MEASURING RACE AND ANCESTRY IN THE AGE OF GENETIC TESTING

By Sasha Shen Johfre, Aliya Saperstein, and Jill A. Hollenbach

ONLINE APPENDIX

In this supplement, we first discuss the details of our propensity-score matching approach to estimate who reports multiple races for self-identification, while accounting for potential selectivity in who takes a genetic ancestry test (GAT). We also present various supplementary tables and figures, numbered according to where they are first referenced in the main text.

Propensity-score matching

In our propensity-score matching approach, we constructed a propensity score by estimating the likelihood of having taken a GAT (see Table A3) using the following categorical predictors: a series of indicators for each type of genealogical research (other than GAT-taking), age, education, region of residence, nativity (U.S. vs. foreign born), and recruitment variables (as a measure of eagerness to participate in genetic ancestry research). These measures in our data were found to be significant predictors of professed (dis)interest in GATs in previous research (Horowitz, Saperstein, Little, Maiers, & Hollenbach, 2019). Using Stata's *psmatch2,* we nearest-neighbor matched our full sample on this propensity score to compare people with similar demographic backgrounds. Our specification achieved sufficient overlap and balance (see Figure A2). Results from our propensity score models were extremely similar in magnitude and significance to results from logistic regression (see Table A4), which suggests that selection into GAT-taking based on these observed factors are not driving our results.

To examine the possibility of unobserved selectivity causing both GAT-taking and the reporting patterns of interest, we ran sensitivity analyses on our p-score models using the Mantel-Haenszel bounds method (Stata's *mhbounds),* which models how strong the association with an unobserved confounder would need to be in order to invalidate the matching results (Becker & Caliendo, 2007; Mantel & Haenszel, 1959). A robust result would indicate that the treatment and outcome are directly causally linked (rather than both caused by a third variable); a sensitive result leaves this question open. When predicting multiracial reporting, our analyses indicated that the p-score model results are somewhat sensitive to potential unobserved confounders. In particular, the estimated average treatment effect would no longer be significant if the odds of differential assignment due to a positively-biased confounder were above 1.15 (see Table A5). Thus, if there is an unobserved confounder that increases the odds of taking a GAT by 15% conditional on the observed covariates (and also causes higher rates of multiple-race reporting), our result may overestimate the true association between taking a GAT and multiracial identification.

One such likely confounder is awareness of racial diversity in one's family history. Although we can measure this in our survey using the number of reported ancestries a respondent selects, for GAT takers it is a post-treatment measure; we do not know the ancestral diversity respondents would have reported prior to taking a GAT. Further, we cannot include reported ancestry counts in our propensity score (or logit) models because of the risk of collider

bias (as it also is related to multiracial reporting, our outcome of interest). This means we cannot definitively rule out the opposite causal interpretation of our results: that people who believe they have mixed ancestry are more likely to take a GAT and are therefore also more likely to report multiple races for self-identification after the fact. As noted in our discussion, we are not overly troubled by this because we expect that causality between GAT-taking and racial identification does run in both directions.

REFERENCES

Becker, S. O., & Caliendo, M. (2007). Sensitivity Analysis for Average Treatment Effects. *The Stata Journal: Promoting Communications on Statistics and Stata*, *7*(1), 71–83.

Horowitz, A. L., Saperstein, A., Little, J., Maiers, M., & Hollenbach, J. A. (2019). Consumer (dis-)interest in genetic ancestry testing: the roles of race, immigration, and ancestral certainty. *New Genetics and Society*, 1–30.

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *Journal of the National Cancer Institute*, *22*(4), 719–748.

TABLE A1: Correspondence between races and ancestries drawn from official racial category definitions provided by OMB (1997) used to create a measure of race-unique ancestries.

Race responses	White	Black	Asian	Hispanic or Latino	Native Hawaiian or Pacific Islander (NHPI)	American Indian
Corresponding Ancestries	• Western Europe • Eastern Europe • Southern Europe • Scandinavia • North Africa • Middle-East	• Sub-Saharan Africa • African American	• East Asia • Southeast Asia • South Asia	• Central or South America	• Pacific Islands	• American Indian

TABLE A2: Descriptive sample comparison to the American Community Survey

Percent of sample in demographic categories				
	Our survey			
	No GAT	GAT	Total	**2015 ACS**
Sex				
Female	79.8	77.0	79.6	50.3
Nativity				
Foreign-born	7.1	6.4	7.0	17.2
Age				
18 to 24	13.4	8.8	13.1	15.6
25 to 34	34.3	31.7	34.2	21.8
35 to 44	26.9	28.6	27.0	20.4
45 to 54	17.5	20.2	17.6	21.6
55 to 64	7.9	10.7	8.1	20.5
Education				
Did not finish high school	0.4	0.3	0.4	11.9
High school grad	18.2	11.9	17.9	51.1
Associate's degree	15.6	12.2	15.4	8.4
Bachelor's degree	37.0	38.5	37.1	18.7
Grad/Professional degree	28.8	37.1	29.3	9.9
Region				
Midwest	23.9	19.6	23.7	21.0
Northeast	17.0	17.5	17.0	17.8
South	34.6	33.0	34.5	37.5
West	24.5	29.9	24.8	23.8

Note: ACS percentages shown are unweighted. Both samples are restricted to 18-64 year-olds to match age restrictions in the marrow donor registry.

TABLE A3: Probit regression predicting whether or not a person took a GAT, as used in propensity-score matching

	Coefficient	Std. Err.	z	P>z	[95% Conf.	Interval]
Female	-0.0810682	0.0167283	-4.85	0	-0.113855	-0.0482814
Foreign born (ref = US-born)	0.0596321	0.0279307	2.14	0.033	0.0048889	0.1143754
Age (ref = 55 to 64)						
18 to 24	-0.131413	0.0318468	-4.13	0	-0.1938317	-0.0689944
25 to 34	-0.0738521	0.0259685	-2.84	0.004	-0.1247494	-0.0229548
35 to 44	-0.0560388	0.026243	-2.14	0.033	-0.1074742	-0.0046034
45 to 54	-0.0386863	0.0276663	-1.4	0.162	-0.0929113	0.0155387
Education (ref = grad/professional degree)						
Did not finish HS	-0.2248703	0.1235671	-1.82	0.069	-0.4670573	0.0173168
High School	-0.3136075	0.0226788	-13.83	0	-0.3580571	-0.2691579
Associate's Degree	-0.2628131	0.0226472	-11.6	0	-0.3072008	-0.2184255
Bachelor's Degree	-0.0941044	0.0165077	-5.7	0	-0.1264589	-0.0617498
Region (ref = Northeast)						
Midwest	-0.102572	0.0225018	-4.56	0	-0.1466747	-0.0584693
South	-0.0523612	0.0205039	-2.55	0.011	-0.092548	-0.0121744
West	0.0856628	0.0211371	4.05	0	0.0442349	0.1270908
Genealogical research activity (ref = none of these activities)						
Ask family member	-0.2754475	0.0233802	-11.78	0	-0.321272	-0.2296231
Viewed family documents	0.090811	0.0158558	5.73	0	0.0597341	0.1218878
Visited genealogy website	0.6003552	0.0153723	39.05	0	0.5702261	0.6304844
Sent away for official documents	0.45826	0.0212249	21.59	0	0.41666	0.49986
Went to a library	0.1481232	0.0216035	6.86	0	0.1057811	0.1904652
Other behaviors	0.4968048	0.02986	16.64	0	0.4382803	0.5553294
Response timing (ref = initial email)						
Reminder email	-0.0294413	0.0183116	-1.61	0.108	-0.0653314	0.0064488
Second reminder email	-0.1087172	0.0193338	-5.62	0	-0.1466108	-0.0708236
Long recruitment email (ref = short recruitment email)	-0.0183411	0.0139074	-1.32	0.187	-0.0455991	0.0089168
Constant	-1.502601	0.0387851	-38.74	0	-1.578618	-1.426583

Note: n = 100,885.

TABLE A4: Results from propensity-score matching predicting multiple-race reporting

Sample	Treated	Controls	Difference	S.E.	T-stat
Unmatched	0.1430	0.1132	0.02979	0.004435	6.72
ATT	0.1430	0.1161	**0.02691**	0.009262	2.91
ATU	0.1132	0.1599	0.04665	.	.
ATE			0.04559	.	.

Note: N = 100,885. Standard errors shown do not take into account that the propensity score is estimated.

TABLE A5: Mantel-Haenszel bounds sensitivity analyses for propensity scoring matching model predicting multiple race reporting

Gamma	Q_mh+	Q_mh-	p_mh+	p_mh-
1	3.10978	3.10978	0.000936	0.000936
1.05	2.45298	3.76828	0.007084	0.000082
1.1	1.82741	4.39728	0.033819	5.50E-06
1.15	1.23026	4.99978	0.1093	2.90E-07
1.2	0.658907	5.57815	0.254978	1.20E-08
1.25	0.111069	6.13447	0.455781	4.30E-10
1.3	0.340627	6.67056	0.366692	1.30E-11
1.35	0.847039	7.18801	0.198487	3.30E-13
1.4	1.33522	7.68825	0.090902	7.40E-15
1.45	1.80655	8.17254	0.035416	1.10E-16
1.5	2.26225	8.64201	0.011841	0

Gamma : odds of differential assignment due to unobserved factors
Q_mh+ : Mantel-Haenszel statistic (assumption: overestimation of treatment effect)
Q_mh- : Mantel-Haenszel statistic (assumption: underestimation of treatment effect)
p_mh+ : significance level (assumption: overestimation of treatment effect)
p_mh- : significance level (assumption: underestimation of treatment effect)

TABLE A6: Logistic Regression Predicting American Indian Ancestry Reporting Among Respondents Who Identified as White (Odds Ratios)

	(1)	(2)	(3)	(4)
Taken a GAT (ref = no GAT)	0.832*** (0.036)	0.795*** (0.035)	0.869* (0.057)	0.878* (0.054)
Ancestry before race condition (ref = race before ancestry)		0.999 (0.019)	1.007 (0.020)	0.999 (0.019)
Ancestry before race condition x Taken a GAT			0.853 (0.075)	
Unprimed condition (ref = knowledge prime)		0.988 (0.019)	0.988 (0.019)	0.997 (0.019)
Unprimed condition x Taken a GAT				0.822* (0.072)
Genealogical research				
Asked family member (ref = did not ask family member)		1.456*** (0.055)	1.456*** (0.056)	1.455*** (0.055)
Viewed family documents (ref = did not view family documents)		1.063** (0.022)	1.063** (0.022)	1.063** (0.022)
Visited genealogy website (ref = did not visit genealogy website)		1.199*** (0.025)	1.199*** (0.025)	1.199*** (0.025)
Sent away for official documents (ref = did not send away for docs.)		1.043 (0.038)	1.043 (0.038)	1.043 (0.038)
Went to a library (ref = did not go to a library)		1.160*** (0.039)	1.160*** (0.039)	1.161*** (0.039)
Other research activities (ref = did not do other research)		1.230*** (0.062)	1.229*** (0.062)	1.231*** (0.062)
Female (ref = male)		1.247*** (0.031)	1.247*** (0.031)	1.247*** (0.031)
Foreign born (ref = US born)		0.396*** (0.027)	0.396*** (0.027)	0.396*** (0.027)
Age (ref = 55-64)				
18 to 24		1.771*** (0.082)	1.771*** (0.082)	1.773*** (0.082)
25 to 34		1.881*** (0.078)	1.881*** (0.078)	1.881*** (0.078)

35 to 44		1.692***	1.692***	1.693***
		(0.071)	(0.071)	(0.071)
45 to 54		1.182***	1.182***	1.183***
		(0.053)	(0.053)	(0.053)
Education (ref = Grad/Professional Degree)				
Did not finish HS		1.634**	1.635**	1.634**
		(0.261)	(0.261)	(0.261)
High School		1.618***	1.618***	1.618***
		(0.047)	(0.047)	(0.047)
Associate's Degree		1.609***	1.609***	1.610***
		(0.047)	(0.047)	(0.047)
Bachelor's Degree		1.115***	1.115***	1.115***
		(0.028)	(0.028)	(0.028)
Region (ref = Northeast)				
Midwest		1.257***	1.257***	1.257***
		(0.042)	(0.042)	(0.042)
South		2.248***	2.248***	2.248***
		(0.069)	(0.069)	(0.069)
West		1.527***	1.527***	1.527***
		(0.051)	(0.051)	(0.051)
Observations	87070	87070	87070	87070

Sample is restricted to respondents who selected "White" as a race response either alone or with another race (results are similar if restricted to monoracial White respondents). Models 2, 3, and 4 also control for recruitment variables (not shown). Standard errors in parentheses.
* p<0.05 **p<0.01 ***p<0.001

TABLE A7: Percent of respondents reporting multiple races among those who report multiple race-unique ancestries, by amount of genealogical research (Panels B, C, and D), age category, and whether the respondent took a GAT.

Panel A. All respondents (n = 22,631)

	GAT	No GAT	Diff.	*p*
18 to 24	53.8	44.8	9.1	0.0258
25 to 34	51.1	38.9	12.2	0
35 to 44	40.5	34.1	6.4	0.0100
45 to 54	34.9	30.9	3.9	0.2296
55 to 64	43.5	29.0	14.5	0.0036
All ages	45.3	37.0	8.2	0

Panel B. Very little research (n = 9,456)

	GAT	No GAT	Diff.	*p*
18 to 24	65.1	45.6	19.5	0.0114
25 to 34	57.3	39.5	17.8	0.0003
35 to 44	34.7	34.3	0.4	0.9476
45 to 54	40.0	30.2	9.8	0.2477
55 to 64	35.7	24.8	10.9	0.3556
All ages	49.2	37.4	11.8	0.0001

Panel C. Some research (n = 9,762)

	GAT	No GAT	Diff.	*p*
18 to 24	45.9	43.4	2.6	0.6339
25 to 34	49.3	38.4	10.9	0.0003
35 to 44	38.2	32.9	5.3	0.1388
45 to 54	33.7	29.6	4.0	0.4266
55 to 64	44.4	28.9	15.5	0.0516
All ages	42.6	36.2	7.4	0.0001

Panel D. A lot of research (n = 3,413)

	GAT	No GAT	Diff.	*p*
18 to 24	61.5	46.5	15.0	0.1377
25 to 34	50.0	39.0	10.9	0.0219
35 to 44	46.7	36.9	9.8	0.0297
45 to 54	34.3	35.4	-1.1	0.8355
55 to 64	45.2	35.8	9.4	0.2423
All ages	45.5	38.4	7.1	0.0056

TABLE A8: Rates of reporting multiple ancestries, multiple race-unique ancestries, 2 or more races, and 3 or more races among people who have not taken a GAT, took a GAT themselves, or reported that they had not taken a GAT themselves but had seen a relative's results.

	Took a GAT	**Viewed a relative's GAT results**	**Did not take a GAT**
Multiple ancestries	3,375 (66.3%)	142 (54.6%)	48,684 (54.1%)
Multiple race-unique ancestries	1,297 (25.4%)	53 (20.4%)	21,281 (23.3%)
Two or more races	748 (14.4%)	33 (12.5%)	10,804 (11.3%)
Three or more races	167 (3.2%)	3 (1.1%)	1,389 (1.5%)

Note: column percentages in parentheses

FIGURE A1: Survey question and response options for ancestry reporting.

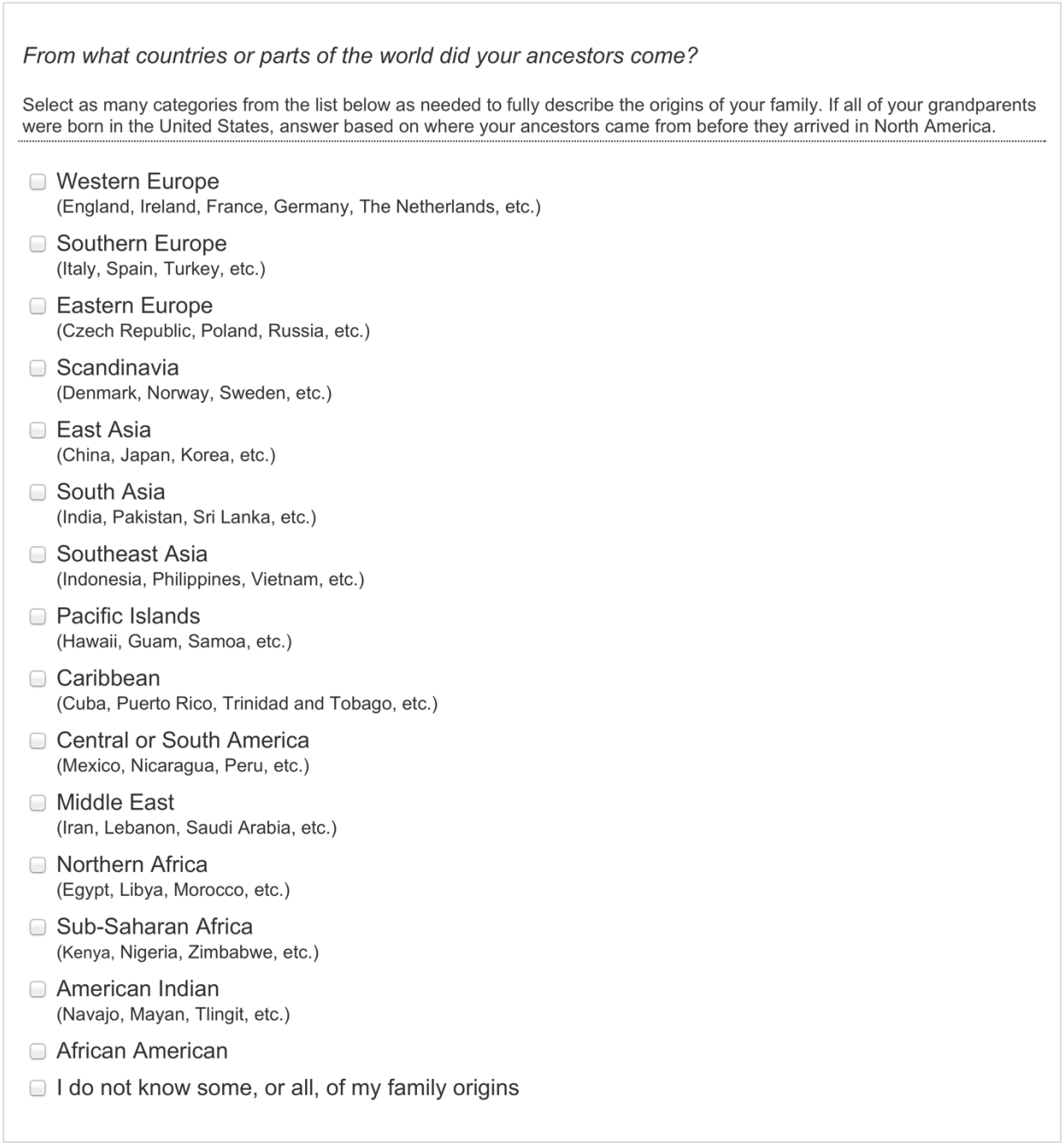


FIGURE A2: Overlap (panel a) and balance (panel b) distributions for our propensity score analysis.

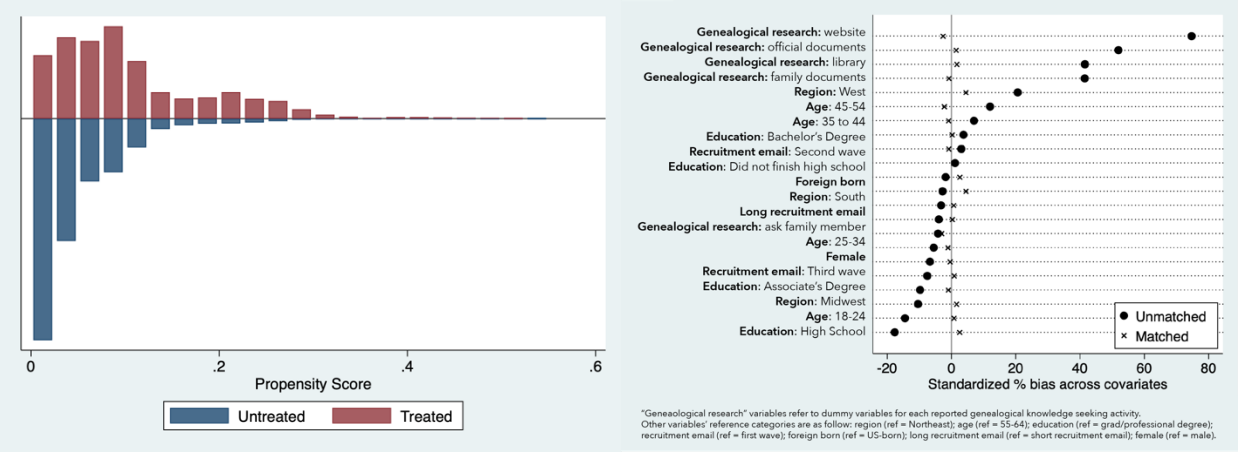
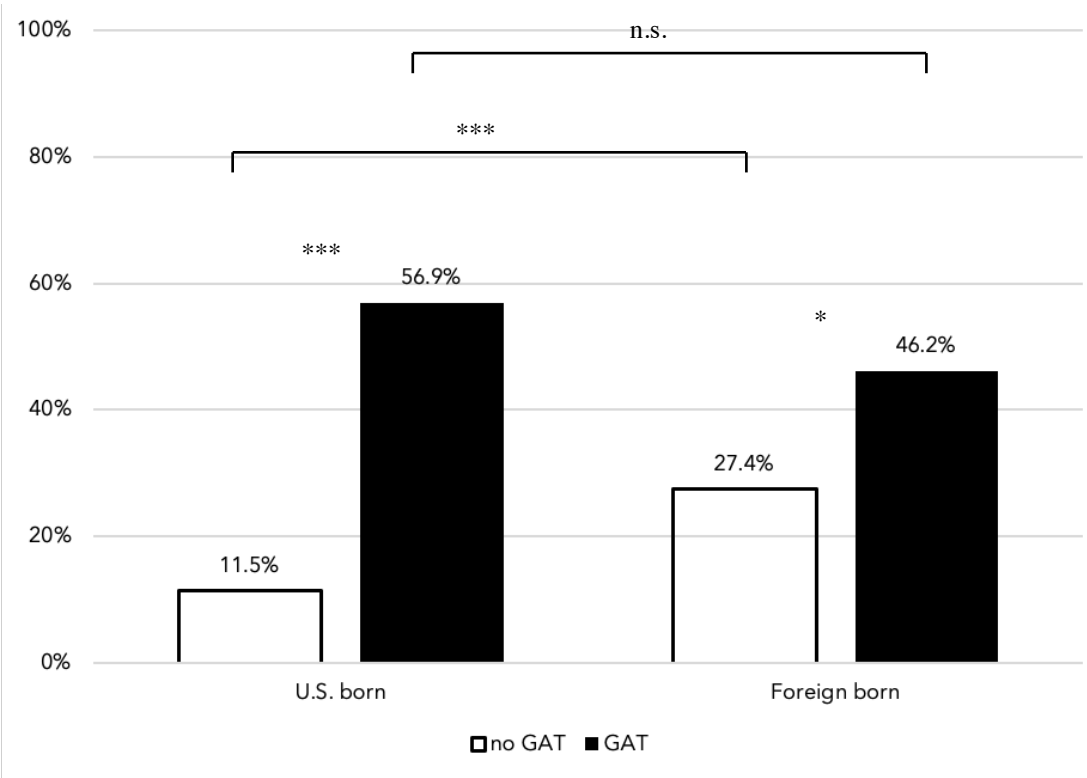
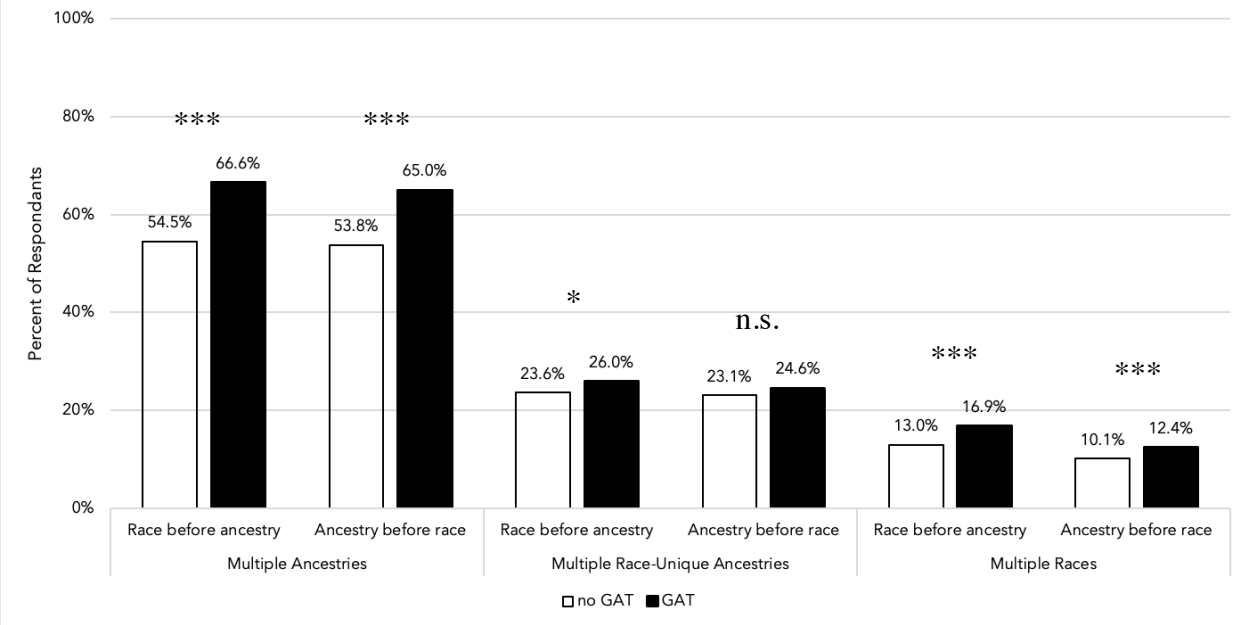


FIGURE A3: Percent who list Sub-Saharan African ancestry among respondents who identified as Black, by nativity and whether they took a GAT.



Note: * p<0.05 **p<0.01 ***p<0.001

FIGURE A4: Rates of selecting multiple ancestries, multiple race-unique ancestries, and multiple races by whether a respondent has taken a GAT and whether they saw race questions before or after ancestry questions.



Note: * p<0.05 **p<0.01 ***p<0.001