Title

Evaluating the Accuracy of 2020 Census Block-Level Estimates in California

Authors

Bozick, Robert

Burgette, Lane F

Sharygin, Ethan

et al.

Publication Date

2023-11-27

DOI

10.1215/00703370-11075209

Peer reviewed

Evaluating the Reliability of 2020 Census Block-Level Estimates in California*

Robert Bozick
Rice University

Lane Burgette
RAND Corporation

Ethan Sharygin
Portland State University

Beverly Weidmer
RAND Corporation

Michael Tzen
University of California-Los Angeles

Regina Shih
RAND Corporation

Aaron Kofner
RAND Corporation

Jennie E. Brand
University of California-Los Angeles

Hiram Beltran-Sanchez
University of California-Los Angeles

*This study was funded by the California Complete Count Committee – Census 2020 Office and supported by the State of California's Department of Finance. All analyses and interpretations are the authors alone and do not express the opinions of the State of California. Please direct all correspondence to Robert Bozick, Houston Population Research Center, Kinder Institute for Urban Research, Rice University, 6100 Main Street, Suite 305, Houston, TX 77005-1892, rbozick@rice.edu

Evaluating the Accuracy of 2020 Census Block-Level Estimates in California

In this study, we provide an assessment of data accuracy from the 2020 Census in California. We compare block-level population totals from a sample of 173 census blocks in California across three sources: (1) the 2020 Census, which has been infused with error to protect respondent confidentiality; (2) the California Housing and Population Sample Enumeration, the first ever independent enumeration survey of census blocks; and (3) projections based on the 2010 Census and subsequent American Community Surveys. We find that, on average, total population counts provided by the U.S. Census Bureau at the block level for the 2020 Census are not biased in any consistent direction. However, sub-population totals defined by age, race, and ethnicity are highly variable and are unlikely to accurately characterize the demographic composition of census blocks. Additionally, we find that inconsistencies across the three sources are amplified in large blocks defined in terms of land area or by total housing units, blocks in suburban areas, and blocks that lack broadband access. We conclude by discussing the implications of our findings for researchers using block-level data from the 2020 Census and for those at the U.S. Census Bureau planning future censuses.

INTRODUCTION

Censuses, which include a complete and total enumeration of a population at a single time point, are indispensable to both formal and social demography. With census data, formal demographers can measure population statics, including the size, distribution, and structure of a population. With multiple censuses conducted over time, formal demographers can characterize population dynamics (Roach and Carey 2020). Social demographers, on the other hand, use

census data to identify the correlates and consequences of both population statics and dynamics. Social scientists from many disciplines also rely on parameters from census data to generate population distributions for weighting sample surveys. Therefore, should censuses be inaccurate, they can have widespread consequences for demography and social science.

In the U.S., the first census took place in 1790 and was thereafter conducted every ten years per a Constitutional mandate. Though every attempt to enumerate a diverse population across a large land area like the U.S. involves challenges that are distinct to the historical period in which it is undertaken, the 2020 Census in the U.S. faced a particularly unusual set of circumstances for censuses in the modern era. Most notably, this was the first census to include an option to participate online, it was preceded by a contentious case brought before the Supreme Court with the intent of suppressing participation among immigrant communities, and it was conducted during a global health pandemic. Additionally, new post-enumeration measures to protect the privacy of individuals have elevated concerns about the accuracy of the data at lower levels of geography, such as census blocks. Given these challenges, there remains great concern about the quality of the 2020 Census and, by extension, demographic analyses based on the resulting data (Ruggles et al. 2019).

In this study, we evaluate the quality of 2020 Census data at the block level across 173 census blocks in California. To do so, we compare enumeration totals from the 2020 Census on these 173 blocks with estimates from two independent sources: (1) a survey administered to all households on those blocks, and (2) population projections based on the 2010 Census and subsequent American Community Surveys. The former source is the first ever independently conducted survey designed to replicate and validate census data collection at scale in the U.S.. The latter source uses an array of demographic and geographic data sets collected and updated

across the intercensal period to estimate the population at the time of the 2020 Census. With data from the 2020 Census as well as from these two independent sources, we can assess the comparability of population estimates across our 173 sample blocks. Additionally, we attempt to identify barriers to enumeration that may have contributed to discrepancies across the different data sources. California is by far the nation's most populous state. More than 12 percent of Americans are Californians. The state's population is diverse, with no racial or ethnic group constituting a majority. Over 10 million Californians are immigrants. The results from our study provide insight into how well the 2020 Census enumerated the population of California, the viability of using block-level data infused with error to protect individual identities, and consequently, what considerations researchers may want to account for when using 2020 Census data for demographic analysis.

In what follows, we first provide a brief overview of the context and unique challenges faced by the U.S. Census Bureau in undertaking the 2020 Census. We then describe the implications of those challenges for our study when enumerating California, which has the largest and most diverse population of any state and provide a brief overview of methods traditionally used in evaluating the accuracy of the decennial census. Next, we list our research questions, describe our data sources and methods, and then present the results of our analysis. We conclude with a discussion of the implications of our findings for analyzing and interpreting block-level data from the 2020 Census.

Challenges Facing the Collection and Analysis of the 2020 Census

The 2020 Census faced challenges that were both anticipated and unanticipated. Regarding anticipated challenges, the U.S. Census Bureau changed the data collection

methodology between the 2010 and 2020 Censuses. Specifically, the 2020 Census was the first in which individuals could respond online. Most households were invited to participate online as the first option, with traditional paper and pencil surveys delivered by mail, interview attempts by phone, and in-person follow-up visits deployed should households refuse to respond to the original online invitation. During these in-person follow-up visits, data collectors relied on smartphone technology to conduct the interview and broadband access to transmit the data successfully. While online surveys and digital technology have been used for many years in the survey research community, this was the first time they were simultaneously implemented in a complete population enumeration at scale. Because of inadequate funding, the U.S. Census Bureau had to drastically scale back its field testing of these technologies, which raised the risk of basic functionality problems, connectivity failures, and cybersecurity threats (Lapowsky 2019).

In addition to anticipated potential technological challenges, there were unanticipated social ones, namely, attempts to suppress the participation of both documented and undocumented immigrants. In the U.S., the census counts all individuals living in the country on April 1st of the census year regardless of whether they are citizens or non-citizens. The Trump administration acted to include a question on the 2020 Census that would ask respondents to report their citizenship status. Research showed a citizenship question would likely deter the participation of immigrant communities and contribute to an undercount (Brown et al. 2018). The Supreme Court ultimately blocked the government from including the citizenship question on the decennial survey (Howe 2019). Still, many critics of the Trump administration's initiative expressed concern that the damage to the credibility of and confidence in the U.S. Census Bureau among immigrant communities may have already been done (Marimow et al. 2019).

These technological changes and court challenges occurred against a broader backdrop of increasing distrust in government in the U.S. Between 2000 and 2020, the percentage of American adults who states they trust the federal government fell by more than half, from 54% to just 20% (Pew Research Center 2020). Distrust in government could translate into Americans hesitant to share personal information with federal agencies, such as the U.S. Census Bureau (Sullivan 2020). Two months into the 2020 Census field operations, a nationally representative survey found that 40% of those who had not yet filled out their census forms expressed an unwillingness to respond to a census worker (Cohn 2020). As this was based on adults willing to respond to a survey, it is likely the 40% estimate among census non-responders is a lower bound estimate of total population hesitancy to participate.

Alongside these challenges, the most serious operational challenge to the 2020 Census was undoubtedly the COVID-19 pandemic, which emerged as a national emergency just as the U.S. Census Bureau began data collection in March 2020. The pandemic delayed fieldwork by approximately three months amidst a flurry of short-term moves and adjustments to living arrangements in response to health and safety concerns of the public as well as to lockdown orders imposed by local and state governments (Supan 2021). Such moves could have substantially affected block-level population totals, particularly for blocks that attracted temporary migrants (such as those in suburban and rural areas that served as short-term havens for teleworkers) or blocks that experienced substantial outmigration (such as those in large, densely populated urban areas where those with the means to leave, even if only temporarily, did so). Once 2020 Census field operations resumed in full after an initial delay, data collectors were required to wear masks and conduct in-person interviews outside from a distance. These safety protocols added further challenges to face-to-face interaction necessary to establish trust and

rapport with hesitant households, while also introducing additional avenues for measurement error (e.g., difficulty hearing during the in-person interview).

All these challenges to data collection are expected to affect data quality in some way. Moreover, the accuracy of the data available to the public is further compromised by new procedures enacted to protect the privacy of survey respondents. A new initiative implemented in 2020, the U.S. Census Bureau's Disclosure Avoidance System (DAS), will now artificially inject noise into its published tables from the decennial census and other federal surveys to prevent the identification of individuals. While top-line population counts at the state level will be unaffected, lower levels of geography will contain error by design and, in some cases, produce erroneous estimates (e.g., individuals living in census blocks with no housing units). This has called into question the usability of data at lower levels of geography, such as counties and census blocks, as well as for key sub-populations of interest (see, for example, Hauer and Santos-Lozada 2022; Mueller and Santos-Lozada 2022; Winkler et al. 2020). Research has found that the quality of the data post-DAS processing as it pertains to accurately producing distributions of racial-ethnic groups within and across lower levels of geography is less than satisfactory (Asquith et al. 2022; Kenny et al. 2021; Mueller and Santos-Lozada 2022; Santos-Lozada al. 2020). With such severe challenges to the implementation of the decennial count on the front end coupled with the injection of noise into the data per the DAS on the back end, demographers have reservations about the utility of the 2020 Census to accurately characterize block-level population statics and dynamics.

Enumerating the Population of California in 2020

As the largest state in terms of population size, and the third largest in land area, California has historically posed challenges for accurate enumeration during the decennial census. In the three previous censuses, the net undercount in California was 1.8% in 1990 (McGhee et al. 2018), 1.5% in 2000 (Ericksen 2001), and 0.26% in 2010 (McGhee et al. 2018). The overall undercount has thus improved in recent censuses. Nevertheless, scholars expressed concern before the census that an undercount in 2020 might be more similar to the 1990 Census as over 70% of California's population now belong to sociodemographic subgroups, such as renters, children, young men, and racial/ethnic minorities, that have traditionally been undercounted in the state (McGhee et al. 2018). Additionally, 27% of California's population is now foreign-born (Johnson et al. 2021), and 50% of children in California have immigrant parents (Population Reference Bureau 2021). These are also two traditionally hard-to-count populations that are furthermore likely to be negatively affected by the Trump administration's attempt to collect data on citizenship status.

Enumeration challenges associated with the demographic composition of the state's population were compounded by a housing shortage. The housing shortage has led an estimated 45% of California residents to "double-up" within single-family units (Anderson 2020) and to a rise in the number of accessory dwelling units that often lack unique addresses, such as basement or garage apartments, "granny flats," and RVs parked semi-permanently at a single address (Chapple et al. 2020). These non-traditional living arrangements pose unique problems for enumeration during a census, as the person filling out the census form on behalf of the housing unit may be confused regarding who to exclude or to include as residents. Underreporting in these situations is likely to be common for those concerned that their "doubled-up" living arrangements or accessory dwelling units would signal building code violations, or that

providing safe harbor for undocumented immigrants could put them in legal jeopardy. Furthermore, accessory dwelling units may be skipped entirely as census forms are only sent to those housing units for which the U.S. Census Bureau has a verifiable address.

Implications of these challenges for the state were clear: An undercount would cost California tens of billions of dollars for support from federal programs like Medicare and the Free and Reduced Lunch Program for economically-disadvantaged students (Wallace et al. 2020) and potentially the loss of a seat in the U.S. House of Representatives (McGhee et al. 2018). Aware of these possible scenarios, state leaders took proactive steps to combat a potential undercount by spending an unprecedented $187 million in community outreach and marketing to encourage participation.

Initial tallies of the 2020 Census show that the total population of California registered at 39,538,223 people, reflecting a 6.1% growth rate since the 2010 Census (U.S. Census Bureau 2021). This rate was lower than the country as a whole, which grew by 7.4% across the decade (U.S. Census Bureau 2021). When considering year-to-year intercensal trends, the state logged a net *loss* of 182,083 people between 2019 and 2020 – its first ever reduction since 1900 when the state began annually documenting its population size (Christopher 2021). As a result of this attenuated growth, the state lost one seat in the U.S. House of Representatives (Christopher 2021).

It is unclear if this attenuated growth reflects a real shift in the state's population, an undercount in the 2020 Census, or a combination of the two. Initial evaluations of the 2020 Census by the American Statistical Association found that "Despite concerns that census numbers could be jeopardized by political interference, the task force found no evidence of anything other than an independent and professional enumeration process by the U.S. Census

Bureau" (American Statistical Association 2021). However, they warned that existing data could not ascertain the accuracy of the enumeration, particularly at lower levels of geography. Of relevance to our present study, California was classified in this evaluation as a state with a "very high risk of error" based on an array of indicators of 2020 Census data quality including a high rate of duplicate questionnaires from households that required rectification by the U.S. Census Bureau, a high rate of proxy responses (e.g., a neighbor), and a high rate of imputed occupied households (Biemer et al. 2021).

Standard Approaches for Assessing the Accuracy of the Census

Researchers typically rely on two sources of data to assess the accuracy of the decennial census: estimates produced by Demographic Analysis (DA) and the U.S. Census Bureau's Post-Enumeration Survey (PES). DA makes use of administrative records on births and deaths (to ascertain natural increase) and estimates of migration from survey data (to ascertain net migration) to project the expected size of the population.[1] This projected value is then compared with the enumerated population total from the census. If the projected value from DA exceeds the enumerated population total from the census, the census is considered to have an undercount. Unlike DA, which relies on existing administrative data, the PES requires additional data collection. The PES is essentially a form of "test-retest reliability" where a sample of households who completed their census forms are invited to complete an identical form a few months after

[1] DA makes use of the demographic balancing equation where: $N_t = N_0 + BIRTHS_{0,T} - DEATHS_{0,T} + IN\text{-}MIGRANTS_{0,T} - OUT\text{-}MIGRANTS_{0,T.}$ In this equation, the size of the population (N) at time t is determined by the initial size of the population at time 0, adding in all the births that occurred during the interval 0 to t, subtracting all of the deaths that occurred during the interval, adding in the number of individuals who migrated into the population during the interval, and subtracting the number of individuals who migrated out of the population during the interval. Detailed information on the U.S. Census Bureau's DA methodology is available in Jensen et al. (2020).

the census is taken. Responses from the household's actual census form are then compared with the household's responses to the PES to ascertain whether there was an undercount or overcount.

Initial findings from the U.S. Census Bureau's DA and PES suggest that, despite all the challenges to undertaking an accurate enumeration of the population in 2020, data from the decennial count appear to be reliable on average. At the national level, both DA and the PES indicate that there was no significant undercount or overcount of the total U.S. population in 2020 (Hill et al. 2022; Khubba et al. 2022). The PES identified a slight *overcount* in California (0.47%), however this was not significantly different from zero (Hill et al. 2022). Taken together, these two sources suggest that 2020 Census data are largely sufficient for characterizing the population of California.

Though informative, both DA and the PES have limitations as tools for evaluating the accuracy of the census. DA uses data from birth and death certificates to measure natural increase, and survey estimates on net-migration to determine the population size at the national level. Generally speaking, data from birth and death certificates in the United States are relatively complete and accurate for population-level analysis. However, as migration totals are based on survey data, they are subject to measurement error. While useful in determining the extent of an undercount at the national level, the U.S. Census Bureau does not produce DA at the state level. Moreover, DA is nearly impossible for small area estimates such as census blocks because migration data is limited or, in most cases, non-existent at more fine-grained levels of geography.

In contrast to DA, the PES is based on data collected from individuals willing to participate in the decennial census and an additional survey. These respondents tend to be a select group of households who are not representative of the full population. By design, the PES excludes those who refuse to participate in the Census, and so it has limited value in determining

the full extent of an undercount. The PES is usually undertaken in the months immediately following the census. Due to the COVID-19 pandemic, however, most fieldwork for the 2020 Census' PES took place between November 2021 and March 2022. For many participating households, this follow-up survey was nearly *two years* after they had completed their original census form. With a considerable amount of time having elapsed between the actual census and the PES, the chance of recall error or the experience of a demographic event increases. Finally, the PES is based on a sample and cannot be used to evaluate the accuracy of estimates for small areas like census blocks.

Research Questions

In this study, we improve upon the PES and DA and contribute to the growing evidence base on the quality of the 2020 Census by addressing the following two research questions:

1) To what extent do population totals from the 2020 Census diverge from totals produced by an independent enumeration and from totals produced by demographic projections?
2) What characteristics of census blocks are associated with divergent population totals when comparing the 2020 Census with an independent enumeration and demographic projections?

To answer these questions, we will compare official 2020 Census population totals from a sample of census blocks in California with population totals from an independent enumeration we conducted and with population totals based on advanced demographic projection methods. We conducted our independent enumeration at the same time as the 2020 Census; thus unlike the PES, it is not subject to recall error. Moreover, participation in our independent enumeration was not contingent upon participating in the 2020 Census, and therefore our data are less affected by

selection effects than the PES. Our demographic projections improve upon the standard DA technique by more explicitly accounting for population trends across the intercensal period measured across an array of data sources, including changes to the number of housing units located on each census block. With our independent enumeration and advanced projection methods, we are better positioned to assess the quality of the 2020 Census at the block level than the PES or DA.

In answering the first question, we will assess the stability of estimates from the 2020 Census at the block level and for key demographic groups across our sample blocks. In answering the second question, we will identify features of census blocks that may have contributed to discrepancies in population totals. At its core, the census is an extensive survey operation that requires a direct physical accounting of all housing units and residents. Blocks vary considerably in size, safety, ease of entry and navigation, capacity for new construction, and broadband access. This creates a wide variety of challenges for census field staff tasked with identifying and confirming the presence of housing units requiring enumeration and performing in-person follow-up visits to non-responders. Similar challenges exist for postal carriers who deliver and return census forms. When analyzing census data at the national or state level, these challenges tend to "average out in the wash." Still, they can introduce substantial error when estimating population totals at lower levels of geography. In identifying which features of census blocks are most strongly correlated with discrepant population totals, our analysis will provide context for researchers and policymakers evaluating small area estimates from the 2020 Census as well as inform efforts to enumerate hard-to-count neighborhoods in future censuses.

METHODS AND MATERIALS

Sample

To answer our two research questions, we analyze data collected from a sample of 173 census blocks in California. Census blocks are the lowest level of geography for which the U.S. Census Bureau provides publicly-available population totals. Census blocks are delineated by visible features, such as streets, roads, streams and other bodies of water, railroad tracks, and by nonvisible boundaries, such as selected property lines and city, township, school district, and county limits. In suburban and rural areas, census blocks may be large, irregular, and bounded by a variety of features, and in remote areas, census blocks can cover hundreds of square miles.

To ensure geographic diversity, we sought a sample that would capture a high level of environmental variation in terms of land cover and climate. To ensure demographic diversity, we sought a sample that would facilitate the construction of representative populations based on age, race, and ethnicity. To meet these dual goals, we designed our sample with two complementary subsamples: a geographic and a demographic subsample. The common starting point for both subsamples is a division of the state into seven regions: the North Coast, the Northern Interior, the Eastern Sierra, the Central Coast, the San Joaquin Valley, Southern California, and the Inland Empire/South Desert.

To draw the geographic subsample, we created strata by interacting the seven regions with 15 consolidated land cover classifications defined by the European Space Agency (e.g., grassland, shrubland, urban area, etc.). Theoretically, this could produce up to 105 potential strata. However, we identified only 57 unique strata because not all land use types were present in each region. To draw the demographic subsample, we first identified five major cities to include alongside the original seven regions to ensure that we would be sampling census blocks from the state's major population centers. These cities included Los Angeles, Sacramento, San

Diego, San Jose, and San Francisco. We used city boundaries to distinguish these cities from the rest of the region within which the city is located. Adding these five cities to the original seven regions resulted in 13 mutually exclusive and exhaustive areas spanning the entirety of the state. We stratified each of these 13 areas using tertiles derived from the California Hard-to-Count Index, a summary index based on an array of demographic, housing, and socioeconomic variables correlated with the difficulty of enumerating different census tracts based on response rates from previous censuses.[2] This process resulted in 36 strata.

In combining the 57 strata from the geographic subsample with the 36 strata from the demographic subsample, we had a total of 93 potential strata from which to sample census blocks. To maximize the number of total census blocks in population areas while also including blocks with every possible land cover type, we identified and excluded 22 strata from the geographic subsample with land cover already represented in the demographic subsample. This resulted in a total of 71 strata to draw the sample. We then randomly selected two census blocks within each stratum. The block probability of selection was proportional to the block's share of the state's total housing units in the 2019 public extract of the 2019 Census Bureau Master Address File. Because we were most concerned with evaluating the hardest-to-count blocks, we sampled four census blocks (instead of two) in the demographic subsample strata defined by the tertile with the highest scores on the California Hard-to-Count Index.

Our final analytic sample consists of 173 census blocks: 70 blocks in the geographic subsample and 103 blocks in the demographic subsample. For this analysis, we combine the geographic and demographic subsamples and analyze them together. For each block, we

[2] The California Hard-to-Count Index was developed by the California Department of Finance to identify parts of the state that would require targeting for community outreach and marketing to encourage participation in the census. More information on the construction of this index can be found at https://census.ca.gov/california-htc/

determine population totals through three sources: the 2020 Census, the California Housing and Population Sample Enumeration, and demographic projections generated from a series of demographic and geographic data sets. We describe each of these sources in turn.

Data Sources

2020 Census

The conduct of the 2020 Census is well-documented by the U.S. Census Bureau, with full details of its operations and methodology made publicly available on the Bureau's website. Of particular relevance for our analysis is the Public Law (P.L.) 94-171 Redistricting Data File that the Census Bureau provides each state's governor to guide the redrawing of districts for the U.S. Congress and state legislatures. This file, delivered to each state approximately one year after the census is taken, shows the population totals by age, race, and ethnicity for all residents on each census block in each state. From this file, we extracted population totals for our 173 sample blocks. These population totals have been subjected to the new DAS procedures, and thus contain an undiscernible degree of artificial error.

California Housing and Population Sample Enumeration

The California Housing and Population Sample Enumeration (CHPSE) is the first ever independent enumeration of a population with the purpose of validating official totals from a decennial census in the United States.[3] Sponsored by the California Complete Count Committee – Census 2020 Office and the California Department of Finance, and undertaken by the RAND

[3] The study is officially named the "California Housing and Population Sample Enumeration." However, "California Neighborhoods Count" was the public-facing name used when communicating with sample members.

Corporation and the California Center for Population Research at UCLA, CHPSE was designed to emulate the methods used by the U.S. Census Bureau as closely as possible. CHPSE had two phases of data collection: an address canvass phase and an enumeration phase. During the address canvass phase, undertaken from January through March of 2020, a team of trained interviewers physically went door-to-door around each sample block to verify the street address and the number of separate housing units at each address. This was crucial in identifying accessory dwelling units often missed when relying on lists of addresses.

In previous decennial censuses, the U.S. Census Bureau undertook a complete in-person physical address canvass (similar to the one undertaken in CHPSE) to establish their address frame in advance of the enumeration of the population. However, for cost-saving purposes, the Bureau deployed a two-tier approach to their address listing for the 2020 Census: 25% of addresses would be subjected to an in-person physical address canvass, as done in previous decennial censuses, and 75% of addresses would be verified "in-office" using geospatial imaging software. A limitation of relying on geospatial imaging software is that it cannot determine multi-family units that share the same address, and it is less capable of appropriately identifying accessory dwelling units. Such omissions in the address frame are known contributors to an undercount. CHPSE, however, included a 100% complete in-person physical address canvassing operation with specific protocols for the field staff to inquire directly about multi-family and accessory dwelling units when interviewing residents.

Across the 173 sample blocks, the address canvass identified 23,913 unique housing units, which is 1,245 more than the U.S. Census Bureau identified in their canvassing efforts. Additionally, the address canvas identified 16 group quarters. During the enumeration phase, all addresses identified in the address canvass phase were sent a form that collected a household

roster identical to that used on the official census form. Residents had the option of filling out the paper form or completing the form online. Non-responders were contacted by telephone and in-person interviewers who visited each block. While the U.S. Census Bureau makes up to six attempts at interviewing households during their non-response follow-up phase, CHPSE made more than six attempts, including up to 11 telephone calls and up to eight in-person visits. Data collection for the enumeration phase yielded a 54.0% response rate of the 23,929 sample addresses. The remaining 46.0% of cases were filled using administrative record allocation and imputation.

We imputed population totals for non-responding households using three data sources that could be linked directly with housing units: (1) real estate tax determinations made by the state, which includes information on the housing unit, such as the number of bedrooms and its square footage; (2) California voter registration data, which includes demographic characteristics of household residents registered to vote; and (3) eligibility data for Medi-Cal, which includes demographic characteristics of household residents participating in the state's health insurance program. With these data sources, along with the data collected on the demographic composition of participating households on the same blocks as non-responders, we applied random tree imputation. This strategy is based on machine learning algorithms that can accommodate non-linearities, interactions, and outliers. We take a staged approach to imputation. First, we impute the total number of residents for each non-responding household. Second, if the imputed number of residents is greater than zero, we impute race/ethnicity and age for the "primary respondent" (i.e., the individual who would have responded to the survey if the household had responded). Finally, if the imputed number of residents is greater than one, we impute the demographic characteristics of all other individuals conditional on the race/ethnicity and age of the primary

respondent. To improve the quality of our race/ethnicity imputations, we use the Bayesian Improved Surname Geocoding tool (Elliott et al. 2009). The tool formalizes the observation that knowing a person's name along with their neighborhood's racial/ethnic composition provides indirect information about household residents' self-identified race/ethnicity.

We considered a wide range of imputation models, including hot deck, k-nearest neighbors, and multiple imputation via chained equations. The conditional models we considered included classification and regression trees and random forests. In combination with the imputation model itself, we considered several design choices that are applicable to many of these methods, including whether imputing missing values in one variable depends on all other observed variables or only a subset; whether all variables are used to impute the total number of residents in the housing unit or only those sufficiently correlated and observed with low rates of missingness; and whether indicators of missingness are allowed in the imputation model.

These options produce 83 distinct imputation approaches that we tested. The various approaches produce substantial differences in the total population estimates. We focus on the configurations that produce population estimates close to the 2010 population totals after applying a statewide inflation factor from 2010 to 2020. Among the top ten algorithms that produce totals that track with the 2010 population totals, we compare the distribution of housing unit counts for observed versus imputed counts. After this process, we selected random forests with tree depth one as our preferred estimate. Given the variability in estimates produced by the different imputation approaches, we augment our preferred estimate by also showing population tabulations produced using classification and regression trees or "CART" (our lower range estimate) and using random forests (our higher range estimate). More detailed information on

these imputation approaches and the data collection procedures used in CHPSE are available in BLINDED FOR REVIEW (2022).

Demographic Projections

Projected population estimates as of April 1, 2020 for the state of California at the block level are derived from an array of data sources, including the 2010 Census Summary File 1 (SF1); the 2020 Census Apportionment Results; U.S. Census partnership programs supporting the 2020 Census; 2020 Tiger/Line boundary files; the American Community Survey (ACS) summary files and public use microdata; ESRI Updated Demographics vintage 2020, and population and housing estimates from the California Department of Finance and the U.S. Census Population Division. The 2010 decennial census block file SF1 data serves as a baseline for average population per housing unit and for starting counts of housing units and population as enumerated in the 2010 Census. We measured block group trends observed across the intercensal period in the ACS and tested household size differences for statistical significance. We adjusted constituent blocks to use the most recent data when significant differences existed in a block group over time.

We used Vintage 2019 Address Count operational data containing housing units per block to estimate initial new population per block using persons per housing unit from the 2010 Census or adjustments made via the ACS in the previous step. We subtract the latest estimates of population in group quarters from total population estimates to generate household population estimates (e.g., from the statewide 2020 apportionment total and residential population total). We add geographic identifiers for summary levels 157 (incorporated place within the county) and 795 (public use microdata area or PUMA) to census blocks to enable reweighting to independent

estimates. We used adjustments to ensure consistency between unit and population counts when possible (e.g., blocks with one or more housing units in 2019 which had no population in 2010 were imputed with average household size and demographics from within the parent block group or tract; blocks with one or more persons in 2010 but no housing units or group quarters in 2019 were set to zero population; etc.).

The 2020 Census Apportionment Results contains the total California state resident population as enumerated in the 2020 Census: 39,538,223 people. This estimate was used for poststratification weighting to control county estimates to sum to the total state population according to the living arrangements (in households or group quarters). Block group quarters population estimates were controlled to county group quarters population estimates without intermediate steps. Household populations were controlled top-down from the adjusted county totals to place within counties and tracts. Counties with two or more PUMAs had an additional intermediate control step so that the adjusted county total was imposed on the PUMA totals and then tract totals. We controlled extrapolated block population counts from the vintage 2019 address county listing data to adjust block group and tract totals, as well as place totals. We also controlled the ACS population by age, race, and ethnicity at the state level to match the total state population and the ACS population by age, race, and ethnicity by county (from 5-year summary file tables) to the adjusted county total estimates. We adjusted the race distribution for Hispanic/Latino/Spanish origin using proportions from the 2010 decennial census due to differences in methodology and results between the decennial census and the ACS. Below the county level, we controlled population by age, race, and ethnicity top-down in the same way as total household populations. We converted population by age, race, and ethnicity into

proportions of the total population and applied these to the previously calculated adjusted block total population.

In a final series of steps, we first calculate the percentage of built-up area in each 2010 block within each intersecting 2020 census block to convert 2010 census block geographies to 2020 boundaries. This step proportionally allocates population totals from each 2010 block into all related 2020 blocks per block boundaries defined in the 2020 Tiger/Line boundary files.[4] Then we calculate marginal totals and round numbers to the nearest integers. We perform resampling to ensure that the rounded totals are summed to the marginal totals stored before rounding.

Empirical Strategy

To determine the extent to which population totals from the 2020 Census diverge from totals produced by our independent enumeration and our demographic projections, we tally and compare population totals for our full sample and subpopulations defined by age, race, and ethnicity. Additionally, we calculate net coverage ratios as follows:

$\left(\frac{Census\,Estimate - Independent\,Estimate}{Census\,Estimate}\right) * 100$. We calculate and compare two sets of ratios: (1) ratios where CHPSE provides the independent estimate, and (2) ratios where our demographic projections provide the independent estimate.

To determine structural characteristics of census blocks associated with divergent population totals when comparing the 2020 Census with CHPSE and demographic projections, we estimate an ordinary least squares regression (OLS) model predicting variation in population

[4] This geographic allocation dataset was provided by Amos, Brian, 2021, "2020 Census Block Crosswalk Data", https://doi.org/10.7910/DVN/T9VMJO, Harvard Dataverse, V2.

totals across the 2020 Census, CHPSE, and our demographic projections. We measure this variation as the block-level standard deviation of the three estimates of the population total. Lower values of this outcome measure indicate greater consistency across the three sources, while greater values indicate greater discrepancy across the three sources. This variation is estimated as a function of six block-level factors which may pose challenges to an accurate enumeration: the size of the census block, number of housing units on the census block, broadband access on the block, urbanicity of the block, the presence of hard-to-count structures on the block, and overall difficulty of accessing the block.

Size of the census block is a continuous measure taken from the California Public Utilities Commission and is expressed in square miles. Number of housing units is a continuous variable taken from U.S. Census Bureau's official 2020 block-level totals. Broadband access is a binary variable taken from the California Public Utilities Commission indicating whether the block is wired for broadband. We represent urbanicity by a set of binary variables indicating whether the block is in an urban, suburban, or rural area, as observed directly by CHPSE data collection staff. The presence of hard-to-count structures is a binary variable indicating whether the block had any gated communities, group quarters, or high-rise apartment buildings, as observed directly by CHPSE data collection staff. Lastly, difficulty of accessing the block is a continuous variable taken from direct observations of CHPSE data collection staff who rated each block on a scale of 1 to 5, where '1' indicates the block was easy to access and '5' indicates the block was difficult to access. We report descriptive statistics for these measures in Table 1.

[Table 1 about here]

Instead of traditional *t*-tests to assess the statistical significance of the parameter estimates in this regression model, we apply Monte Carlo permutation tests. Researchers often use these

tests for parametric inference from small, nonprobability samples such as ours (Good 2013). Following conventional standards, we base our analyses on 10,000 permutations per model (Good 2013). *P*-values produced from these permutations indicate the probability of obtaining a result at least as extreme as the test statistic given that the null hypothesis is true.

RESULTS

Our first analytical task is to determine the extent to which population totals from the 2020 Census diverge from totals produced by an independent enumeration and from totals produced by demographic projections. To do so, we first show comparisons at the block level. In Figure 1, we plot 2020 Census population totals against our CHPSE population totals. In Figure 2, we plot 2020 Census population totals against our projected population totals. We fit a regression line to these bivariate distributions and report Pearson correlation coefficients.

[Figure 1 & Figure 2 about here]

Both figures show considerable alignment in the estimates, as evidenced by a nearly 45-degree regression line and strong correlations (r = .95 in both figures). However, there appears to be greater consistency in census blocks that have smaller populations. When the census block population is less than 500 residents, the plots are more closely clustered near the fitted line. There is greater variability in the estimated population totals when the census block population is greater than 500 residents. This provides suggestive evidence that smaller blocks may produce more accurate totals and/or be less affected by the application of the DAS procedures than larger blocks – an issue that we further explore in our multivariate analysis.

It is worth pointing out that both figures contain a noticeable outlier in which the 2020 Census total far exceeds the independently estimated total. The largest outlier when comparing

the 2020 Census population totals with the CHPSE population totals in Figure 1 is for a block in Palo Alto where the 2020 Census counted 1,246 individuals while CHPSE only counted 263 individuals. On this particular block, the U.S. Census Bureau address canvass identified 552 housing units, while our CHPSE address canvas identified only 124 housing units. Given this large discrepancy, the CHPSE field staff re-canvassed this block to validate the total. While we cannot unequivocally ascertain the reason for this discrepancy, we speculate the U.S. Census Bureau may have made an error. The error was either in their documentation of total housing units or in their delineation of block boundaries, as they relied on "in-office" verification using geospatial imaging software for 75% of blocks instead of the traditional in-person canvass.[5] The largest outlier when comparing the 2020 Census population totals with the projected totals in Figure 2 is for a block in the Bay Area, where the 2020 Census counted 1,314 individuals while we projected only 69 individuals. On this particular block, the U.S. Census Bureau address canvass identified 569 housing units while we only projected 25 housing units. When we remove these two outliers, the correlation in Figure 1 improves to .97, and the correlation in Figure 2 improves to .98.

Next, we tally and compare population totals for our full sample and for subpopulations defined by age, race, and ethnicity. These tabulations are shown in Table 2 using the categories provided in the Public Law 94-171 Redistricting Data File. The 2020 Census counted 53,295 individuals living in our 173 sample blocks, which is 1,483 more than counted by CHPSE (translating to a 2.8% overcount) and 32 less than we projected (translating to a 0.1% undercount). Though we focus on our preferred CHPSE estimate, it is important to note that our

[5] We are unable to determine if the block in question was subject to in-person or in-office validation on the part of the U.S. Census Bureau.

various imputation strategies yield considerable variation. Our low CHPSE estimate was 6,298 *fewer* individuals than the 2020 Census, and our high CHPSE estimate was 5,210 *more* individuals than the 2020 Census. In sum, the different sources yield non-negligible variation, but the official 2020 Census enumeration is near the center of these ranges. Taken together alongside the plots shown in Figure 1 and Figure 2, our analysis suggests that the total population counts provided by the U.S. Census Bureau at the block level for the 2020 Census are not biased in any notable direction.

[Table 2 about here]

The situation changes when moving from the total population to key subpopulations. For ease of interpretation, we shift our focus from raw totals to the net coverage ratios presented in Table 3. Net coverage ratios are computed as 100*(census-CHPSE total)/CHPSE total or 100*(census-projected total)/projected total. Positive ratios indicate the percentage by which each subpopulation is undercounted, while negative ratios indicate the percentage by which each subpopulation is overcounted – assuming the independent estimate (CHPSE total or projected total) used in the calculation is correct. While the ratios are close to zero for the total population, the ratios are larger for subpopulations defined by age, race, and ethnicity. In some cases, they are considerably larger.

[Table 3 about here]

In terms of age, demographers are typically concerned with the undercount of the young. However, our independent sources suggest the opposite for our sample blocks in California. Our CHPSE preferred estimate shows a 11.4% overcount for those under 18 years of age, while our projected totals show a much more modest overcount of 1.9%. As with our top-line population totals, the CHPSE low and high estimates evidence a considerable range, with our low estimate

indicating an 18.6% overcount of the youngest ages and our high estimate indicating a 17.6% undercount of the youngest ages. In terms of race/ethnicity, the ratios comparing the 2020 Census with CHPSE suggest that the 2020 Census Bureau considerably undercounted Native Hawaiian and other Pacific Islanders (-731.4%) and considerably overcounted those reporting two or more races (43.5%) and Asians (43.1%). The ratios comparing the 2020 Census with our projections suggest that the 2020 Census Bureau considerably overcounted those reporting two or more races (58.2%) or some other race (22.2%) while considerably undercounting Whites (-30.5%). Given that these ratios are large and produce inconsistent patterns across the different data sources, we have less faith in the reliability of the 2020 Census subpopulation totals by age and race and ethnicity at the block level.

[Table 3 about here]

Our second analytical task is to identify characteristics of census blocks associated with divergent population totals when comparing the 2020 Census with an independent enumeration and with demographic projections. To do this, we estimate an OLS regression predicting the average deviation in population totals across the 2020 Census, CHPSE, and our projections as a function of block-level characteristics that may have posed challenges to an accurate enumeration. We present parameter estimates from this model in Table 4. In the model, two variables yield coefficients that meet conventional levels of statistical significance: the size of the block in square meters ($p < .05$) and the number of housing units on the block ($p < .01$). The corresponding coefficients are positive, indicating that larger blocks – both in terms of land area and in terms of the number of housing units – show more variation in their population estimates across our different sources than smaller blocks. This complements the findings in Figures 1 and 2, which show less consistent population totals for larger overall populations. Lastly, because the

possibility of making a Type II error is elevated with small sample sizes such as ours, we highlight two coefficients that approach conventional levels of statistical significance ($p < .10$): Blocks without broadband access produce more heterogenous population totals compared to blocks with broadband access, and suburban blocks produce more heterogenous population totals compared to rural blocks. We discuss the implications of these four significant coefficients in our discussion.

[Table 4 about here]

DISCUSSION

The U.S. Census Bureau faced unprecedented obstacles in undertaking its decennial enumeration in 2020. Prior to data collection, the Trump administration attempted to add citizenship questions to the survey form, which would likely suppress participation among immigrant communities. During data collection, which included the first-ever option to participate online with limited pilot testing, the COVID-19 pandemic disrupted and delayed nearly every phase of the operation. Following data collection, new procedures to protect the privacy of individuals infused the data with artificial noise. This constellation of factors elevated concerns about the accuracy of the data – particularly at lower levels of geography such as census blocks. In this study, we assess the accuracy of block-level population totals in the 2020 Census using data from the CHPSE, the first ever independent enumeration survey of census blocks, as well as from demographic projections. We also identify characteristics of census blocks that may lead to unreliable population totals. By doing so, we can provide some background and guidance to social scientists who use block-level data from the 2020 Census for their analyses.

Our study has three key findings. First, we find that in our sample of 173 census blocks in California, population counts provided by the U.S. Census Bureau at the block level for the 2020 Census are not biased in any consistent direction. Our preferred CHPSE estimate indicates that the 2020 Census produced a 2.8% overcount, while our projections indicate that the 2020 Census produced a 0.1% undercount. To put these findings in context, recall that the U.S. Census Bureau's PES detected a 0.5% overcount in California. Further, the correlation between the 2020 Census and CHPSE totals was .95, and the correlation between the 2020 Census and our projected totals was also .95. Given these strong correlations and given that the official 2020 Census total falls within the range of our CHPSE and projected totals, we conclude that block level population total estimates from the 2020 Census are largely reliable.

Second, we find that block level totals for subpopulations defined by race and ethnicity are highly variable, with the largest discrepancies differences observed for Asians, Native Hawaiians, other Pacific Islanders, and those reporting multiple races. One plausible explanation for the discrepancies between the 2020 Census and our projections is that the latter are functionally derived from the 2010 Census and ACS data collected during the 2010 to 2020 intercensal period. Across the decade, the share of babies who are ethno-racially mixed increased (Alba 2020), and alongside the rise of genetic testing services, there is evidence that Americans are increasingly likely to report being of mixed race and ethnicity (Johfre et al. 2021). Relying on data collected earlier in the intercensal period may fail to pick up these changes and allocate mixed-race individuals to single race categories. Regardless of the reason for these differences, our findings cast doubt on the reliability of 2020 Census data to accurately characterize the racial-ethnic composition of census blocks. This corroborates other research which finds unreliable racial-ethnic distributions produced from the 2020 Census at lower levels of

geography (Asquith et al. 2022; Kenny et al. 2021; Mueller and Santos-Lozada 2022; Santos-Lozada al. 2020). Taken together, these findings suggest that researchers studying racial/ethnic population statistics and dynamics at the block level with 2020 Census data should do so with extreme caution.

Third, we find that three structural features of census blocks are associated with discrepant population totals across our different sources: block size (measured in terms of land area or in total housing units), urbanicity, and broadband access. Blocks that are large in land area and housing stock, blocks that are in suburban areas, and blocks that do not have broadband access appear to have the least reliable population totals. This has implications for researchers using block-level data from the 2020 Census and those at the U.S. Census Bureau planning for future censuses. Researchers using block-level population totals as key variables in their analysis, either as a predictor or as an outcome, may want to consider including controls for block size, urbanicity, and broadband access in their models. Controlling for these factors will not eliminate error caused by faulty enumeration or from the application of the DAS, but it will help provide clearer estimates. In planning for future censuses, the U.S. Census Bureau should prioritize large blocks, blocks in suburban areas, and blocks lacking broadband in their address canvass in their outreach activities and non-response follow-up procedures. Additionally, developers of the DAS may want to consider these block-level factors when infusing noise so that data for large blocks, blocks in suburban areas, and blocks lacking broadband are not distorted more than necessary to ensure respondent confidentiality.

There are two notable limitations of our analysis. First, in previous censuses, concerns about data quality centered entirely on the completeness of the data collection operation. However, with the introduction of the DAS, we cannot ascertain if the discrepancies observed in

our study are due to data collection failures or from the artificial noise infused into the data. Therefore, our findings can only speak to the general reliability of block-level totals provided by the U.S. Census Bureau from the 2020 enumeration. Second, CHPSE was, by and large, a successful effort by contemporary survey research standards, but still lagged the 2020 Census in terms of participation. In the 2020 Census, data were directly collected from 76.1% of households, with administrative record allocation and imputation needed to estimate the population for the remaining 23.9%. In CHPSE, data were directly collected from 54.0% of households with administrative record allocation and imputation needed to estimate the population for the remaining 46.0%. However, the findings we report from CHPSE, with its higher administrative record allocation and imputation rate, are instructive for future censuses. This is especially true as initial planning for the 2030 Census indicates a move toward increasing reliance on administrative data instead of direct surveys administered to households (MITRE Corporation 2016).

In closing, if we liken demography to photography, censuses can be considered a point-in-time snapshot of the population. If two photographers were to take a picture of the same scene at the exact same time or if the same photographer shot the same scene twice, but minutes apart, the resulting snapshots would not be identical. This becomes even further complicated when the scene in question is of a population – a fluid construct that changes size and shape by the seconds. In the case of the 2020 Census, the DAS can be thought of the same way as post-hoc applying a filter to a photograph: It maintains the overall composition of the scenery but artificially obscures and amplifies different details. Our study, with the first-of-its-kind independent enumeration survey, provides a rare opportunity to compare multiple snapshots of the same population with different "cameras" and "filters." While we do not expect the results to

be identical, they provide different angles from which to make comparisons to construct an overall evaluation of the scene – i.e., the population living in a census block. In evaluating these various population snapshots, we highlight relative consistency at the population level with substantial inconsistencies at the sub-population level. Further, we observe systematic patterning in these inconsistencies across different structural dimensions of census blocks. Based on our assessment of these multiple snapshots, we urge demographers and other social scientists to proceed with extreme caution when using block-level subpopulation totals from the 2020 Census. Finally, we urge the U.S. Census Bureau to consider the block-level features we identified as correlated with inconsistent estimates when planning for the 2030 Census.

REFERENCES

Alba, R. (2020). *The Great Demographic Illusion: Majority, Minority, and the Expanding American Mainstream*. Princeton, NJ: Princeton University Press.

American Statistical Association. (2021). 2020 Census state population totals:
A report from the American Statistical Association task force on 2020 Census quality indicators. Washington, DC: American Statistical Association.

Anderson, M. (2020, November 18). Cities whose residents are most likely to live with roommates. *Porch*. Retrieved from https://porch.com/advice/cities-whose-residents-likely-live-roommates

Asquith, B., Hershbein, B., Kugler, T., Reed, S., Ruggles, S., Schroeder, J., Yesiltepe, S., & Van Riper, D. (2022). Assessing the impact of differential privacy on measures of population and racial residential segregation. *Harvard Data Science Review*, (Special Issue 2).

Biemer, P., Salvo, J., & Auerbach, J. (2021). The quality of the 2020 Census. An independent assessment of Census Bureau activities critical to data quality. Working paper. arXiv:2110.02135

Brown, J.D., Henneness, M.L., Dorinski, S.M., Warren, L., & Yi., M. (2018). *Understanding the Quality of the Alternative Citizenship Data Sources of the 2020 Census*. Washington, DC: U.S. Census Bureau, Center for Economic Studies.

Chapple, K., Lieberworth, A., Ganetsos, D., Valchuis, E., Kwang, A., & Schten, R. (2020). *Accessory Dwelling Units in California: A Revolution in Progress*. Berkeley, CA: Center for Community Innovation.

Christopher, B. (2021, May 7). California's population shrank in 2020, but don't call it an exodus. *Cal Matters*. Retrieved from https://calmatters.org/politics/2021/05/california-population-shrink-exodus/

Cohn, D. (2020, July 28). Four-in-ten who haven't yet filled out U.S. census say they wouldn't answer the door for a census worker. Pew Research Center. Retrieved from https://pewrsr.ch/39GX30L

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., & Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, *9*(2), 69.

Ericksen, E. (2001). *An evaluation of the 2000 Census. Final report to Congress*. Suitland, MD: U.S. Census Bureau.

Good, P. (2013). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York, NY: Springer.

Hill, C., Heim, K., Hong, J., & Phan, N. (2022). *U.S. Census Bureau, 2020 Post-Enumeration Survey Estimation Report, PES20-G-02RV, Census Coverage Estimates for People in the United States by State and Census Operations*. Washington, DC: U.S. Government Publishing Office.

Hauer, M.E. & Santos-Lozada, A.R. (2021). Differential privacy in the 2020 census will distort COVID-19 rates. *Socius*, 7.

Howe, A. (2019, July 11). Trump administration ends effort to include citizenship question on 2020 census. *SCOTUSblog*. Retrieved from https://www.scotusblog.com/2019/07/trump-administration-ends-effort-to-include-citizenship-question-on-2020-census/

Jensen, E.B., Knapp, A., King, H., Armstrong, D., Johnson, S.L., Sink, L., & Miller, E. (2020). Methodology for the 2020 Demographic Analysis Estimates. Suitland, MD: U.S. Census Bureau.

Johfre, S.S., Saperstein, A., & Hollenbach, J.A. (2021). Measuring race and ancestry in the age of genetic testing. *Demography*, *58*(3), 785–810.

Johnson, H., Perez, C.A., & Mejia, M.C. (2021). *Immigrants in California*. San Francisco, CA: Public Policy Institute of California.

Kenny, C.T., Kuriwaki, S., McCartan, C., Rosenman, E.T.R., Simko, T., & Imai, K. (2021). The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census. *Science Advances*, 7(41), eabk3283.

Khubba, S., Heim, K., & Hong, Y. (2022). *U.S. Census Bureau, 2020 Post-Enumeration Survey Estimation Report, PES20-G-01, National Census Coverage Estimates for People in the United States by Demographic Characteristics.* Washington, DC: U.S. Government Publishing Office.

Lapowsky, I. (2019, February 6). The challenge of America's first online census. *Wired*. Retrieved from https://www.wired.com/story/us-census-2020-goes-digital/

Marimow, A.E., Zapotosky, M., & Bahrampour, T. (2019, July 2). 2020 Census will not include citizenship question, Justice Department confirms. *Washington Post*. Retrieved from https://www.washingtonpost.com/local/social-issues/2020-census-will-not-include-citizenship-question-doj-confirms/2019/07/02/0067be4a-9c44-11e9-9ed4-c9089972ad5a_story.html

McGhee, E., Bohn, S., & Thorman, T. (2018). *The 2020 Census and Political Representation in California*. San Francisco, CA: Public Policy Institute of California.

MITRE Corporation (2016). *Alternative Futures for the Conduct of the 2030 Census*. McClean, VA: MITRE Corporation.

Mueller, J. T. & Santos-Lozada, A. R. (2022). The 2020 US Census differential privacy method introduces disproportionate discrepancies for rural and non-white populations. *Population Research and Policy Review*.

Pew Research Center. (2020). Americans' views of government: Low trust, but some positive performance ratings. Washington, DC: Pew Research Center,

Population Reference Bureau. (2021). Children living with foreign-born parents. Washington, DC: Population Reference Bureau. Retrieved from https://www.kidsdata.org/topic/573/foreign-born-parents/table#fmt=786&loc=2,1&tf=108&sortColumnId=0&sortType=asc

Roach, D., & Carey, J. (2020). *Biodemography: An Introduction to Concepts and Methods.* Princeton, NJ: Princeton University Press.

Ruggles, S., Fitch, C., & Schroeder, J. (2019). Differential privacy and census data: Implications for social and economic research. *AEA Papers and Proceedings*, *109*, 403-408.

Santos-Lozada, A. R., Howard, J. T., & Verdery, A. M. (2020). How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences of the United States of America, 117*(24), 13405-13412.

Sullivan, T. A. (2020). Coming to our census: How social statistics underpin our democracy (and republic). *Harvard Data Science Review,* 2(1).

Supan, J. (2021, May 27). Pandemic moving study: How remote work spurred moves out of big cities. *Allconnect*. Retrieved from https://www.allconnect.com/blog/covid-moving-trends

U.S. Census Bureau. (2021). California remained most populous state but growth slowed last decade. Retrieved from https://www.census.gov/library/stories/state-by-state/california-population-change-between-census-decade.html

Wallace, S.P., Khan, A., & del Pino, H.E. (2020). *Health and Social Service Implications of a Census Undercount in Los Angeles*. Los Angeles, CA: UCLA Center for Health Policy Research.

Winkler, R. L., J. L. Butler, K. J. Curtis, & Egan-Robertson, D. (2022). Differential privacy and the accuracy of county-level net migration estimates. *Population Research and Policy Review,* 41(2), 417-435.

Table 1. Descriptive statistics for sample census blocks

	Mean/proportion	Standard deviation
Average deviation in population totals across the 2020 Census, CHPSE, and demographic projections	35.75	64.26
Size of the block (square meters)	0.13	0.33
Number of housing units on the block	131.03	173.54
Block does not have broad band access	0.09	--
Block is urban	0.55	--
Block is suburban	0.28	--
Block is rural	0.17	--
Block has hard-to-count structures	0.16	--
Difficulty of accessing block	1.16	0.54

N = 173

Table 2. Comparison of population totals from different sources

	2020 Census totals	CHPSE totals (low)	CHPSE totals (preferred)	CHPSE totals (high)	Projected totals
Total Population	53,295	46,997	51,812	58,505	53,327
Age					
Under 18	11,125	9,052	9,856	13,078	10,919
18 and older	42,170	37,945	41,956	45,427	41,101
Race					
White	23,365	24,512	30,896	30,866	30,498
Black or African American	3,097	2,770	2,849	3,631	3,103
American Indian and Alaska Native	842	822	935	1,106	762
Asian	8,206	4,834	4,670	5,681	7,551
Native Hawaiian and Other Pacific Islander	207	1,740	1,721	1,911	248
Some other race	10,565	7,875	6,776	9,415	8,231
Two or more races	7,013	4,444	3,965	5,895	2,934
Hispanic or Latino					
Hispanic or Latino	19,215	17,401	22,236	21,865	19,086
Not Hispanic or Latino	34,080	29,596	29,576	36,640	34,241

Table 3. Net coverage error ratios of 2020 Census based on comparisons with CHPSE totals and projected totals

	2020 Census compared with CHPSE totals (low)	2020 Census compared with CHPSE totals (preferred)	2020 Census compared with CHPSE totals (high)	2020 Census compared with projected totals
Total Population	11.8%	2.8%	-9.8%	-0.1%
Age				
Under 18	18.6%	11.4%	-17.6%	1.9%
18 and older	10.0%	0.5%	-7.7%	2.5%
Race				
White	-4.9%	-32.2%	-32.1%	-30.5%
Black or African American	10.6%	8.0%	-17.2%	-0.2%
American Indian and Alaska Native	2.4%	-11.0%	-31.4%	9.5%
Asian	41.1%	43.1%	30.8%	8.0%
Native Hawaiian and Other Pacific Islander	-740.6%	-731.4%	-823.2%	-19.8%
Some other race	25.5%	35.9%	10.9%	22.1%
Two or more races	36.6%	43.5%	15.9%	58.2%
Hispanic or Latino				
Hispanic or Latino	9.4%	-15.7%	-13.8%	0.7%
Not Hispanic or Latino	13.2%	13.2%	-7.5%	-0.5%

Table 4. Parameter estimates from an OLS regression model predicting the avarage deviation in population totals across the 2020 Census, CHPSE, and demographic projections, with Monte Carlo permutation tests

	Coefficient	Permutation *p*-value
Size of the block (square meters)	38.73*	0.03
Number of housing units on the block	0.20**	0.00
Block does not have broad band access	72.95+	0.09
Block is urban vs. rural	18.00	0.14
Block is suburban vs. rural	27.19+	0.05
Block has hard-to-count structures	-13.83	0.88
Difficulty of accessing block	-10.91	0.95
Constant	-0.49	0.98
$R^2 = 0.32$		
N = 173		

$+p < 0.10$ $*p < 0.05$ $**p < 0.01$

Figure 1. Scatterplot comparing population totals from CHPSE with population totals from the 2020 Census for 173 sampled blocks in California

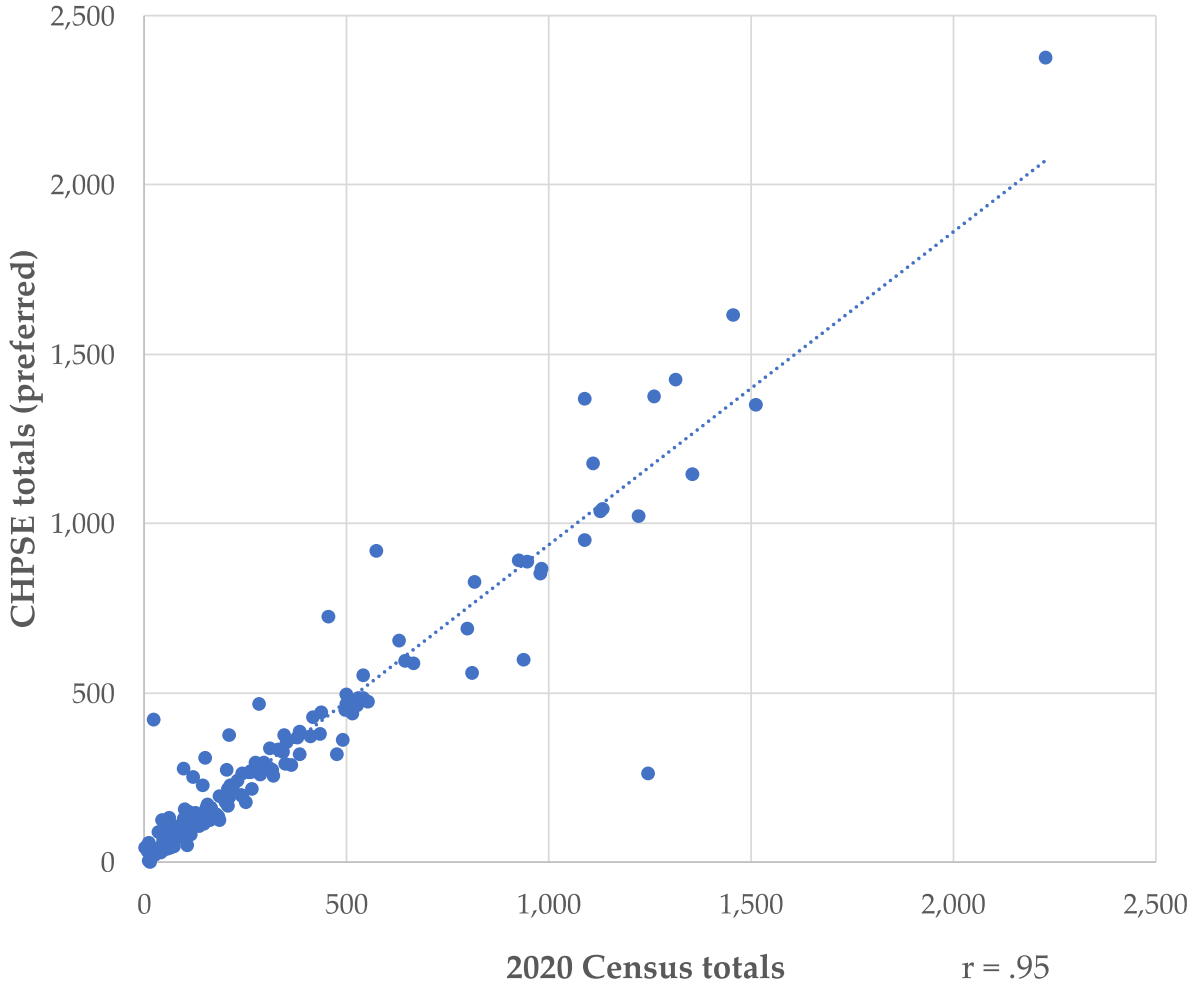

Figure 2. Scatterplot comparing projected population totals with population totals from the 2020 Census for 173 sampled blocks in California