Technical Report: Duplicate Respondents + Asian and Hispanic subsamples
Date: 06/09/2022
Authors: Pia Deshpande and Jeremiah Cha

The objective of this project is to explore the political characteristics of Asian and Hispanic ethnic subgroups as well as key differences by generation within these samples. Using the 2020 Cooperative Election Study (CES), we find that Asians and Hispanics, albeit in different ways, diverged from other racial identifiers in age, education, regional distribution, and immigration status. While both Asians and Hispanics are younger than the rest of the country, there are stark educational disparities between the two groups. Ideologically, the distribution of non-white identifiers resembled one another but split from their white counterparts. Unlike white respondents, a plurality of Black, Asian, and Hispanic respondents are "middle of the road." Larger shares of these groups identify as liberal.

Although the CES boasts 60,000 observations, sample size issues for intergenerational analyses within ethnic subgroups still exist. Below, we will detail the specific sample size issues for any individual ethnic subgroup. Asian and Hispanic subgroups presented several challenges, especially as the sample was divided up multiple times for demographic breaks. This is troubling for researchers, especially because compared to the GSS or the ANES, the CES has far more Hispanic and Asian respondents, meaning obstacles we find in this report may exist in other national surveys with smaller ethnic minority samples.

To potentially ameliorate these issues, we attempted to combine the 2020 CES with the 2018 and 2016 CES samples. However, YouGov **alerted us that there are a sizable number of respondents who answered surveys in both 2018 and 2020. 25% of 2020 respondents also took the 2018 survey, for a total of 15k respondents.**[1] We attempted to prune these duplicate observations as YouGov cannot identify the repeat respondents due to PII considerations. We attempted to prune the sample using covariates, which proved inconclusive. You may read more about this and its implications for analysis in the Appendix.

Prior to conducting analyses, we compared the 2020 CES with other large-scale surveys (e.g. Pew, GSS) and Census data (e.g. American Community Survey). In general, the CES sample resembled the ACS 2019 5-year data with regard to key demographic breaks. We are still working through comparisons with the November 2020 Current Population Survey (CPS) and its voter supplement, but expect similar results concerning voter registration and behavior.

The rest of the memo is organized as follows. We begin with the results of our validation against the ACS 2019 5-year data and other large surveys. We follow this up with descriptive statistics on Asian and Hispanic samples in the 2020 CES. Finally, we review potential avenues for further research on Asian and Hispanic respondents in the CES.

Validation

American Community Survey (2019, 5 year)[2]

Validation was done with the ACS microdata from IPUMS. All variables were at the person level (as opposed to household level), so person-weights were applied. The 2019 ACS is

[1] This information was provided to us by YouGov, which can only document duplicates from YouGov panels. It is possible that there are slightly more duplicates than this estimate, but YouGov cannot tell who these individuals are.
[2] ACS data is subsetted to the universe of respondents ages 18+

compared to the 2020 CES (non-voter-validated) data. The 2020 CES voter-validated data will be compared with the CPS supplement.

Overall, the 2020 CES sample of Asians and Hispanics lines up well with Census estimates, albeit with some discrepancies. Regarding the total share of Asians and Hispanics, the 2020 CES has a slightly smaller proportion of both Asians and Hispanics. When compared to the samples of other large surveys like Pew and the GSS, the 2020 survey captures a more comparable percentage of Asians and Hispanics.

The 2020 CES performs well when compared to the 2019 ACS across demographic breaks, except concerning education and regional distribution. Table 1 shows comparisons between the two surveys with highlighted cells emphasizing categories where the two studies deviated by 5 or more percentage points. Among both Asians and Hispanics, those who had high school educations or less were undersampled. Moreover, about a 5 percentage point difference exists when comparing Hispanics in the Western Census region.

Table 1: 2020 CES and 2019 ACS

	2020 CES	2019 ACS		2020 CES	2019 ACS
Asian	4.2%	5.8%	**Hispanic**	13.1%	15.9%
Male	44%	47%	Male	49%	49.7%
Female	55%	53.0%	Female	51%	50.2%
18-29	23%	22.3%	18-29	32%	28.7%
30-44	32%	31.1%	30-44	28%	32.3%
45-64	31%	31.2%	45-64	28%	28.6%
65+	14%	15.4%	65+	13%	10.4%
HS or less	21%	28%	HS or less	49%	58.1%
Some college	23%	21.9%	Some college	33%	27.3%
College grad	34%	28.8%	College grad	13%	10.2%
Postgrad	22%	21.0%	Postgrad	5%	4.4%
Midwest	13%	11.8%	Midwest	9%	8.6%
Northeast	21%	19.8%	Northeast	17%	14.3%
South	25%	22.7%	South	40%	37.6%
West	41%	45.7%	West	34%	39.4%

In terms of ethnic subgroup analyses, the samples in the ACS and the 2020 CES are not directly comparable The 2020 CES includes an option for respondents to identify as from the United States, which both Asian and Hispanic respondents often choose. This diverges from the ACS options for their RACE and HISPAN variables, which does not include the U.S. as an option. As a result, the two diverge from each other, albeit with worse results for Hispanic respondents. Other ACS variables such as ANCESTR1, which asks about ancestry or ethnic origin, are worse comparison variables. For example, only about 0.225% of Asians identify as having ancestry or ethnic origin in the United States.

Table 2: Subgroup proportions of total sample for the 2020 CES and the 2019 ACS

	Asian		Hispanic		
Country of Origin	CES	ACS	COO	CES	ACS
China	27.7%	22.6%	Mexico	43.1%	60.2%
Philippines	14.1%	18.9%	Puerto Rico	17.3%	10.0%
India	14.3%	20.5%	Spain	17.6%	2.0%
			South America	6.4%	7.0%
United States	**13.0%**		United States	**33.5%**	

Below are the weighted sample sizes for our subgroups of interest in the common and voter validated 2020 CES. There are robust Hispanic subpopulations, particularly among Mexican and Spanish respondents. However, subsetting across demographic breaks or looking at voter-validated data would pose problems. The n's for Asian subpopulations are much smaller (with the tenuous exception of Chinese respondents). Though they are more robust in Hispanic subpopulations, voter-validated data and further subsetting across demographic breaks would pose problems as well.

Table 3: Weighted Ns -- Asians (2020 CES)

Subgroup	**Frequency**	**Frequency (VV)**
China	718.7	402.3
India	370.6	161.9
Philippines	365.8	156.0
United States	336.7	125.9

Table 4: Weighted Ns -- Hispanics (2020 CES)

Subgroup	**Frequency**	**Frequency (VV)**
Mexico	3375.4	1604.7
Puerto Rico	1351.4	662.6
South America	500.8	257.6
Spain	1381.8	906.5
United States	2624.5	1174.3

Descriptive Analysis

Immigration status

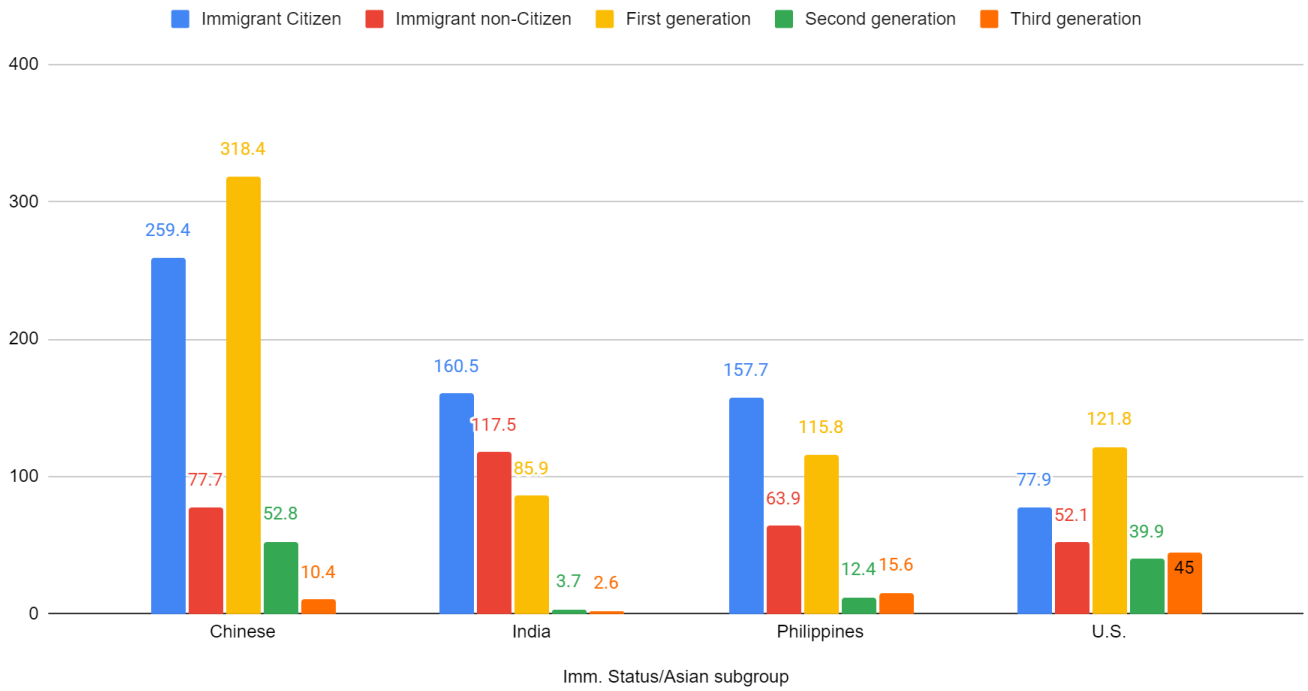
Table 4 details the breakdown of immigration status across Asian and Hispanic respondents. When compared to the national sample, it is apparent that both groups have larger citizen and non-citizen immigrant populations. Indeed, the starkest difference can be found in the third generation category, which is attributable to the relative novelty of immigration, especially from Asian countries.

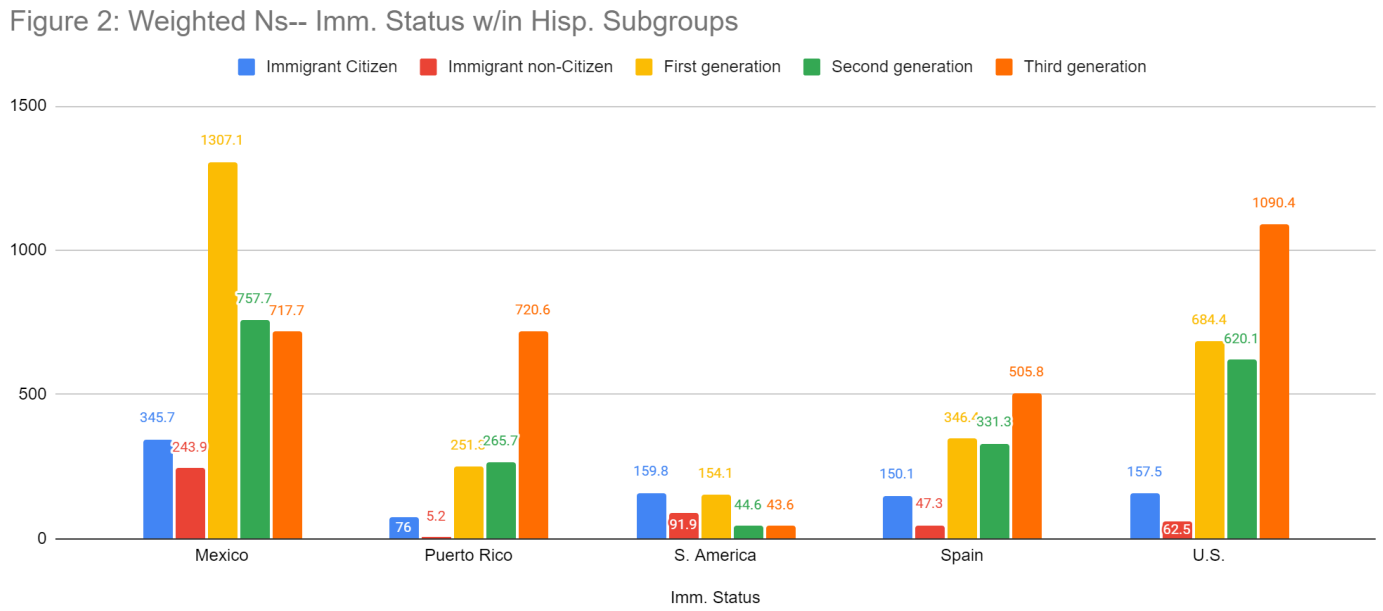
Table 5: Breakdown of Immigrant Status

Immigrant Background	**All**	**Asian**	**Hispanic**
Immigrant Citizen	6%	39%	14%
Immigrant non-Citizen	3%	17%	8%
First generation	10%	32%	30%
Second generation	18%	7%	19%
Third generation	63%	5%	29%

Figures 1 and 2 present the weighted Ns for Asian and Hispanic subgroups of interest.

Figure 1: Weighted Ns-- Imm. Status w/in Asian Subgroups



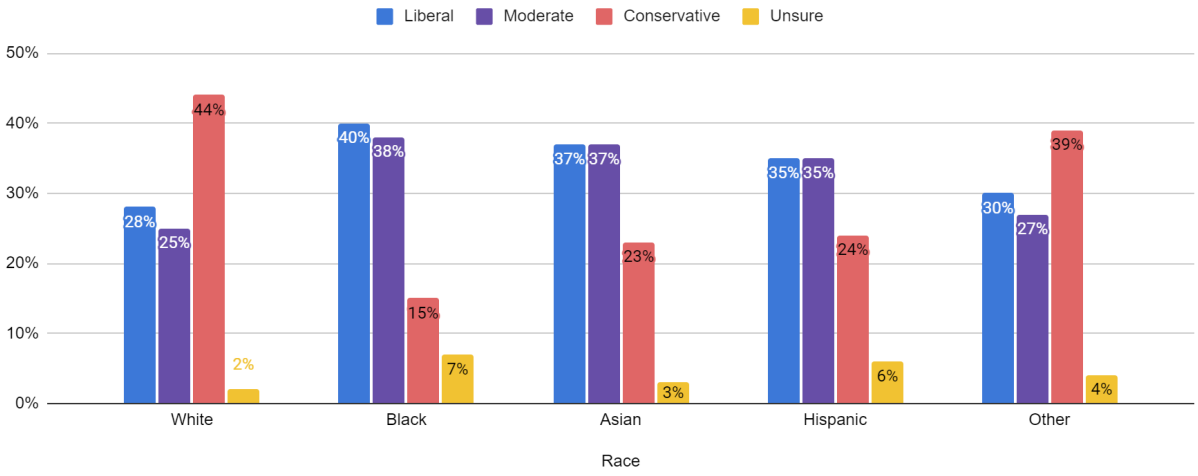


The distinctions the CES makes regarding immigration history splices the data into relatively small categories, particularly for Asian subgroups (even Chinese respondents, see the weighted N for third generation Chinese participants). The story is a bit better for Hispanic respondents, with the subgroups of those who listed the United States or Mexico as countries of origin being robust enough for potential further analysis. For example, there are only 10.4 respondents who identify as third-generation Chinese Americans. Making conclusions about generational differences with respect to ideology or other political attitudes would be difficult, considering these small sample sizes. Although it is possible to produce statistics about Asians and Hispanics broadly along these breaks, heterogeneous immigration experiences (e.g. groups that entered the U.S. following the 1965 Immigration Act versus those with longstanding ties in the country) make these conclusions difficult to generalize to any given ethnic group.

Political ideologies

Figure 3 details the distribution in ideology among validated voters in the 2020 CES. Asian and Hispanic voters generally mirror each other and Black identifiers quite closely, although deviating slightly with regard to the share of conservative identifiers. Notably, Asian voters are the surest about their ideological leanings with only about 3 percent of respondents indicating they are unsure about their ideological identification. Among all three major minority groups, conservative identifiers are the smallest group. This is the direct opposite of their white counterparts, where a plurality of respondents identify as conservatives.

Figure 3: Ideological distribution by race



Voting record

Below is an overall breakdown of voting behavior from the five major racial groups and the two sexes polled by the CES (this is replicated from the demographic table included in the 2020 CES guide):

Table 6: 2020 Presidential Vote by Demographic Group				
	Group	**% Biden**	**% Trump**	**% of Electorate**
Sex				
	Male	46%	51%	47%
	Female	55%	43%	53%
Race				
	White	43%	55%	72%
	Black	89%	10%	11%
	Hispanic	64%	33%	10%
	Asian	69%	30%	4%
	Other	43%	53%	3%

Diving further into this data can be done, but only for a limited selection of groups. Among Asians, only Chinese Americans have a sample size larger than 300. The sample size of Hispanic respondents among voter-validated respondents is larger in comparison. Mexico and Hispanic United States identifiers voted heavily in favor of Biden.

Hispanic: Mexican respondents

Table 7: Presidential Vote Choice, Mexican respondents	
Presidential Vote Choice	**Frequency**
Biden	66.2%
Other	3.2%
Trump	30.5%

Hispanic: United States respondents

Table 8: Presidential Vote Choice, Hispanic: U.S.-identified respondents	
Presidential Vote Choice	**Frequency**
Biden	60.5%
Other	3.9%
Trump	35.5%

Hispanic: Spanish respondents

Table 9: Presidential Vote Choice, Hispanic: Spanish respondents	
Presidential Vote Choice	**Frequency**
Biden	57.7%
Other	3.0%
Trump	39.2%

Hispanic: Puerto Rican respondents

Table 10: Presidential Vote Choice, Hispanic: Puerto Rican respondents	
Presidential Vote Choice	**Frequency**
Biden	66.0%
Other	3.3%
Trump	30.6%

Asian: Chinese respondents

Table 11: Presidential Vote Choice, Asian: Chinese respondents	
Presidential Vote Choice	**Frequency**
Biden	73.2%
Other	1.1%
Trump	25.7%

Future work

Increasing sample size

To do any meaningful deep-dives into the relationship between immigration status, ideology, and voting records among Asians and Hispanics, a stacked CES sample across 2018 and 2020 is likely necessary.[3] However, there are some respondents who take the CES multiple years in a row. Though these respondents are few and far between and don't make a meaningful difference in aggregate analysis, it creates a very difficult problem to overcome in subgroup analysis. We cannot know for sure who repeat respondents are, because respondents are entitled to their privacy. Among 2020 CES respondents, there are 900 Hispanic respondents who also took the 2018 CES compared to 350 duplicate Asian respondents. This includes 100 Chinese, 60 Indian, 40 Filipino, 450 Mexican, 200 Puerto Rican, and 100 South American. When subgroup n's are already small, these duplicates pose a real problem.

Although our efforts to identify these double respondents ourselves for pruning have produced mixed results, a successful stack would give us the necessary power to make conclusions about generational differences among Asian and Hispanic ethnic subgroups. Pruning on certain covariates may have a negligible effect on certain subgroups, but increasing our effective sample size would allow us to answer questions in the literature about socialization and differences across generations with regard to political development. It is also possible we could look into pursuing differential privacy, which may let us filter out these respondents and still find meaningful results at the larger subgroup level.

It is possible that we could look into using multilevel regression and poststratification (MRP) to interpolate more data and conduct public opinion analyses. However, MRP provides its users no way to gauge the level of uncertainty associated with each new estimate. Moreover, none of the researchers involved in this project have experience with MRP, but we would welcome feedback on whether or not this is a path we should explore.

Survey validation/accuracy metrics

The entire process of validating the CES against other surveys has piqued our interest in how accurate surveys are when generalizing their results to Asian or Hispanic populations. Most of

[3] We cannot stack past the 2018 CES, since the options available to respondents regarding immigration status are markedly different in their wording.

the work we have done so far is a thorough audit of how the CES compares to other surveys and the 2019 ACS, and we hope this work can be used to inform recruiting Asian and Hispanic respondents in the future.

Right now the solution to our data woes seems to be an expensive one: studying minority populations effectively and with detail takes targeted sampling. There aremany unexplored questions that even the CES, with its 60,000 respondents and sizable Asian and Hispanic respondent pools, could not answer.

1. How does ideology relate to immigration status and ethnic subgroups? Are Chinese Americans whose families immigrated to the United States prior to the Asiatic Barred Zone different from Chinese Americans whose families immigrated after the 1965 immigration liberalization? If so, why?
2. What are we systematically not capturing when clumping together heterogeneous Hispanic and Asian subpopulations? It is true that smaller Asian subpopulations tend to be poorer, less educated, and less politically active than the two largest Asian subgroups in the United States by far: Indian and Chinese Americans? What does this mean about our estimates of Asian public opinion? How likely are those other groups to respond to a survey request?

There are methodological complications even in sample validation. Focusing specifically on Asian Americans, the CES nationality question allows for respondents to select “United States.” We think this is a substantively good and interesting measure - if respondents are listing themselves as Asian and ascribing their nationality to the United States, it tells us a lot about how long it may take Asian immigrants to feel like a part of the United States (Lien et al. 2003). However, the Census includes no such category for its Asian Americans, so comparing the subgroup breakdowns of Census Asians and CES Asians are somewhat stymied. Comparing within groups (gender breakdown of CES Chinese respondents vs. Census Chinese respondents) is still possible.

Appendix: Methodological Note

Unfortunately, the results of the pruning were inconclusive. We attempted to prune double respondents on categories, such as age (calculated with the variable *birthyear*), sex, and zip code. We also performed this analysis with other permutations of demographic characteristics, such as education and age. These attempts did not prune enough Hispanic respondents in particular. As a result, we kept our analysis to the 2020 CES. *This means researchers should exercise caution when stacking CES samples to analyze Asians and Hispanics, and should not examine Asian subgroups between years at all.*

We attempted other ways to prune the sample for duplicates and could judge our accuracy using YouGov's reported numbers — 900 Hispanic repeat respondents and 350 Asian repeat respondents between CES 2018 and CES 2020. The permutation using age, sex, and zip code proved to be the most accurate for finding duplicate Hispanic respondents. YouGov also provided us with the number of duplicates for some ethnic subgroups. We report two attempts at pruning below, one of which includes country of origin and one of which does not.

Pruning with age, sex, zip code, and country of origin

Table A1: Pruning Duplicate Respondents (Asians)

Country of Origin	**Stacked N**	**YouGov Duplicates**	**Pruning with age, sex, zip code, and country of origin**	**Pruning with age, sex, zip code**
China	795	100	122	134
India	448	60	55	57
Philippines	332	40	70	75

Table A2: Pruning Duplicate Respondents (Hispanics)

Country of Origin	**Stacked N**	**YouGov Duplicates**	**Pruning with age, sex, zip code, and country of origin**	**Pruning with age, sex, zip code**
Mexico	3346	450	250	353
Puerto Rico	1423	200	126	175
South America	756	100	66	90

Appendix Tables

Table A3: Weighted Ns -- Ideology w/in Asian subgroups

Ideology	Asian subgroup	Frequency
Conservative	China	128.0
Liberal	China	201.0
Moderate	China	331.5
Unsure	China	58.2
Conservative	India	61.9
Liberal	India	112.1
Moderate	India	172.0
Unsure	India	24.5
Conservative	Philippines	125.0
Liberal	Philippines	100.2
Moderate	Philippines	112.0
Unsure	Philippines	27.7
Conservative	U.S.	81.2
Liberal	U.S.	108.8
Moderate	U.S.	102.7
Unsure	U.S.	43.9

Table A4: Weighted Ns -- Ideology w/in Hisp. subgroups

Ideology	Hisp. subgroup	Frequency
Conservative	Mexico	701.4
Liberal	Mexico	963.0
Moderate	Mexico	1188.2
Unsure	Mexico	520.2
Conservative	Puerto Rico	264.1
Liberal	Puerto Rico	365.0
Moderate	Puerto Rico	523.9
Unsure	Puerto Rico	198.4
Conservative	S. America	81.4
Liberal	S. America	178.1
Moderate	S. America	164.0
Unsure	S. America	77.2
Conservative	Spain	350.5
Liberal	Spain	399.9
Moderate	Spain	512.6
Unsure	Spain	118.8
Conservative	U.S.	576.4
Liberal	U.S.	748.0
Moderate	U.S.	811.8
Unsure	U.S.	487.3