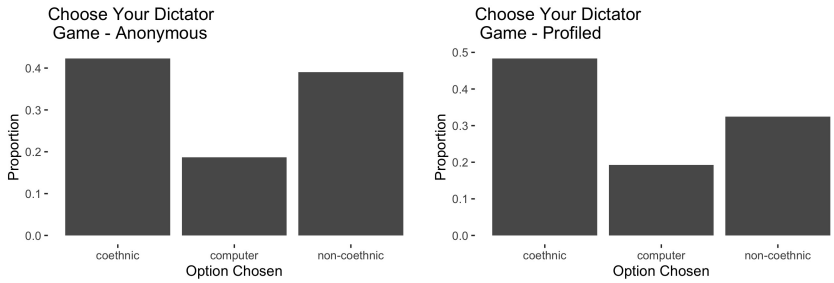
In two of the economics games—DG and PGG—individual level effect estimates can be computed by contrasting conditions in which they play with coethnics or non-coethnics. As this provides only one trial of each type, we do not expect the person-level effects to be very well estimated. Nevertheless, the distribution of person-level effects provides a transparent summary of the data (Figure A2). Both show strong modes at zero, with little evidence of second modes elsewhere, and means indistinguishable from zero.

Figure A2. Distribution of Individual Effects, Economics Games



(a) *Note:* Distribution of individual level ethnic effect estimates in the DG and PGG games.

Figure A3. Distribution of Responses, CYD



Note: Distribution of individual responses in the CYD tasks.

B Results Using Full Sample

In our pre-analysis plan, we registered our intention to analyze the data without rejecting participants based on their scores on the Raven's matrix or literacy tests. We indicated that we would run separate analyses in which we dropped participants based on these factors, treating them as robustness tests. For the reasons provided in the paper, we have reversed this prioritization: the main results we present in the paper are those in which we *do* reject participants on the basis of literacy. In this section, we report the results of our pre-registered specification, without any rejection of individuals. The main difference worth noting is that the WMT no longer reports a significant result.

Table A1. Average Contributions in Economics Games

	Full Sample	Final Sample
Dictator Game	33.23	32.71
Public Goods Game	45.77	44.55

Table A2. Average Proportion of "Positive" Responses

	Full Sample	Final Sample
AMP Universal	0.57	0.55
AMP Town	0.68	0.72
AMP Region	0.67	0.67
WMT	0.42	0.42
SFA	0.53	0.54
FFA	0.61	0.61

Table A3. Ethnic Effect Matrix, Full Sample, No Participant-wise Exclusions

Task	mean(effect)	SE(effect)	z-score	p value
AMP-Universal	0.233	0.019	11.987	0.000
AMP-Town	0.075	0.014	5.474	0.000
AMP-Region	0.103	0.016	6.245	0.000
WMT	0.009	0.010	0.958	0.338
FAA-Slow	0.004	0.013	0.328	0.743
FAA-Fast	0.013	0.018	0.731	0.465
DG	-0.011	0.011	-0.992	0.321
PGG	0.004	0.015	0.265	0.791
CYD-Anonymous	0.028	0.031	0.878	0.415
CYD-Profiled	0.076	0.031	2.433	0.018

C Results Using Full Sample, Dropping Participants With Many Fast Trials

Here we present the main results of a pre-registered additional analysis in which we drop participants from an entire task if their reaction time was less than 200 ms in more than 20% of trials (note that individual trials faster than 200 ms are already excluded from the main analyses, as pre-specified). The results are substantively unchanged, except for the WMT, which loses its statistical significance.

Table A4. Ethnic Effect Matrix, Full Sample, Dropping Participants with Many Fast Trials

Task	mean(effect)	SE(effect)	z-score	p value
AMP-Universal	0.239	0.020	12.228	0.000
AMP-Town	0.080	0.014	5.614	0.000
AMP-Region	0.103	0.017	6.174	0.000
WMT	0.011	0.009	1.171	0.242
FAA-Slow	0.004	0.013	0.305	0.761
FAA-Fast	0.013	0.018	0.721	0.471
DG	-0.011	0.011	-0.992	0.321
PGG	0.004	0.015	0.265	0.791
CYD-Anonymous	0.028	0.031	0.878	0.415
CYD-Profiled	0.076	0.031	2.433	0.018

D Additional Analyses

In this section we response on additional (planned) analyses that space limitations prevented us from including in the main text.

D.1 Correlations in Individual-Level Behavior Across Measurement Strategies

We first consider correlations in the individual-level responses across the games and tasks to determine whether these varied measures capture similar underlying constructs.[1] Table A5 provides a matrix of correlations across all behavioral games and social psychology tasks (and also explicit survey questions, discussed below) with p-values for each comparison.

1. A principal components analysis would also be a natural approach for this question, and we did indeed conduct one. However, since the correlation among items is low, it showed very little loading except between multiple versions of the same task (i.e. versions of the AMP, or versions of the FAA).

All AMP tasks correlate well with each other. The AMP-Region and AMP-Town are especially well correlated at 0.52, as would be hoped: they differ only in whether ethnicity was cued by naming towns or naming regions. Both AMP-Town and AMP-Region also correlate with the AMP-Universal. This could be because general comfort levels with the computerized interface and compliance with the specific task instructions may factor into effect sizes in all three tasks. Alternatively, some individuals may be motivated to avoid any influence of the prime on their response to the target by adopting strategies such as pseudo-randomizing their responses (Hazlett and Berinsky 2018).

Table A5. Ethnic Effect Correlation Matrix

	Task	Oppose Intermarriage	Believe Not Peaceful	Distrust	Segregation Support	CYD-Profiled	CYD-Anonymous	PGG	DG	FAA-Fast	FAA-Slow	WMT	AMP-Region	AMP-Town
Social Psych Tasks	AMP-Universal	-0.033 (0.662)	-0.018 (0.812)	-0.050 (0.505)	-0.006 (0.940)	**-0.156*** (0.037)	**-0.188*** (0.012)	-0.107 (0.154)	-0.048 (0.529)	0.015 (0.842)	0.082 (0.277)	-0.083 (0.273)	**0.240*** (0.006)	**0.304*** (0.000)
	AMP-Town	0.049 (0.542)	0.098 (0.221)	0.146 (0.069)	0.009 (0.914)	0.006 (0.940)	0.088 (0.275)	**-0.174*** (0.029)	0.079 (0.326)	0.132 (0.101)	0.080 (0.318)	-0.043 (0.592)	**0.526*** (0.000)	–
	AMP-Region	0.051 (0.560)	0.085 (0.331)	0.119 (0.174)	-0.103 (0.238)	-0.004 (0.962)	0.065 (0.458)	**-0.179*** (0.039)	0.088 (0.316)	0.145 (0.095)	0.143 (0.101)	0.136 (0.120)	–	
	WMT	-0.003 (0.963)	-0.132 (0.078)	-0.070 (0.348)	-0.092 (0.218)	0.004 (0.954)	0.046 (0.538)	-0.021 (0.781)	0.026 (0.726)	**-0.223*** (0.003)	**-0.300*** (0.000)	–		
	FAA-Slow	0.032 (0.665)	**0.223*** (0.003)	**0.163*** (0.029)	-0.036 (0.628)	0.035 (0.636)	0.132 (0.077)	0.067 (0.373)	**0.161*** (0.030)	**0.524*** (0.000)	–			
	FAA-Fast	0.092 (0.218)	**0.227*** (0.002)	0.133 (0.075)	0.007 (0.926)	0.010 (0.898)	0.141 (0.059)	0.065 (0.386)	0.120 (0.107)	–				
Behavioral Games	DG	0.067 (0.371)	0.111 (0.137)	0.102 (0.170)	0.002 (0.979)	-0.002 (0.980)	-0.065 (0.384)	-0.033 (0.658)	–					
	PGG	0.063 (0.401)	-0.072 (0.336)	0.088 (0.240)	0.057 (0.444)	0.033 (0.659)	-0.058 (0.434)	–						
	CYD-Anonymous	0.001 (0.987)	0.066 (0.377)	0.095 (0.204)	0.044 (0.558)	**0.240*** (0.001)	–							
	CYD-Profiled	0.015 (0.844)	0.070 (0.347)	**0.146*** (0.049)	-0.045 (0.546)	–								
Survey	Segregation Support	-0.097 (0.194)	0.045 (0.548)	0.063 (0.397)	–									
	Distrust	0.043 (0.560)	**0.472*** (0.000)	–										
	Believe Not Peaceful	0.061 (0.413)	–											

Note: Table of cross-task correlations. Values shown in each cell are correlation coefficients, with the corresponding p-value in parentheses. *p<0.05; **p<0.01; ***p<0.001.

We see similar correlations across versions of the FAA (the FAA-Fast and FAA-Slow, $\rho = 0.52$) and versions of the CYD (CYD-Anonymous and CYD-Profiled, $\rho = 0.24$). This suggests that, as expected, the different versions of the FAA and CYD tasks are each measuring similar things (or, alternatively, capture particular tendencies that affect the way individuals engage with each of these tasks). With respect to the FAA, we note that the distribution of individual ethnic effects is relatively bimodal, with two modes on opposite sides of zero (see Appendix A). This suggests that something is being measured at the individual level on this task that we do not currently understand.

Turning to the correlations between tasks of wholly different types (i.e., leaving aside comparisons among different versions of the same tasks), we see little consistent pattern. We find negative relationships between the WMT and both FAA tasks, between the AMP-Region and the PGG, and between the AMP-Universal and both CYDs. Meanwhile, we find positive, significant correlations between the PGG and the FAA-Slow and between the AMP-Town and the CYD-Anonymous. The inconsistent sign on these relationships, combined with the multiple correlations we calculate (40 cross-task/game correlations, of which we find just two positive, statistically significant associations) leads us to conclude that the games and tasks are either not measuring the same underlying biases or that some of them are simply poor measurement tools.

We also compared responses in the behavioral games and psychological tasks to responses to a set of survey questions that asked participants directly about their attitudes toward members of the other ethnic group:[2] We consider four explicit survey questions:

Intermarriage Attitude: "Would you be upset if your son or daughter were to marry a person from [Central/Nyanza]?"

Segregation Support: "Would you support measures to keep people from [Central/Nyanza] from living in your neighborhood?"

Perceived Trustworthiness "How much do you trust people from [Central/Nyanza]?"

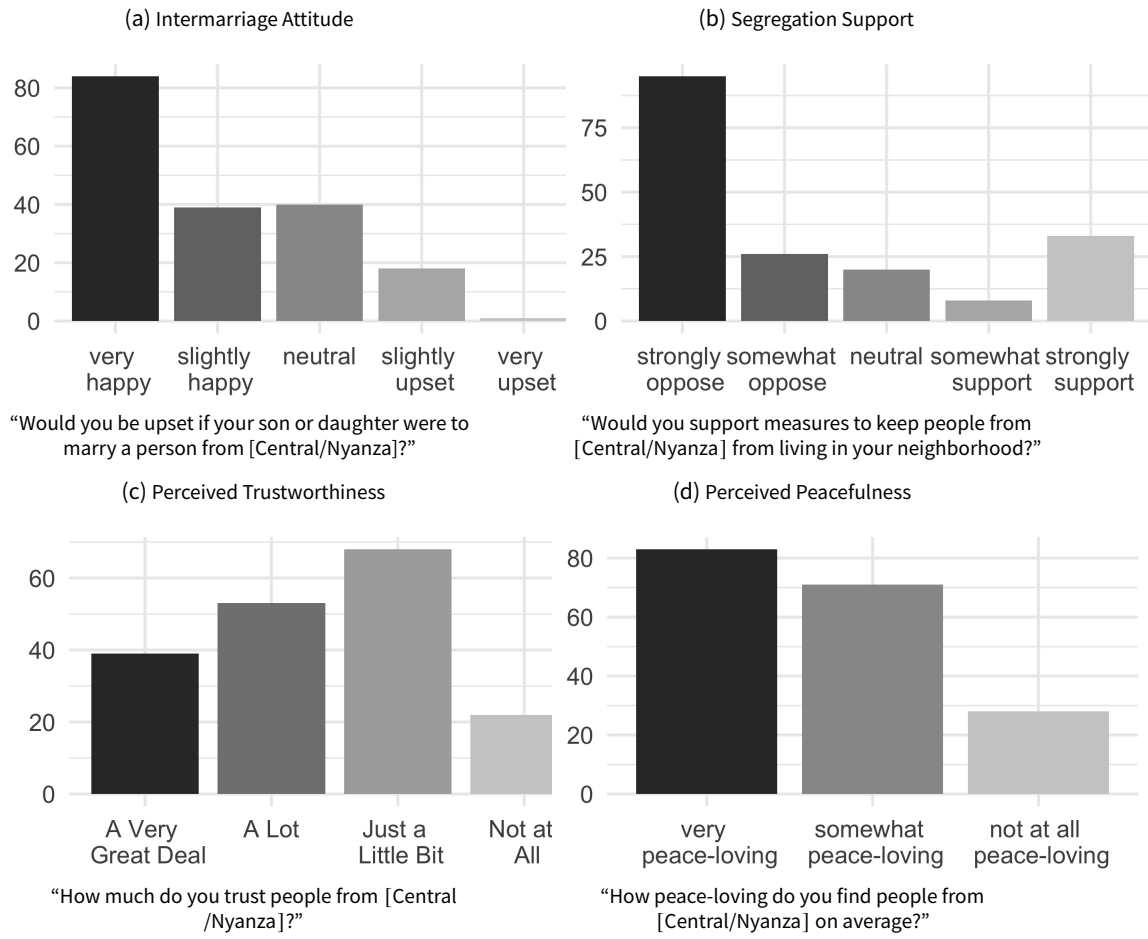
Perceived Peacefulness: "How peace-loving do you find people from [Central/Nyanza] on average?"

The distribution of responses is presented in Figure A4. A majority of respondents either do not have especially hostile attitudes toward members of other ethnic groups or are unwilling to express such attitudes in surveys. However, there were some participants who did express negative attitudes or perceptions. For example, about half of respondents said they trusted people from the other ethnic group either "just a little bit" or "not at all." Unfortunately, we do not have data

2. Here, ethnicity was not cued directly but rather through region, as in the AMP-Region. Kikuyus were asked about their attitudes toward people from Nyanza, while Luos were asked about their attitudes toward people from Central. The question regarding intermarriage was asked during the recruitment survey. The other three questions were asked in the post-survey administered at the end of the lab session.

showing attitudes toward respondents' own ethnic groups to use as a basis of comparison.

Figure A4. Responses to Explicit Questions



To permit comparisons with the behavioral games and social psychology tasks, we transformed the survey responses from ordinal to numerical values, with larger values corresponding to greater ethnic bias.[3] The results, shown in Table A5, indicate that responses to the explicit survey questions are not correlated with responses in the social psychology tasks or behavioral games. We calculate 40 correlations and find only two to be statistically significant at the $p<0.05$ level, which is precisely the number of significant correlations we would expect to find based on chance alone. We conclude that the explicit questions and the laboratory-based measurement approaches are either measuring different dimensions, or that at least one is a very poor measurement approach. Indeed, one reason not to expect correlations is precisely because we are concerned about social desirability bias in explicit survey questions (e.g., Clark and Schober 1992; Bertrand and Mullainathan 2001; Berinsky

3. As noted, most of these questions capture reactions toward non-coethnics without collecting information about comparable reactions toward coethnics. They therefore do not allow for the construction of "ethnic effects" in a way that is analogous to the other approaches. Nevertheless, if more biased individuals tend to give less favorable answers about non-coethnics, then these responses may correlate across individuals with the ethnic effects measured in the games and tasks.

2004). If responses on explicit questions are encumbered by strong social desirability bias as might be feared, then we should not expect to find strong correlations between these and any of the tasks, even if the latter are performing as hoped.

D.2 Follow-up Survey

We conducted a follow-up mobile phone survey of lab participants two months after the completion of the lab sessions. The purpose of the follow-up survey was to test whether behavior in the lab was correlated with attitudes expressed in reaction to the annulment of the August 2017 presidential election by the Kenyan Supreme Court.[4] Given the sharp ethnic polarization surrounding the election, we expected to find attitudes toward both the Court's ruling and the behavior of the Independent Electoral and Boundaries Commission (IEBC) to be strongly correlated with ethnicity—Kenyatta, the incumbent president and original winner, was Kikuyu and Odinga, the primary opposition candidate and the petitioner in the Supreme Court case, was Luo. The question was whether members of each community who had shown themselves to be more ethnically biased in the behavioral games and social psychology tasks expressed more extreme attitudes in the follow-up survey. To operationalize this, we recoded individual responses to indicate whether the participant gave the "ethnically congruent" response, (i.e., a Kikuyu responding "no" or a Luo responding "yes" is coded as 1). We asked the following questions:

Support Rerun: "Do you think that the presidential election ought to be re-run?" [yes; no]

One Side Responsible: "Do you think the two presidential campaigns are equally responsible for the reported irregularities in the most recent presidential election?" [yes; no]

Reform IEBC: "In light of what happened in the August election, the primary opposition candidate has called for significant changes in the electoral commission. To what degree do you support these changes?" [strongly support; somewhat support; somewhat oppose; strongly oppose]

Reform Judiciary: "In light of what happened in the August election, the President has called for significant changes in the judiciary. To what degree do you support these changes?" [strongly support; somewhat support; somewhat oppose; strongly oppose]

In keeping with expectations, participants' ethnic group memberships were strongly predictive of the answers they gave to these questions. Luos were much more likely than Kikuyus to say that they thought the presidential election should be re-run and to support significant changes to the electoral commission. Kikuyus, meanwhile, tended to reject changes to the electoral commission and to support changes to the judiciary.

4. The lab sessions were conducted a month prior to the election and the follow-up survey was conducted between the annulment of the election and its re-running. For a useful summary of the election and its aftermath, see (Chege 2018).

However, as shown in Table A6, there is only one statistically significant correlation between survey responses and our laboratory-based measures of ethnic bias, and it points in the wrong direction (i.e., is negatively signed). While we do find a handful of significant correlations between the responses to the follow-up survey and the direct questions about bias asked in the recruitment and post-lab surveys, two of the four are negatively signed. Taken together, these results suggest that, at least in our case, neither lab behavior nor direct survey questions about ethnic bias are good predictors of explitiely reported attitudes toward a set of ethnically charged real-world events. Our findings are thus in keeping with the skeptical perspectives of Levitt and List (2007) and Kessler and Vesterlund (2015) regarding the inability of experiments measuring social preferences to predict attitudes reported outside of the laboratory.

D.3 Priming Threat and Fear

Finally, we examine how exposure to an experimental prime intended to remind participants of their experience with prior violence may influence the effects picked up by the behavioral games, social psychology tasks, and explicit questions.[5] Participants were randomly assigned to either a control or a priming treatment.[6] Those assigned to the priming treatment were presented at the start of the lab session with a short on-screen quiz (the contents of which were also read to them over the headphones) that asked whether they felt that they or their family were ever at risk of harm during the violence that followed the 2007 election and, if so, what kind of harm they feared most.[7] Those assigned to the control treatment listened to a recording of a bird chirping. Participants assigned to the priming treatment also received a reinforcement prime later in the lab session, between the AMP and the FAA tasks.[8]

Table A7 shows how the individual level ethnic effects differ between those who were primed to fear/threat and those who were not. We find no evidence that the priming affected any of the ethnic effect measures or responses to the explicit questions. As we noted in our pre-analysis plan, while a shift in effects due to such a prime would provide evidence of validity, it is certainly not a

5. Studies using similar experimental priming involving exposure to recollections of a traumatic experience include Lerner *et al.* (2003) and Callen *et al.* (2014).

6. All results in the paper except those reported below collapse across primed and unprimed participants.

7. The first question asked: "Before we start the other activities, we want to ask you a question about your recollection of the 2007 elections. As you know, our country experienced terrible violence in the weeks after the 2007 elections. Thinking back on that time, did you feel that you or your family were ever at risk of harm? Please indicate your answer by pressing the button for "yes," "no," or "I would prefer not to answer" on the screen." The follow-up question then asked: "What kind of harm did you fear the most?" and provided the following options: "violence against you personally; violence against other family members or your community; destruction of property; displacement of you or your family members; disruption of peace in the country; I did not feel fearful of any harm; I would prefer not to answer."

8. The reinforcement prime was administered similarly to the original prime. The first question asked: "Before the next activity, we want to take a break again to ask you another question about your fear of violence in the country. Earlier we asked about the 2007 election. As you know, there is another election scheduled to take place next month. Do you fear that there will be violence during the coming election? The follow-up question then provided the same options, with the addition of "I do not think there will be violence in the coming elections."

Table A6. Follow-up Survey and Ethnic Effect Correlation Matrix

LeftColOct	Task	Support Rerun	One Side Responsible	Reform IEBC	Reform Judiciary
Social Psych Tasks	AMP-Universal	0.037	-0.035	-0.048	-0.016
		(0.642)	(0.660)	(0.550)	(0.838)
	AMP-Town	-0.087	0.133	**-0.176***	0.055
		(0.311)	(0.120)	(0.039)	(0.520)
	AMP-Region	-0.032	0.064	0.031	-0.063
		(0.731)	(0.492)	(0.737)	(0.499)
	WMT	-0.038	-0.084	0.016	-0.105
		(0.633)	(0.291)	(0.841)	(0.189)
	FAA-Slow	0.012	0.010	-0.022	-0.062
		(0.880)	(0.897)	(0.780)	(0.438)
	FAA-Fast	-0.046	0.080	0.028	-0.096
		(0.566)	(0.317)	(0.722)	(0.229)
Behavioral Games	DG	0.075	-0.049	0.057	0.002
		(0.341)	(0.535)	(0.473)	(0.983)
	PGG	0.043	0.056	-0.043	-0.046
		(0.587)	(0.478)	(0.590)	(0.565)
	CYD-Anonymous	0.015	0.038	0.073	0.081
		(0.855)	(0.636)	(0.357)	(0.306)
	CYD-Profiled	-0.132	0.066	0.138	0.055
		(0.095)	(0.403)	(0.081)	(0.490)
Survey	Segregation Support	**-0.186***	-0.103	-0.107	-0.085
		(0.018)	(0.194)	(0.178)	(0.284)
	Distrust	-0.034	**0.166***	0.057	**0.174***
		(0.667)	(0.036)	(0.469)	(0.027)
	Believe Not Peaceful	-0.092	0.125	-0.086	0.006
		(0.246)	(0.114)	(0.276)	(0.937)
	Oppose Intermarriage	-0.069	0.105	-0.139	0.002
		(0.382)	(0.183)	(0.080)	(0.975)

Note: Correlation between laboratory tasks and survey responses and answers to follow-up questions surrounding the 2017 election re-run. Values shown are correlation coefficients, with the corresponding p-value in parentheses. $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

requirement, as there are many reasons this prime may fail to alter ethnic effect estimates even if they were very sensitive to constructs of fear or threat perception. For example, in the WMT, having primed the participant to think about prior violence may simply make her more likely to say that *all* the target images show a weapon, rather than increasing the effect of ethnicity on that tendency.

Additionally, the theory of misattribution that gives rise to the AMP, FAA, and WMT posits that these tools measure associations developed over a long period of time (Payne *et al.* 2005). They may therefore simply be very hard to manipulate in the short-run.[9]

Table A7. Effect of Priming on Coethnic Preference

	Dependent variable:											
	AMP-Region	AMP-Town	WMT	FAA-Slow	FAA-Fast	DG	PGG	CYDA	CYDP	Segregation	Distrust	Peaceful
primed	−0.017	0.032	0.024	0.004	−0.093	0.485	−4.245	−0.020	−0.086	0.183	0.095	−0.068
	(0.056)	(0.036)	(0.024)	(0.034)	(0.050)	(2.404)	(3.568)	(0.074)	(0.075)	(0.230)	(0.142)	(0.107)
Observations	133	157	180	181	181	182	182	182	182	182	182	182

Note: *p<0.05; **p<0.01; ***p<0.001
Robust standard errors in parentheses.

9. We note that a somewhat different set of priming conditions in Berge *et al.* (2020)—to ethnic identity, political competition, and national identity, rather than to past exposure to violence—also failed to generate changes in participants' behavior in the DG, PG or CYD.