A Cascade Model: How Latino Immigrants' Lowered Response Will Lead to Differential Undercount in Census 2020

SAN JOAQUIN VALLEY CENSUS RESEARCH PROJECT
SAN JOAQUIN VALLEY HEALTH FUND

BY EDWARD KISSAM

SAN JOAQUIN VALLEY HEALTH FUND

THE CENTER at Sierra Health Foundation

CALIFORNIA INSTITUTE FOR RURAL STUDIES

January 2019

Table of Contents

Executive Summary

A Cascade Model: How Latino Immigrants' Lowered Response Will Lead to Differential Undercount in Census 2020

This is the second in a series of six reports by the San Joaquin Valley Census Research Project based on Fall 2018 survey research assessing the likely impact that adding a citizenship question to Census 2020 will have on Latino first- and second-generation immigrant undercount in the region.

The Census Bureau has been consistently optimistic about the viability of streamlined census procedures introduced as part of modernizing and re-engineering Census 2020. However, the new procedures, while cost-effective and probably satisfactory for easier-to-count populations and communities, are likely to have serious limitations when utilized in neighborhoods and communities such as those of the San Joaquin Valley.

Most problematic, the reliability of the planned data collection strategies has not been assessed or tested in the distinctive societal context where addition of the sensitive question about citizenship is expected to lead directly to pronounced response bias within an already hard-to-count population. We expect the re-engineered procedures will exacerbate differential undercount, at least in immigrant communities, and quite possibility in others with high proportions of low-income households.

The cascade model presented here draws both on the research done in the San Joaquin Valley and presented in the first of these reports and on previous analysis and research. It describes how lowered response rates are likely to affect each stage in the census process of data collection/processing.

This San Joaquin Valley Census Research Project-based model explains how lowered response is transformed into undercount. It is referred to as a "cascade" model because the level of success and failure at each stage in decennial census operational procedures determines the parameters for census operations at the next stage. Level of self-response, for example, determines extent of reliance on enumerator efforts to secure information from the households that fail to self-respond. Enumerator success in this endeavor then determines the extent of reliance on proxy interviews for information on households. Cumulative success at this stage, then, determines the

extent of efforts to secure data from administrative records. And finally, cumulative success determines the need for reliance on count and whole-person imputation. The accuracy of the census count (and demographic profile of the population) depends on the level of reliance on each data-collection or analytic operation, since some (e.g. proxy interviews) are known to be more error-prone than others.

The cascade of census stages in decennial census collection/imputation are visualized in the cascade model of undercount of the San Joaquin Valley Latino first- and second-generation immigrants in Figure 1.

Figure 1—The Cascade Model of Diminishing Data Quality in Successive Census Operations

Compilation of Master Address File

Delivering Invitations to Respond (Internet Choice and Internet First) and Securing Self-Response

Non-Response Follow-up: Enumerator Household Interviews

Non-Response Follow-up: Proxy Interviews with Neighbors if no "Direct" Interview is possible

Search for "High Quality" Matching Administrative Records for HH Imputation

Hot Deck Imputation to estimate size and characteristics of HH's not enumerated otherwise

Census Tabulations Reporting Population Count and Demographic Characteristics

The current paper incorporates empirical data collected in the San Joaquin Valley Census Research Project survey to estimate the Census Bureau's success/failure in the following stages of the process: MAF-building, self-response, NRFU direct interview response, and proxy interview response. It also relies on the project's survey data to estimate the systematic undercount resulting from use of hot-deck imputation due to differences in size between the households likely to respond to the census and those likely not to respond.

Despite the valiant but compromised efforts by the Census Bureau to generate accurate census tabulations in the face of greatly elevated and uneven non-response, we believe the result will be attrition in data quality that ultimately results in flawed tabulations of both the size and demographic characteristics of the region's population.

The stakes are high for the San Joaquin Valley because the hard-to-count population of Latino immigrants makes up more than one-third of the entire population in the region. The model estimates the level of Latino immigrant undercount in the region as being 11.7% if the citizenship question is added. Given the size of the Latino immigrant population likely to be undercounted, it is reasonable to expect a 4.1% undercount in the total population of the San Joaquin Valley.

One of the issues it will be particularly important to consider is not just the overall flawed tabulations, but the differential undercount of sub-groups within the Latino population. Different levels of census response among undocumented immigrants, legal residents, naturalized citizens and the U.S.-born second-generation will skew the census-derived demographic profile of Latinos, as well as give rise to geographic disparities in census count. The model also identifies some potential ways for the Census Bureau to collaborate with local stakeholders in combined efforts to ameliorate likely undercount. This paper makes it clear that "Get Out The Count" campaigns focused primarily on impacting respondent motivation will not yield adequate results unless they also incorporate strategies to improve operational processes of census data collection.

It appears that the Census Bureau's view about the efficacy of its procedures to "cure" widespread

non-response stemming from inclusion of the citizenship question in Census 2020 is misplaced. The cascade model in its present (initial) stage is essentially an exercise in hypothesis generation—tracing how patterns of non-response ripple onward through NRFU into flawed tabulations. We cannot yet definitively determine the model's predictive accuracy, in part because details on some aspects of Census 2020 operations (particularly those relating to reliance on administrative records and algorithms for hot-deck imputation) are unclear or unavailable.

We also recognize that the San Joaquin Valley Census Research Project initial findings about the prevalence and structure of complex households need to be further researched, due to the variety of housing accommodation and living arrangements, and the need to better understand how adding a citizenship question would exacerbate pre-existing patterns of partial household undercount in these sorts of crowded housing. The contribution of the project's initial research in this specific area is to highlight issues that have not yet been adequately addressed by the Census Bureau.

The current analysis and estimate of Latino immigrant undercount can and should be refined as Census 2020 operations are finalized. Nonetheless, we think it is critical at this juncture of census planning to think clearly and practically about operational adjustments that might contribute to an accurate and fair census.

Our hope is that the analysis presented here provides a useful framework to re-assess how the re-engineered decennial census operations will affect differential undercount in different regions and among ethnic groups with specific demographic profiles. The model's projection of the likely magnitude of differential undercount in Latino immigrant communities, even if subsequently adjusted, suggests the need for a commitment to carry out the research needed to yield fine-grained measurement of Census 2020 differential undercount and to use ethnographic research and demographic analysis in addition to dual-system estimation. It is unfortunate that the Census Bureau's ethnographic research efforts, which so powerfully illuminated crucial understanding of multiple causes of differential undercount, have languished over the past decade. Such research might well have shown, as the

San Joaquin Valley Census Research Project has sought to do, that the dynamics of census undercount cannot be adequately understood in isolation, that real-world context and operational implementation need to be carefully considered concurrently.

Being a work in progress, the San Joaquin Valley Census Research Project will update the analyses in the model to incorporate forthcoming survey-based findings about patterns of census response among non-Latino immigrants in the San Joaquin Valley when they become available in February 2019.

Meanwhile, we encourage readers who are concerned about the possibility of differential undercount in communities with high concentrations of low-income minority and immigrant households to consider using the cascade model analytic framework in combination with local survey and ethnographic research to examine the distinctive configuration of operational risks they face if Census 2020 includes the citizenship question.

Introduction

An important issue in projecting the impact of the Department of Commerce's efforts to add a question on citizenship to Census 2020, given consensus that adding the question will decrease census response rates among Latino and other immigrants, is to determine the extent to which the problem of lowered response can be overcome in the course of non-response follow-up (NRFU).[1] This report describes a "cascade" model of census undercount developed to demonstrate how dramatically increased levels of non-response among certain populations in certain community contexts would be transformed into differential undercount.

The cascade model of undercount provides a basis to generate a sound empirically based estimate of the regional impacts that a 2020 decennial census with an added citizenship question (CQ) might have on undercount of Latino first- and second- generation immigrants. The report is drawn from interviews with Latino immigrants in the eight counties of the San Joaquin Valley and, based on their responses, shows how widespread differential undercount of this population arises and how it would affect the total Census 2020 accuracy for the region.

The analysis presented here details the development of the cascade model based on survey data from the San Joaquin Valley, on review of previous research on census undercount describing how multiple causes of undercount interact, and on examination of the likely impact of the Census Bureau operational plans for implementing a re-engineered Census 2020 on census enumeration. It shows that, even if there were more funding, and vigorous efforts by the Census Bureau and local stakeholders, the lowered response rates of Latino first- and second-generation immigrants in the San Joaquin Valley will still result in serious differential undercount.

At the same time, by closely examining the extent to which different factors might contribute to undercount, the cascade model provides guidance for strategic efforts to adapt census operations to improve census accuracy in the region—by identifying operational pressure points where collaborative efforts with local stakeholders might, at least, mitigate serious differential undercount.

The model presented here also provides insights for designing an alternative enumeration research initiative, whereby states and other census stakeholders might independently evaluate census coverage of hard-to-count populations in geographic areas with concentrations of immigrants where already-problematic standard census operational procedures may fail due to unprecedented high levels of non-response occasioned by the CQ.

Overview of the Cascade Model and Its Utility

During the non-response follow-up (NRFU) process, the Census Bureau works hard to implement a methodological strategy designed to compensate for household non-response. Nonetheless, census data quality is eroded in communities when there are high levels of non-response among some sub-populations, despite the Census Bureau's best efforts to secure complete enumeration. This is because, when confronted with high levels of household non-response, the Census Bureau is forced to rely on additional operational and statistical procedures—most notably proxy interviews, recourse to administrative records and, finally, imputation—to generate published tabulations of raw census data. Each of these efforts, while partially compensating for non-response, introduces errors into the eventual tabulations of census data that provide the official basis for apportionment and for allocation of federal funding.

The cascade model of undercount described here draws on earlier researchers' powerful insights that patterns of differential undercount do not stem from certain populations being intrinsically hard to count, but rather from the interactions between the census system of enumeration and the population being enumerated, as well as on findings from the San Joaquin Valley Census Research Project of Latino first- and second-generation immigrants to estimate the extent of undercount in this population and the resulting patterns of regional undercount.

The San Joaquin Valley Census Research Project Survey and the Cascade Model: Understanding the Dynamics of Census Undercount

The Census Bureau has been consistently optimistic about the viability of streamlined census procedures introduced as part of modernizing and re-engineering Census 2020. Unfortunately, there are reasons to believe that the new procedures, while cost-effective and satisfactory for easier-to-count populations and communities, have serious limitations when utilized in neighborhoods, communities, counties, regions and states with higher-than-average concentrations of non-citizens.

1 See the January 15, 2019, decision by the District Court, Southern District of New York in New York Immigration Coalition et. al. v. United States Department of Commerce "Findings of Fact and Conclusions of Law."

The specific ways in which the Census Bureau's system of data collection and analysis leads to differential undercount stem in part from the characteristics of the population being enumerated, but also from the structural characteristics of a geographic area—housing patterns and living arrangements and local socioeconomic context.[2] For better or worse, the population and housing characteristics of the San Joaquin Valley make it a natural laboratory for exploring the extent to which adding a sensitive question such as the citizenship question to the decennial census gives rise to differential non-response, which then results in severe differential undercount of, at least, immigrants and, presumably, other socioeconomically defined groups too.

Of particular concern in the San Joaquin Valley and regions with dense concentrations of low-income immigrant households are use of "in office" address canvassing as a substitute for "in field" address canvassing, efforts to rely on administrative records as a source of information on household size and characteristics, and ultimately, reliance on hot-deck imputation when other efforts fail. The re-engineered Census 2020 procedures may sometimes be more cost-effective than old-fashioned operational processes, but the apparent cost-effectiveness of these operational innovations may well undermine census accuracy—especially in geographic areas where non-response is extremely high. This appears to be a likely outcome in California's San Joaquin Valley.

The Department of Commerce's decision to add a citizenship question (CQ) to Census 2020 was clearly bound to result in differential undercount in the San Joaquin Valley. It is a large region with an expected 2020 population of about 4.6 million that is 52% Hispanic and where slightly more than one-third (35%) of the adult Hispanic population of potential census respondents, i.e. "householders" (P1), are first- or second-generation Hispanic immigrants. What was less clear was how to develop a quantitative estimate of the dynamics, whereby elevated non-response among immigrants resulting from adding the question might interact with existing structural barriers and thus undermine efforts to secure a fair and accurate census count.

The current iteration of the cascade model is specific to estimating the undercount of Latino first- and second-generation immigrants in the San Joaquin Valley because it

incorporates survey-based coefficients from the October-November 2018 San Joaquin Valley Census Research Project survey. However, the model framework can be adapted for use in any community or geographic area to generate an estimate of the magnitude of undercount of any identified vulnerable hard-to-count population in the context of any specific geographic area, assuming appropriate survey-based coefficients are available to describe a particular hard-to-count population's propensity to self-respond, respond to enumerators, respond to proxy interviews, size of responding and non-responding households, and likely representation in administrative records.

Litigation Seeking Adjustment for Erroneous Census Tabulations that Arise from Differential Undercount[3]

Differential undercount of minorities has been a longstanding statistical problem in the decennial census and a practical policy problem because the flawed census data resulting from differential undercount leads to misallocation of federal funding that relies on census-derived data. At the same time, it also reduces political equity.

In 1980, several major cities sued the Census Bureau and sought to have census data statistically adjusted in order to correct for what was universally recognized to be not just random errors and omissions, but systematic errors in census enumeration. These systematic errors gave rise to flawed tabulations of the population in geographic areas, cities and states with higher concentrations of minorities and immigrants. There was similar litigation around the 1990 decennial census and, once again, efforts to secure statistical adjustment failed.[4]

2 Census Bureau ethnographic research began to provide useful insights in the 1980s, but the most important research for understanding the multiple causes of undercount stems from the Bureau's 1986 TARO (Test of Adjustment-Related Operations) research in the Los Angeles Basin. Analysis by David Fein and Kirsten West of finding from that initiative, particularly their analysis of data from the "Causes of Undercount" survey component have been crucial (Fein and West 1988; Fein 1989; West and Fein 1990). The Census Bureau's subsequent ethnographic research program in connection with the 1990 decennial census also has provided very important insights. I reviewed this research in detail and relied on it in several studies of differential undercount of migrant and seasonal farmworkers (Kissam and Jacobs 2006; Kissam 2012) and to estimate the overall undercount of Mexican immigrants in the United States (Kissam 2017).
3 Differential undercount is analyzed here in the context of census tabulations of data because census operations always include a number of procedural steps that seek to augment, enhance and/or adjust for non-response or erroneous response. What is often thought of as "census data" are actually the result of a sequence of data analysis procedures.
4 See Linda Greenhouse, "High Court Rules Results Are Valid in Census of 1990," New York Times, March 21, 1996, for details on the decision. The 1990 plaintiffs included the cities of New York, Los Angeles, Chicago, the U.S. Conference of Mayors, Dade County (Miami), Florida, the states of California and New York, and national groups such as LULAC and NAACP. With an expected 2020 population of about 4.6 million, the San Joaquin Valley's population is larger than the cities of Chicago and Los Angeles, and Dade County, Florida.

The plaintiffs did not prevail in the litigation seeking statistical adjustment for the 1980 or the 1990 Census, in part because it was ultimately determined that technical limitations in efforts to adjust would not necessarily yield a more accurate result. However, the Census Bureau did, then, undertake a program of research to better understand the causes of census undercount and, if necessary, go forward with statistical adjustment to correct the systematic undercount.[5] Review of that research indicates that in a census with a much higher level of census undercount than was observed in 1980 or 1990 (as is likely in Census 2020 if the citizenship question is included), statistically reliable analysis of differential undercount would be feasible.

What is new with respect to the attempt to add the CQ to Census 2020 is that adding the question will decrease census response among certain sub-populations of respondents—most definitively non-citizens and, among the non-citizens, Hispanic non-citizens and others in their social networks. There also will be negative impacts from other re-engineered but inadequately tested census operations.[6]

With improved research/analysis methodology, statistical adjustment may be feasible, but would require re-examination of old assumptions and research due to the dramatic changes in census questionnaire design and operations.

The quest for a fair and accurate census requires not only national estimates of differential undercount, but also robust efforts for smaller geographic areas, most obviously the political jurisdictions where census accuracy has the greatest impact on equity—counties, sub-state regions such as the San Joaquin Valley, as well as entire states.

If it were to rely on multiple methodological approaches, supplementing standard dual-system (DSE) analysis currently planned by the Census Bureau, with enhanced demographic analysis (DA) and ethnographic analysis (EA), such research might make an important contribution toward overcoming what will almost certainly be regional disparities resulting from shortcomings in Census 2020 design and implementation.

Understanding the Dynamics of Differential Undercount

In the mid-1980s, as a result of growing policy concern and litigation about the impact of differential undercount of minorities, the Census Bureau initiated a research program to better understand patterns of undercount. This program resulted in greatly improved understanding about the multiple causes of census undercount.

Two lines of research were particularly fruitful in improving researchers' understanding of the dynamics of undercount—a multi-stranded research initiative in the Los Angeles basin as part of the 1986 Test of Adjustment Related Operations (TARO) and a subsequent national initiative as part of the 1990 Decennial Census—the ethnographic alternative enumeration research conducted by the Census Bureau's Center for Survey Research.

Based on the Los Angeles research, a sound theoretical framework to understand the causes of census undercount, not simply correlations, emerged (Fein 1989, Fein and West 1988, West and Fein 1990). The powerful insight stemming from this research is that undercount results from conflict between the census system processes for enumeration/ data collection and the social universe in which respondents live.

The ethnographic research program made important contributions to understanding undercount. It analyzed how different interactions between the socioeconomic and cultural context of potential census respondents' lives and census operations gave rise to differential undercount of low-income minority and immigrant families (De La Puente 1992, De La Puente 1993). The program meticulously documented undercount in 31 distinct ethnic groups in different communities across the U.S.

Moving Beyond a Focus on Self-Response Toward Understanding Multiple Causes of Undercount

A major preoccupation in the Census Bureau, which has emerged from a decade of budget constraints and constant

effort to cut the costs of non-response follow-up (NRFU), is self-response—because it is seen as the first stage in the census process.[7]

Differential undercount in the decennial census, and the American Community Survey (ACS), stems not only from lack of motivation among respondents but also from operational shortcomings embedded in census operations. Effective strategies for implementing a decennial census that both accurately counts the population and provides a reliable demographic and socioeconomic profile of the population require not only attention to self-response in the decennial census, but also careful in-depth attention to the multiple causes of differential census undercount.

The ubiquitous mapping of census tracts based on their projected low-response score (LRS) and corresponding plans for outreach to hard-to-count populations provides valuable insights for census strategy, but has distracted the Census Bureau and census advocates alike from adequately addressing structural causes of undercount stemming from the mismatch between decennial census operational strategies for Census 2020 and community context described by Kirsten West and David Fein more than three decades ago.

Although the bulk of public discussion of the impact of the citizenship question on census fairness and accuracy has focused on the citizenship question's impact on self-response, in actuality, eventual differential undercount is determined not simply by self-response, but also by subsequent enumeration efforts—enumerator visits to non-responding households and efforts to secure proxy interviews—as well as by subsequent statistical procedures used by the Census Bureau to compensate for non-response.

Widespread attention has been given to promoting census response among less-motivated households, but the most effective strategies to work toward a fair and accurate census in 2020 will need to also include a firm understanding of the ways in which census non-response initiates a cascade of subsequent operational steps to overcome the initial problems of non-response, and what may be possible pressure points for intervention once the cascade has begun.[8]

In addition to longstanding nonresponse follow-up operations seeking to increase enumeration of

households—including enumerator follow-up visits and proxy interviews—the Census Bureau has made a notable change in its re-engineering for Census 2020. Plans for implementing nonresponse follow-up now include the proposition that it will be possible to rely on administrative records to determine the size and characteristics of non-responding households. This notion is problematic as part of a strategy to accurately enumerate immigrant neighborhoods. In hard-to-count areas such as those where the San Joaquin Valley Latino immigrants live, diligent as they are, these efforts to assure census accuracy will inevitably fall short.

The analysis presented here of the cascade of valiant but compromised Census Bureau efforts to generate accurate census data provides a basis for examining how each step in the census enumeration process in hard-to-count areas in regions such as the San Joaquin Valley contribute to a cumulative regional undercount.

Using the Cascade Model to Estimate Differential Undercount of Latino First- and Second-generation Immigrants in the San Joaquin Valley

An important consideration in the San Joaquin Valley Census Research Project (SJVCRP) survey design, given well-justified and widespread concern about the impact of adding the CQ to Census 2020, was to examine undercount not simply for an ethnic/racial group, i.e. Hispanics, but for sub-populations distinguished on the basis of immigration and citizenship status:

- undocumented immigrants,
- foreign-born legal residents,
- naturalized citizens,
- second-generation immigrants (the adult U.S.-born children of Latino immigrant parents).

Consequently, the Latino immigrant population surveyed in the research consists of potential census respondents 18+ years of age, referred to in census terminology as the "householder" or "P1," the person who is the census

---

7 In actuality, as discussed subsequently in this paper, "MAF-building," the Census Bureau's process of developing its Master Address File presumed to represent the universe of housing units in the U.S., is the first stage in census data collection and an important cause of census undercount, even before households have begun to respond (or not).

8 The Census Bureau's framing of its findings regarding NRFU in its 2016 Census Tests in Los Angeles and Houston and the 2018 End-to-End Test in Providence, Rhode Island, presume that observed self-response rates and NRFU interview success rates will be substantially improved in the 2020 decennial census. There is not much evidence to support this overly optimistic assessment. Messaging helps, but is not a magic bullet. Self-response rate in the 2016 test in Harris County (Houston) Texas, was 39.3% and, in Los Angeles, 52.5% (Jennifer Reichert, "Findings from the 2016 and 2017 Census Tests," presentation to State Data Centers Annual Meeting, April 6, 2018).

respondent and to whom the household census roster is referenced.[9]

Based on survey findings and assumptions about MAF completeness, availability and accuracy of administrative records for non-responding households, partial household undercount in "complex" households, and errors arising from hot-deck imputation, the cascade model provides the basis for generating the San Joaquin Valley Census Research Project estimate of undercount.

The model projects there will be a 21.1% undercount of Latino households headed by undocumented immigrants, 7.5% for those headed by foreign-born legal residents, 5.9% for those headed by naturalized citizens, and 10.3% for those headed by second-generation citizens.

The aggregate undercount in any geographic region is the sum of undercount among households in each undercounted population within the region.

The first- and second-generation Latino immigrants surveyed and for whom we can empirically project self-response, response to enumerator visits and response to requests for proxy interviews is large—making up about 35% of the entire universe of potential census respondents in the San Joaquin Valley. The ancillary component of the cascade model computes regionwide undercount for each sub-population surveyed, by weighting the sample to approximate the prevalence of each of the survey sub-populations in the region. This indicates that the undercount of Latino first- and second-generation immigrant households will result in a 4.1% undercount of the entire population in the region—an undercount of about 188,000 persons.

How Variations in Response Rate and Structural Causes of Undercount Will Be Transformed Into Eventual Differential Undercount in Census 2020

The ways in which non-response erodes census accuracy are complex and vary in relation to specific populations being enumerated, survey design and community context. But there is no doubt that when any survey—decennial census included—encounters high levels of non-response, accuracy is degraded because the process comes to rely less on "data" collected, actual empirical data, than on statistical processes utilized to adjust for lack of data or erroneous information supplied by reluctant respondents.[10]

Higher levels of non-response predictably result in incremental errors and uneven levels of non-response among different groups. This, in turn, inevitably results in differential undercount of hard-to-count sub-populations. However, it cannot be stressed too often that many of the causes of undercount are structural. That is, they derive from the way in which census operations play out in different neighborhoods, with distinct housing patterns and ethnicities. Essentially, dramatically heightened levels of non-response among the Latino immigrant population studied in the SJVCRP set the cascade of flawed enumeration in motion—because the errors arise from standardized but imperfect efforts to compensate for non-response.

Factors in the cascade of census operations, the model described here, identifies as determinants of the eventual accuracy of the census count for a neighborhood, community, county or state include the following:

- Success/failure in MAF-building, i.e. generating a complete address list with omissions of housing units leading to total HH omissions
- Success/failure in securing self-response (via return mail or online)
- Success/failure in securing an in-person NRFU interview with a non-responding HH
- Success/failure in securing a proxy interview if a non-responding neighbor's HH is unavailable for an interview or refuses
- Success/failure in securing a high-quality administrative record match for HH's that did not self-respond, respond to an enumerator or be "enumerated" via proxy interview
- Bias in HH size and characteristics stemming from under-reporting in complex HH's (partial HH undercount)
- Bias in HH size and characteristics due to available administrative records omitting some HH members (partial HH undercount)

9 In the real world of day-to-day household life in Latino immigrant communities, a decision to respond or not respond to the census may often be more a family decision than the decision of a particular person. The notion that there is an easily identifiable "householder," possibly a male "head of household," who fills out the census form is antiquated and ethnocentric. For example, a number of second-generation Latino immigrants told the project's researchers, as part of survey response or in focus groups, that they would advise about response and help them if they were to respond.
10 See Joseph Salvo, "Expert Rebuttal Report: Errors in the Local Census," November 11, 2018). Interestingly, an important legal consideration in the 1980 and 1990 strategy to oppose statistical adjustment was the constitutional reference to "actual enumeration." The cascade model suggests the situation will be subtly different in Census 2020 because Census Bureau re-engineering has shifted census data collection methodology away from actual in-field data collection (in-field address canvassing, survey design to encourage higher response rates, robust NRFU) toward efforts that, however characterized, involve some form of alternative to actual enumeration (proxy interviews, reliance on administrative records, and count and whole-person imputation).

- Bias in HH size and characteristics stemming from recourse to hot-deck imputation of non-responding HH size due to systematic differences in HH size between donor HH's and non-responding HH's

The cascade model indicates that, at each stage, there is erosion in data quality despite Census Bureau efforts to secure information from respondents and from inaccuracies in the surrogate sources of information it utilizes in lieu of household census responses.

The major sources of error are omission of housing units from the decennial census sampling frame (either due to not being in the MAF or being erroneously identified as vacant) and systematic bias in hot-deck imputation. Data quality is further eroded by partial household undercount stemming from incomplete/out-of-date administrative records being used as a basis for enumeration, from under-reporting in complex households and, quite possibly, from systematic bias in proxy interviews.

Moreover, in areas such as the San Joaquin Valley (and most other immigrant communities across the U.S.) where the CQ is widely believed to represent a threat or to be improper, there is decreased willingness to participate in an apparently politicized census and data may be incomplete or partially falsified. Here it is particularly important to remember that incomplete or partially falsified responses affect the enumeration of different individuals in the households where householders filling out the census form modify their responses due to apprehension about the consequences of providing information to the government.

The Cascade Model of Census Undercount in the San Joaquin Valley

The San Joaquin Valley Census Research Project estimate of an 11.7% undercount of Latino immigrants in the region is conservative. It does not, for example, attempt to estimate the additional negative impact that constrained access to the Internet and low digital literacy, coupled with print literacy, might have on self-response rates.[11]

The model, first, gives attention to the Census Bureau's limitations in generating a complete address list that includes low-visibility unconventional or hidden housing units (the "bad MAF" problem). It does not include an estimate of possible erroneous deletions of occupied housing units that are incorrectly believed to be vacant (due either to errors in administrative records or enumerator judgment).

The model also gives special attention to the extreme reluctance observed in the San Joaquin Valley Census Research Project survey to proxy interviews used by the Census Bureau to secure information on non-responding households. As well as being difficult to secure, it is crucial to understand that proxy interviews are error-prone because they are, at best, estimates by neighbors—some accurate, but some inaccurate—of the number and characteristics of neighboring households. Another important cause of undercount is due to the serious problems with enumerations based on securing an apparently matching administrative record that the Census Bureau envisions using to determine the characteristics of a household that has not self-responded, has not been contacted by an enumerator or removed from the NRFU workload by a proxy interview with a neighbor. These characterizations of household size and demographic profile based on reference to administrative records are, of course, not actually enumerations and are an element of Census 2020 operations that has not been adequately tested. Even when an apparently matching administrative record is found for an immigrant household, it is suspect as a source of information about household size and composition because it may be out of date and is very unlikely to include peripheral household members who are not part of the primary core family living in the housing unit.

Finally, at the end of the cascade of Census Bureau efforts to secure information on non-responding households, there is the last-ditch attempt to use hot-deck imputation to determine the characteristics of households that have not responded, for which there is no information from a proxy interview, and where no matching administrative record can be found.

This final stage in the decennial census process is particularly problematic in the San Joaquin Valley and probably in other communities with dense concentrations of immigrants. The ubiquitous problem in use of hot-deck imputation in the San Joaquin Valley is that the non-responding Latino households are systematically larger than those that do respond.[12] Therefore, when a nearby responding household is considered to be the

11 The data on survey respondents' level of educational attainment is relevant here. About half of the population has only an elementary school education. A number of respondents' comments about considerations entering into their response included reference to their inability to read or write (in Spanish or in English). Interestingly, even some who had access to the Internet and used applications such as Facebook, for example, said they were illiterate.
12 Hot-deck imputation refers to efforts to impute the size and characteristics of a non-responding household where information from a proxy interview or an administrative record is also unavailable by assigning it the characteristics (size and/or demographic profile) of nearby households.

donor for imputing the size and characteristics of a non-responding household, each imputation will contribute to systematic differential undercount.

There are inevitably some uncertainties in the cascade model of undercount used here to generate an estimate of regional undercount projections. It has not yet been possible to adequately explore a full range of assumptions/scenarios regarding availability of administrative records or variations in errors introduced by hot-deck imputation. The assumptions incorporated in the model where there is the greatest uncertainty relate primarily to two factors: availability and accuracy of administrative records to be used in lieu of actual enumeration (i.e. via self-response or household response to an enumerator), and the composition of the pool of households used as donors for imputing the characteristics of non-responding households for which administrative records are not available.[13]

Despite uncertainties inherent in the cascade model, there is, nonetheless, no doubt that non-response will result in very elevated levels of differential undercount in the San Joaquin Valley and other similar areas of the U.S. For example, in the neighborhoods where there are the densest concentrations of households headed by undocumented immigrants, hot-deck imputation may not generate errors as serious as in more mixed neighborhoods—but in these neighborhoods, more of the households will lack matching administrative records so relatively more will need to be imputed.

Coefficients used in the cascade model will be adjusted in subsequent iterations of the model, as further details on Census Bureau implementation of its 2017 Operational Plan become available (scheduled for March 2019), additional findings from the 2018 End-to-End Test of census operations in Providence County, Rhode Island, are reported, and a proposed Census Bureau split-panel mail survey of response and non-response planned for summer 2019 are published.

The model framework for analyzing undercount and factors that affect undercount that enter into the cascade model projection of undercount are discussed in more detail below.

Survey-Derived and Estimated Coefficients in the Cascade Model

The cascade model of undercount incorporates empirical data from the San Joaquin Valley Census Research

Project survey for modeling expected success rates for self-response, response to enumerators, and proxy response in the census process among the first- and second-generation Latino population.

The model also incorporates empirical data from the survey regarding the size of non-responding and responding Latino households. These data provide the basis for estimating inaccuracies resulting from use of hot-deck imputation (because non-responding households are generally larger than those likely to respond).

The model also incorporates survey-derived estimates of the proportion of persons in complex households that do respond to the census, but where "extra" non-family members are likely to be left off the household census roster.

The cascade model also incorporates assumptions regarding the Census Bureau's success in securing administrative records for non-responding households where no proxy interview can be conducted, and the reliability of the information in the administrative records is questionable. It also projects the impact of hot deck imputation on the eventual undercount.

The Cascade of Semi-Successful Efforts in the Nonresponse Follow-up Process Meant to Compensate for Households' Failure to Self-Respond

The success of NRFU depends on many factors—some of which can be predicted more reliably than others. Key considerations are discussed below.

Receiving an Enumerator Visit During NRFU if the Housing Unit is not in the MAF?

An important consideration, in addition to a householder's willingness to self-respond, has to do with their having an opportunity to self-respond or respond to an enumerator during the course of NRFU.

One-fifth of the SJVCRP Latino survey respondents who were in the U.S. in 2010 say they never received a census form in 2010 and were not contacted by an enumerator. Their recollection may be imperfect, but surely there is reason to be concerned about thoroughness of NRFU due to errors in the MAF.

13 Andrew Keller has reported details of Census Bureau research on imputation (Keller 2015). What is most relevant in the current context is that, although extent of reliance on imputation was low at the national level in 2010, it can be expected to be much higher in 2020 and that it will be higher in some geographic areas (such as the San Joaquin Valley) than others. Moreover, imputation of household characteristics was more often necessary than count imputation (of HH size). See Keller, A., "Imputation Research for the 2020 Census," Statistical Journal of the IAOS 32 (2016) 189–198.

Operational/Logistics Shortcomings in Implementing NRFU in Areas with Extremely High Non-Response Rates?

Operational failures in NRFU, a likely consequence of overly optimistic Census Bureau projections of 2020 self-response rates, would, inevitably, make the actual undercount higher than the cascade model currently projects.

Salvo and Lobo (2013) argue, for example, that an unmanageable NRFU workload in parts of New York City resulted in many occupied housing units being incorrectly classified as vacant—just to get them removed from an enumerator's work assignment. Enumerators are also able to remove a housing unit from their workload if it is deemed unsafe—an understandable provision, but worrisome as an option for an individual enumerator without the requisite communication/social skills—to reduce their workload.

A December 2018 GAO report, for example, also points to Census Bureau research showing that rushed enumeration where there is a higher-than-expected nonresponse follow-up (NRFU) workload contributes to enumeration errors.[14] Problems encountered by the Census Bureau in accurately gauging the extent of non-response in the Latino immigrant neighborhoods and staffing NRFU operations may contribute to differential undercount.

Impact of the Logistics Challenges Involved in Enumerating Households of Working Poor?

Another uncertainty is that it is not known exactly how many of NRFU enumerator visits may fail to yield an interview simply because the enumerator visit took place when there was no household respondent at home. The current operational plan (as per the Census Bureau's June 8, 2018, Federal Register Notice) is that enumerators will be required to make three attempts to contact a non-responding household; after three unsuccessful contact attempts, three efforts will be made to conduct a proxy interview.

We know, from research on 2010 Census coverage in hard-to-count tracts in agricultural areas of California, for example, that enumerators' ability to establish rapport with non-responding households (likely to be similar to those in our survey who responded that "maybe" they would answer the census) will affect completion both of NRFU direct interviews (with reluctant households) and proxy households. Another challenge is that, as discussed in the report on San Joaquin Valley Census Research Project survey findings, reluctance to participate in proxy interviews is extremely high.

Limitations of Reliance on Administrative Records to Enumerate Households in NRFU?

One of the most serious potential problems connected to Census Bureau operational plans for Census 2020 implementation in the San Joaquin Valley is the viability of using administrative records to enumerate non-responding households. The uncertainties here stem from the fact that there is no empirical data on the proportion of Latino immigrant-headed households in the region for which there will be high-quality matching administrative records. The cascade model projects that there will be serious limitations on finding matching administrative records for non-responding households.

Moreover, it is assumed that, despite the appearance of a match between a household and an administrative record for the address in some cases, apparently matched records will systematically omit some of the actual household members because they are out of date or underlying information is inaccurate. The entire Census Bureau proposition of relying on administrative records to impute household size (and characteristics) of non-responding households is a novel and untested one introduced only after efforts were made to add the citizenship question.[15]

Key Threats to San Joaquin Valley Census Data Quality and the Components of the Cascade Model

Below are details on key components of the cascade model used in estimating how non-response translates into undercount.

The model predicts Census Bureau <u>success</u> rate at each step in the census process and estimates the percent of actual population enumerated at each stage. This prediction then provides, as the analysis moves through each stage in the process, an estimate of <u>cumulative</u> enumeration—after self-response, after response to enumerator NRFU visits, after enumeration via proxy interview, and enumeration via reference to administrative records.

---

14 Government Accountability Office, ""2020 Census: Additional Steps Needed to Finalize Readiness for Peak Operations, GAO Report 19-140, December, 2018
15 Census Bureau research on use of administrative records has focused primarily on using such records to reduce NRFU workload (and cost) by identifying vacant housing units that do not need to be enumerated.

It is then assumed the remainder of non-responding households that have not been enumerated in any of these operations will need to be imputed. Although, historically, it has not been necessary to rely extensively on imputation, there are many reasons to believe the situation will be different in Census 2020, at least in communities of Latino and other immigrants, because of the anticipated high levels of non-response among these populations, along with serious difficulties to be expected in "refusal conversion" efforts if the CQ is included. There are reasons to believe that some of the standard census procedures such as reminder postcards will be minimally effective in the sociopolitical context of a census with the CQ.

Having determined the proportion of households in each sub-population where size and household characteristics will have to imputed, the model then examines under-count that stems from erroneous hot deck imputations of the size and characteristics of the remaining residents of households that have not been enumerated.

The cascade model also takes into account errors introduced in the course of enumeration—most notably errors stemming from undue reliance on inaccurate administrative records, but also enumeration errors stemming from complex household respondents' omission of peripheral household members.

Before Enumeration Begins—Housing Units Omitted from the Census Bureau's Address List

It is generally agreed that the sampling frame for the decennial census always omits some low-visibility unconventional and/or hidden housing units. Although there is limited data on the pervasiveness of this problem, we have recently documented the prevalence of missing housing units in several major California counties and communities (Kissam, Quezada, and Intili, 2018). This research generated relevant empirical data on the completeness of the Census Bureau's address list (MAF – Master Address File) in the San Joaquin Valley.

Community-based address canvassing linked to LUCA was conducted in Stockton and in Fresno in areas where unconventional and/or hidden housing units were prevalent. MAF quality varied from census tract to census tract in the community-based canvassing, but prevalence of hidden housing units averaged 4.8% in canvassed areas (Kissam, Quezada, and Intili 2018). The San Joaquin Valley cascade model assumes that for U.S-born citizens, naturalized citizens and legal residents, 3% of the

population live in unconventional and/or hidden housing and that, for undocumented immigrants, 5% of households live in this type of low-visibility housing.

Therefore, the San Joaquin Valley cascade model begins with the assumption that, as a result of incomplete address canvassing at the first stage in the census process, only 95% to 97% of the Latino immigrant study population live in housing units included in the Census Bureau's MAF and can be enumerated.

Persons living in housing units that are not included in the MAF do not generally get a mailed invitation to respond to the census, a follow-up paper form if they fail to respond or an enumerator visit. A few hidden housing units not in the MAF may be identified in U/L (Update–Leave) areas of the San Joaquin Valley, but, nonetheless, the assumption of 3%-5% total household omissions due to low-visibility occupied housing units not being on the Census Bureau's address list appears to be well-founded.

If there were to be higher-than-expected designation of TEA's[16] for U/L and increased use of U/E (update–enumerate), it would ameliorate the impact of this particular cause of undercount, but current plans and budget constraints make this unlikely.[17]

The cascade model is conservative in its estimate of undercount in that it does not seek to further adjust for barriers to self-response noted in the discussion of the SJVCRP survey findings, e.g. substantial numbers of households without their own mail address (28%). The 13% of households who report they have a PO Box may or may not receive an invitation to respond to the census (depending on Census Bureau determination regarding TEA), and the 15% who share a mailbox with others also may or may not receive an invitation or mailed paper form.

Conceivably, innovative and aggressive messaging campaigns to urge households living in hidden housing units and those without their own mailbox to proactively respond via the online non-ID processing option (essentially, an online "Be Counted" option) might have a positive impact, but this only would have promise if there were also facilities to assist these households in submitting an online response.

16 TEA's—Types of Enumeration Areas is a classification that reflects the various operations and methods of enumeration used to collect addresses and enumerate areas.
17 The Census Bureau's 2017 Operational Plan and quarterly Program Management presentations make it clear that there will be virtually no use of U/E and limitations on use of U/L. While reductions in utilization of these old-fashioned enumeration approaches were curtailed as part of cost containment efforts, estimates of the negative impact on data quality are sketchy.

It is unclear to what extent online non-ID (NID) processing might contribute to inclusion of households living in the hidden housing units and those without their own mail delivery, but what is clear is that there are many barriers in the way—both inadequate motivation to take proactive steps to be counted in the context of a census perceived as being unfriendly to immigrants and lack of Internet access.[18]

Level of Self-Response—the Key Determinant of NRFU Workload

The cascade model uses self-response rates derived from the San Joaquin Valley Census Research Project survey. These are computed based on expressed willingness to respond minus 5%—half the CBAMS-based adjustment of 10% accounting for the gap between stated willingness and eventual response behavior.

The resulting presumed self-response rates are: 58% for legal residents, 65% for naturalized citizens, 44% for U.S.-born second-generation immigrants, and 20% for undocumented immigrants. These rates of self-response are not inconsistent with those observed in the 2016 Los Angeles and Houston test censuses or the 2018 End-to-End Test. In fact, they might be considered optimistic given the fact that both of these tests took place before the decision to add the CQ to the census was announced.

The actual fall-off from expressed willingness to respond to eventual self-response might be lower if assurances about confidentiality of census data could be framed so as to gain widespread acceptance. Conversely, fall off might be higher if concerns about possible government misuse of census data continue to rise despite repeated assurances from the Census Bureau and census promoters regarding confidentiality of individual responses and disclosure avoidance. A very specific challenge is that the Latino survey respondents did not view the issue of confidentiality in a context specific to the Census Bureau, but, rather, as part of their overall environmental scan of federal government behavior and presidential rhetoric about immigrants and, specifically, about Mexicans.[19]

Even if overall levels of non-response among Latino immigrants were to be reduced somewhat, the survey findings suggest that there would continue to be disparities in response between sub-groups of Latino immigrants. Naturalized citizens might be more likely to

respond and there might even be slightly higher rates of response than the survey and focus group discussions indicate among second-generation immigrants. But it is very unlikely that response among undocumented households, a substantial sub-population of immigrants, would increase.

NRFU Step 1: Adequacy of Administrative Records for Identifying Occupied vs. Vacant Housing Units

Census Bureau research from Census 2010 shows that, nationally, the NRFU workload lacked 4% of all actual housing units (based on results of Vacancy Deletion Checks).[20] This may be a problem in the San Joaquin Valley in Census 2020—particularly in areas with major fluctuations in seasonal occupancy and sub-standard housing conditions. This potential problem is not, however, explicitly included in the cascade model calculations at this point because we do not yet have a basis for assessing how serious this problem will be. It is a consideration, however, in working to refine estimates of the proportions of actual occupied housing units that will be enumerated in NRFU.

NRFU Step 2: Direct Interview Completion (Interview with a non-responding household, excluding proxy interviews)

The cascade model assumes that respondents' willingness to respond to an enumerator who comes to interview a non-responding household is the general willingness to respond expressed in their answers in the San Joaquin Valley Census Research Project survey answers to Q. 5. In actuality, the rate of success in securing direct interviews rests not only on a household's willingness to respond to the enumerator, but also the basic logistic challenge faced by enumerators in making contact and in finding an adult householder willing and able to respond at home.

GAO's report on the 2016 Census Tests conducted in Harris County, Texas, and Los Angeles County, California, mention a NRFU interview completion rate of 70% (Harris County) to 80% (Los Angeles County). However, the definition of "NRFU interview completion" included both direct enumerator interviews with non-responding households and proxy interviews (which made up 25% of all NRFU interview completion in 2010). Therefore, the

---

18 For a full description of this process, see Census Bureau, "Census 2020 Detailed Operation Plan for: 13- Non-ID Processing (NID)," June 2018. There would appear to be undercount risks in areas with impaired Master Address File quality in response processing to the extent this relies on matching Non-ID ISR (Internet Self-Responses) to addresses in administrative records.

19 The SJVCRP report on Latino immigrant survey respondents' perspectives on census response or non-response, "Troubled Reflections: San Joaquin Valley Immigrant Perspectives on Census 2020" (forthcoming, February 2019) shows that response is likely to be very sensitive to sociopolitical context and media coverage of administration policies vis-à-vis immigrants.

20 Thomas Mule, Census Coverage Measurement Memorandum Series #2010-G-01, May 22, 2012.

cascade model estimate can be compared to the test census results by recognizing that the direct interview rate in Harris County was about 52.5% and in Los Angeles County about 60%. A still more worrisome recent report from GAO is that in the Providence, Rhode Island, End-to-End 2018 Test, there was a 33% non-interview rate in the NRFU workload.[21]

Consequences of Truncation of Enumerator Return Visits?

As noted in the prior discussion of NRFU plans, the current Census Bureau operational plan (as per its June 8, 2018, Federal Register Notice) is that enumerators will be required to make three attempts to contact a non-responding household. After three unsuccessful contact attempts, three efforts will be made to conduct a proxy interview.

The Census Bureau's decision to truncate the number of enumerator return visits seeking a direct interview may be particularly problematic in the Latino immigrant neighborhoods in 2020. The Census Bureau's 2010 NRFU Contact Strategy experiment showed that nationally about 42% of non-responding households had been successfully interviewed after three NRFU contact attempts. Although refusals were low (< 3%) at the third visit, a substantial number of contact attempts (26%) resulted simply in a "notice of visit" being left or recorded as "no contact" (Compton and Bentley 2012). It is also important to note in the tabulation of results from the experiment that the level of proxy interviews for occupied households was very high (30%) at the third contact attempt.

The Census Bureau's basic NRFU plan vis-à-vis number of visits to non-responding households apparently may also involve reliance on a business model to determine the cost-effectiveness of return visits by enumerators. A NRFU algorithm for authorized number of enumerator return visits to a non-responding household based on cost-effectiveness is worrisome because it might predict that return visits were less cost-effective in neighborhoods such as those that San Joaquin Valley immigrants live in, where willingness to respond is lower and where making contact is more challenging because low-income family members may work longer hours.

In some cases, a household may only be visited once before an attempt is made to enumerate it via proxy interview or by recourse to administrative records. Even in cases where more than one visit may have been made to a household that is, in fact, willing to respond, a NRFU

interview may still not be completed because the business model required the enumerator to give up too soon and because the alternative strategies for securing data from the household (e.g. a reminder postcard left at the door) may be ineffective.

The Census Bureau alleges that its enumerator deployment software will optimize enumerator visits to make it as likely as possible for the enumerator to find an adult at home. However, it is very unlikely that the standard optimization model will do well in the San Joaquin Valley Latino immigrant community, where work hours are often long and where weekend work is common. It deserves note also that the 2018 End-to-End Test showed that unaccountably there were fewer enumerator visits scheduled for Saturdays, the day when actual chance of contacting a household respondent was highest.

Consequences of Limited Availability of Waivers to Hire Non-Citizens as Enumerators?

Another area of uncertainty vis-à-vis enumerator success in securing household interviews stems from concerns about the Census Bureau's ability to hire enough linguistically competent/culturally competent enumerators to successfully persuade undecided households to respond. Current management priorities in the Census Bureau are focused on hiring enough enumerators to get the job done and there is less attention to reliably assessing enumerators' ability to persuade reluctant households that have failed to self-respond to go ahead and consent to an interview with the enumerator.

The survey findings suggest that refusals of enumerators' interview requests in NRFU may turn out to be higher in 2020 than ever before based not only on apprehension about the consequences of providing information in the course of a census that includes the CQ, but also on survey respondents' frequent comments that they have learned not to open the door to strangers—due to a variety of commercial scams and guidance from immigration legal advisors regarding ICE visits.

The Model Estimate of Enumerator Success in Securing Interviews with Households That Failed to Self-Respond

For the purpose of projecting undercount, the cascade model uses the specific levels of response for undocumented and legal resident non-citizens and for naturalized and U.S.-born Hispanic second-generation immigrants for the

---

21 See page 22, GAO Report 19-140, December 2018.

respective groups as a proxy for estimating direct interview success rate. The aggregate level of response for all Latino immigrants will probably be about 41% for the overall population of Latino immigrant households, depending on the eventual behavior of those in each sub-population of potential census respondents who said that "maybe" they would respond to the census if it included the citizenship question.

NRFU Step 3: Efforts to Secure Proxy Interview Response Rate

Proxy interviews are an important component of NRFU. The 2017 GAO report suggests they made up 25% of enumerations in the 2017 census tests in Los Angeles and Houston. The 2010 NRFU contact strategy experiment showed that 30% of NRFU interviews after the third contact attempt were proxy interviews. The 2018 End-to-End Test showed that they accounted for 27% of NRFU interviews.[22]

The San Joaquin Valley Census Research Project survey data are clear-cut in projecting response rate to enumerator requests for information on neighboring households as being only 8% affirmative if Census 2020 includes the citizenship question. There will also be variations that may further decrease proxy interview completion rate—depending on proxy interview respondents' knowledge regarding a specific household they might be asked about. The SJVCRP survey findings, including survey respondents' discussion of their perspectives of willingness and ability to provide information on neighboring households to a census enumerator, indicate that the 8% rate is the best that can possibly be expected.

It is also worthwhile to note that the Census Bureau's discussion of its procedures for securing proxy interviews is extraordinarily ill-suited to the San Joaquin Valley, apparently being based on assumptions about urban neighborhoods (e.g. enumerators requesting information from local real-estate agents or landlords).[23] However, Joseph Salvo, New York City's Demographic Unit Director for the city's Planning Department, a leading census expert, recently explained that this sort of effort is not well-suited to urban neighborhoods either.

NRFU Step 4: Using Administrative Records to Impute Household Size and Characteristics

Perhaps the greatest challenge is the Census Bureau's

ability to compensate for dramatically increased levels of non-response due to the citizenship question by relying on administrative records to secure information on non-responding households. There are many questions regarding the eventual viability of such record-matching and using such records to enumerate non-responding households.

Extent to Which the Census Bureau Will Attempt to Use Administrative Records?

The Census Bureau's Federal Register Notice regarding its proposed Census 2020 operations states that administrative records will be used as follows, "enumerating households that do not self-respond and whom we were unable to contact after six mailings and one in-person field visit."[24]

In the neighborhood/community context in the San Joaquin Valley where Latino immigrant households are concentrated, if there is a very low initial response rate and contacts are difficult due to many adults in households working long hours, there will be very heavy reliance on administrative records for enumeration.

It is only recently that the idea of using administrative records to impute characteristics of entire households was adopted (see *Wall Street Journal* April 3, 2018, story by Paul Overberg and Janet Adamy, *"Trump Administration Plans to Check Your Answer On Citizenship Question.")* This element of the Census Bureau's planning was only introduced after publication of its 2017 Operational Plan, where administrative records were to be used only for identifying vacant housing units to be excluded from the NRFU workload. It was not part of the census test in Rhode Island and is, therefore, almost completely untested.

The Census Bureau has said in various public statements that it is very optimistic about being able to secure high-quality administrative records to provide information on non-responding households. However, Census Bureau research over the decade has focused on a specific and justifiable (though possibly flawed) use of administrative records—to identify and eliminate from the NRFU workload housing units that are not occupied—not on use of administrative records as a source of information on household characteristics.

22 See page 24, GAO Report 19-140, December 2018.
23 Described in the Census Bureau's July 2018 Federal Register Notice requesting comments on planned census operations.
24 Federal Register, June 8, 2018 (Vol. 83, No. 111, page 26643).

Although the *Wall Street Journal* article focuses on Secretary Wilbur Ross's announcement that the Census Bureau would refer to administrative records to check the correctness of answers provided by respondents regarding citizenship status, it is obvious that such an endeavor presumes the possibility of securing comprehensive household data from administrative records. Census Bureau Scientific Director John Abowd is quoted in the article, referring to Secretary Ross's statement that Social Security applications would be used for this purpose, as saying that "the bureau would have to create a statistical model but hadn't begun to figure out how."

Availability of High Quality Matching Administrative Records for Latino Immigrant Households?

The Census Bureau's own research shows there will be limited availability of administrative records for the kinds of households most prevalent in San Joaquin Valley immigrant communities, because those who are least likely to respond are also those for whom there is least likely to be an administrative record match.[25] Census Bureau researcher Rhenuka Bhaskar and her colleagues explain that matching a household to an administrative record requires a Personal Identification Key (PIK). Bhaskar's research on administrative records shows that there are much lower levels of PIKs for foreign-born households than U.S-.born populations.

Characteristics associated with lacking a PIK include: number of persons in household, living in a tract where more than 45% of the population is foreign-born, Hispanic race/ethnicity, not being a U.S. citizen, limited English or no English, and being a recent immigrant (<10 years in the U.S.). These characteristics are, of course, prevalent in the San Joaquin Valley Latino immigrant networks and the study population. Administrative records will often be unavailable. It should also be remembered that Bhaskar's excellent analysis is based on examination of availability of matching records for households that are ACS respondents. It is more likely that high-quality matching administrative records are available for the immigrant households that do respond to the ACS than for those that do not—so the ACS-based estimate of availability is probably high.

Securing High Quality by Matching Administrative Records with Households or Housing Units?

Bhaskar's analysis of the availability of administrative records to be used in determining the household size and characteristics examines the Census Bureau's ability to match an administrative record to a household that has responded to the ACS, not a housing unit. It is difficult to understand how a match could easily be made at all between a non-responding household and an administrative record—based simply on housing unit information.

In the context of NRFU, the Census Bureau must secure a matching administrative record for a non-responding housing unit. Since the Census Bureau's plan is to rely on an administrative record when there is no response from a household, nothing is known about the household characteristics. All that is available is a non-responding housing unit address.

Similarly, it is entirely unclear how the Census Bureau might propose to assess the quality of an apparent administrative record match based simply on an address. Serious discrepancies can be expected in neighborhoods where low-income renter households move often. Especially in the low-income neighborhoods in the San Joaquin Valley, where families may often be forced to move due to cash flow crises stemming from seasonal unemployment, the year-old SSA or IRS record for a household may often not match up with the household currently living at an address.

Presumably, attempts would be made to secure SSA or IRS records for the address of a non-responding housing unit to an income tax filer or individual in the SSA Numident database whose record has that address. But in many cases, especially for undocumented immigrants, the Latino immigrant sub-population most likely to fail to self-respond, respond to an enumerator, or be enumerated via proxy interview, reliably matching Social Security Administration or IRS records will not be possible or will be unreliable. Moreover, such records are very likely to be out-of-date so that, consequently, newly born children will be disproportionately omitted.

The Institute of Taxation and Economic Policy has estimated that about half of all undocumented workers in the U.S. file income tax returns (Gee et. Al 2017). It is likely that still fewer of the undocumented workers in the San Joaquin Valley—many of whom are farmworkers—are likely to actually file income taxes and, thereby, generate an administrative record with the IRS.

25 Rhenuka Bhaskar et al, "Assimilation and coverage of the foreign-born population in administrative records," SJIAOS (2018).

Although employer reports to the SSA of employees' earnings may be quite complete, they provide limited and unreliable information about household size (as defined by the Census Bureau OMB residence rules to refer to everyone living under the same roof—irrespective of economic/social relationships).

The Cascade Model's Conservatively Projected Level of Census Bureau Success in Relying on Administrative Records for Enumeration

Provisionally, the cascade model very optimistically assumes that matching administrative records are available for 80% of the U.S.-born non-responding householders, 70% of the naturalized citizen non-responding householders, 60% of the legal resident non-responding householders, and 30% of the undocumented non-responding householders.[26] However, as discussed above, we believe that the Census Bureau's actual ability to match administrative records to housing units is very unlikely to achieve this level of success.

The latest Census Bureau estimate is that only about 5% of U.S. households would be enumerated via use of administrative records.[27] However, this assumption seems extraordinarily optimistic with respect to the San Joaquin Valley. The cascade model projects, based on the assumption there will be high proportions of households that fail to self-respond, as well as limited success in securing direct interviews with households or proxy interviews, and limitations on securing high-quality matching administrative records, that about 8% of Latino naturalized citizens, 9% of legal residents, 23% of U.S.-born second-generation immigrants, and 17% of undocumented immigrants might eventually be successfully enumerated using administrative records.[28]

Even if administrative records are more readily available for the Hispanic first- and second-generation immigrants than is assumed in the cascade model, there remain serious questions about the impact Census Bureau efforts to rely on administrative records might have on the accuracy of census data in regions such as the San Joaquin Valley. This is because they may be out-of-date or reflect only a single family/budgetary unit in an extended family or complex household.

San Joaquin Valley Census Research Project research team members' experience in three decades of national farm labor research, as well as published reports, suggest that Social Security data on farmworkers, a population made up almost entirely Mexican-origin, often undocumented, immigrants is compromised.[29] It is common in the agricultural employment sector (and in other immigrant-dominated segments of the labor market) for unauthorized new arrivals to secure a falsified green card (mica) or to work using one borrowed from a friend or relative or purchasing one.

Moreover, in cases where matching records <u>are</u> available, it is unwise to assume that the administrative record will include everyone actually living in a housing unit where the householder has failed to self-respond to the census. The culturally misguided assumption that all or most immigrant households are <u>reliably</u> profiled in any administrative record system, including the SSA Numident database the Census Bureau envisions using or IRS records, is a serious source of potential undercount (since administrative records are unlikely to show <u>more</u> people living at an address than actually live there but often show fewer).

Erroneous Enumerations in the Course of NRFU: Partial Household Omission in Complex Households that Do Self-Respond or Respond to an Enumerator Visit

The prevalence of complex households, where non-family members are very commonly excluded from census responses by householders, means there will be a substantial partial household undercount due to omission of "extra" household members within the complex households that do respond to the census.

The problem here is that the OMB/Census Bureau residence rules continue to be indifferent to the ways in which cultural concepts of household in immigrant communities differs from the official definition. Census form instructions ask respondents to be sure they have remembered to include everyone living in a housing unit (referred to by the Census Bureau as "household") on their

---

26 The SJVCRP survey data show that many of the complex households and unconventional hidden housing units at an address are occupied by undocumented immigrants, making it very unlikely that the landlord's or property manager's administrative records will somehow include these economically and socially distinct households.

27 Census Bureau submission for OMB clearance for the 2020 Census—Request for Comments, December 28, 2018.

28 Differences in projected reliance on administrative records for the different Latino immigrant sub-populations stems from differing levels of self-response and enumerator response in each and availability of administrative records for each sub-population.

29 Data analyst Richard Mines has conducted research on California farmworkers for more than four decades and was the Department of Labor Project Officer for the National Agricultural Worker Survey for more than a decade. Researcher Ed Kissam has conducted research on farmworkers and immigrant settlers in rural communities for more than three decades.

census form. In fact, the prevailing concept of household is that it consists of a core family/budget unit. Other family units living under the same roof are typically distinguished as not being household members.

The SJVCRP survey provides a good estimate of the prevalence of complex households in San Joaquin Valley Latino immigrant communities (22% of all households), but there remain uncertainties about the extent of partial undercount in these places since it is not clear whether, in some cases, how many of the "extra" persons living at the place might be included on the primary household's census response—even though generally they are not.

The current model assumes that 20% of the non-family "extra" members in undocumented complex households <u>are</u> included on householders' census rosters, but that the remaining 80% are omitted. In households of legal residents, naturalized citizens, and U.S.-born immigrants, it is assumed that only 60% are left off the household roster.

Statistical Efforts after NRFU: Enumeration Errors from Hot-Deck Imputation of the Size and Characteristics of Non-Enumerated Households

An important cause of differential undercount is the reliability of hot-deck imputation used to determine the characteristics of households that have not self-responded, have not been successfully interviewed by an enumerator, have not been enumerated via proxy interview, and for which there is no high-quality matching administrative record. In such cases, hot-deck imputation is the last resort for the census count.

Household characteristics of non-responding households are imputed from the characteristics of nearby households that have responded—the donor pool. If these households are similar in size, imputations will, at least, provide an acceptable estimate of the census count. However, if they are systematically different in size, the hot-deck imputation process will lead to a systematic undercount of the population residing in non-responding households.

The problem in this sort of imputation is that the Latino immigrant households that do respond to the census are smaller than those that <u>fail</u> to respond, as well as the fact that the average San Joaquin Valley non-immigrant household is much smaller than the non-responding Latino households. The error introduced through hot-deck imputation will depend on whether a nearby responding

Latino household is chosen as the donor household or whether a non-Latino household is chosen.

The San Joaquin Valley Census Research survey shows that the average household size for the non-responding Latino immigrant households is 4.6 persons. This contrasts sharply with overall average household size in the San Joaquin Valley region of 3.24 persons. The relative size of the non-responding and the responding Latino households varies by legal and citizenship status.

Model Coefficients for Estimating the Impact of Hot-Deck Imputation from Relying on Responding Latino Immigrant Household as the Information Source for Imputing Size and Characteristics of Non-Responding Latino Immigrant Households

The cascade model incorporates estimates of errors introduced by hot-deck imputation where a responding Latino immigrant household is the donor for a non-responding household by comparing the average household size of those willing to respond to Census 2020 with the citizenship question and those unwilling to respond based on analysis from the San Joaquin Valley Census Research survey.

- The Latino U.S.-born second-generation immigrant households not willing to respond are 0.67 persons larger than the responsive ones.
- The naturalized citizen headed households not willing to respond are 0.6 persons smaller than the responsive ones.
- The households of legal residents not willing to respond are 0.15 persons larger than the responsive ones.
- Finally, the households headed by an undocumented immigrant not willing to respond are .45 persons larger than the responsive ones.

Estimating the Impact of Hot-Deck Imputation from Relying on Average San Joaquin Valley Households for Imputing the Size and Characteristics of Non-Responding Latino Immigrant Households

The San Joaquin Valley Census Research Project survey data shows that about one-third of the survey respondents live in census tracts with 28% or more non-citizens, another one-third in census tracts with 20%-27% non-citizens, and the remaining third in census tracts with 20% or less non-citizens.

The cascade model assumes that the donor households for hot-deck imputation will be about a 50/50 mix of Latino immigrant and non-Latino average-sized households. In the 50% of the cases where a non-Latino household is used as the donor household for imputing the size of the non-responding Latino household, the size differential is greater.

Consequently, the model assumption that there is a 50/50 mix of Latino and non-Latino responding households as the donor pool for imputing the size of Latino immigrant non-responding households is conservative. Ongoing modeling will be done to further examine how residential patterns would affect hot-deck imputation once details of the Census Bureau's final methodology become available.

The model uses the average San Joaquin Valley household size as the estimated household size for donor households in this case: 3.2 persons per household. In contrast, the mean household size for non-responding households in the Latino survey population is 4.6 persons. Therefore, the size differential in these imputations where an "average" household is the "donor" for imputing non-responding household size is 1.4 persons per household.

Weighting Survey-based Estimates of Latino Immigrant Household Census Participation and Household Size to Project Undercount for the San Joaquin Valley Region

The cascade model weights calculated undercount rates for each sub-population of survey respondents and adjusts for survey over-sampling of undocumented immigrants and under-sampling of naturalized immigrants and U.S.-born citizens in relation to the proportion of the San Joaquin Valley population they represent.

Weights for estimating regionwide population undercount are derived by weighting the calculated undercount rate for each of the San Joaquin Valley Census Research Project survey sub-populations to the estimated proportion of each in the entire San Joaquin Valley population of potential census respondents (18+ years of age).

The weighting is based on the following estimated proportion of the actual population for each of the survey's Latino immigrant sub-population categories of legal and citizenship status: 8.5% undocumented first-generation Hispanic immigrants, 5.3% legally resident first-generation Hispanic immigrants, 6.2% foreign-born Hispanic naturalized citizens, and 14.8% second-generation

U.S.-born Hispanic immigrants. These estimates of the composition of the Latino non-citizen population are based on 2017 ACS data on the Hispanic foreign-born population 18+ years of age and estimates by the Center for Migration Studies of New York demographer, Robert Warren, of the undocumented population in the San Joaquin Valley—by county and by national origin. The estimate of the size of the naturalized Hispanic population is directly available in the ACS 2017 dataset. The estimate of the size of the Hispanic second-generation adults is estimated based on Census Bureau demographic research.

Survey and Cascade Model Implications for Undercount of Latino First- and Second-generation Immigrants throughout California

Projections of anticipated undercount among different sub-groups within the Latino immigrant population in California are an important part of public discussion and decision-making in efforts to work as effectively as possible toward a fair and accurate census. Such analysis is also important in planning strategic response to the possibility of a deeply-flawed Census 2020. The cascade model provides a basis for assessing the extent to which different facets of decennial census enumeration will contribute to undercount, the varying levels of undercount within the Latino population and geographic patterns of differential undercount.

The key technical/data analysis questions relating to the generalizability from the cascade model results for Latino first- and second-generation immigrants in the San Joaquin Valley are likely to relate primarily to the extent to which educational attainment, occupation and media exposure to news about anti-immigrant government actions, actual ICE enforcement and other contextual factors affect propensity to respond to the census.

Although there are uncertainties inherent in generalizing from the San Joaquin Valley regional survey to assess the impact of a census with the CQ on California as a state, there is some reason to undertake this endeavor, because other efforts to project undercount are already underway.[30]

30 See Eric McGhee, Sarah Bohn, and Tess Thurman for a technically sophisticated estimate of possible undercount and implications for reapportionment using a different methodology, "The 2020 Census and Political Representation in California," Public Policy Institute of California, October 2018. A critical consideration is to assess the extent to which differential undercount of immigrants will affect other states with substantial immigrant populations such as Florida, Texas and New York. Demographer Robert Warren and his colleagues at the Center for Migration Studies of New York, a well-known research and policy organization with deep expertise regarding the national distribution of undocumented immigrants and legal permanent residents, is examining this issue. Their assessment of the ways in which state and regional variations in immigrant population mix affect overall population undercount will be important in understanding California impacts.

Examination of Neighborhood Effects as a Factor in Extrapolating from the San Joaquin Valley Census Research Project Survey Data

In principle, the San Joaquin Valley findings about undercount of first- and second-generation Latino immigrants can be used as a basis for estimating statewide undercount of this large sub-group of California's Hispanic population. An issue to examine in relation to such a projection is to determine if non-response and the resulting undercount it engenders is context-sensitive, that is if there are neighborhood effects that affect levels of household response and non-response. This particular issue is discussed below.

Respondents to the San Joaquin Valley Census Research Project survey live in areas with varying density of immigrant settlement and predicted self-response (LRS).[31] Among the factors that might affect responsiveness, the study examined the willingness of undocumented, legally resident, naturalized citizen, and U.S.-born citizens to respond to the census with a citizenship question in census tracts with different concentrations of non-citizens—0-20% (low), 21-27% (medium), 28%+ (high).[32] This analysis seemed useful to assess the extent responsiveness to Census 2020 with the citizenship question might be affected not only by household characteristics, but also by neighborhood and community environment.

As part of our examination of the generalizability of the survey findings, we analyzed respondents' willingness to respond to a census with the citizenship question in relation to the proportion of non-citizens in the community and average LRS score of the tracts in the area in which they lived. We hypothesized that these variations in neighborhood context might affect willingness to respond if, as seems to be the case, there is, or if there comes to be, some discussion within social networks about deciding to respond or not respond.

Overall levels of response/non-response among the U.S.-born second-generation immigrants, the legal residents and the naturalized citizens were not significantly correlated with concentration of non-citizens in a census tract. However, undocumented immigrants' willingness to respond was lower in the communities with higher concentrations of non-citizens. In areas with higher concentration of non-citizens (>28% non-citizens in an area) only 23% of the undocumented respondents said they were willing to respond—implying an 18% eventual response rate, while in the areas with lower concentrations of

non-citizens, 38% expressed a willingness to respond—implying an eventual 33% response rate among undocumented households.

Therefore, we believe that there is a modest relationship between individuals' propensity to respond and "neighborhood" involved in, at least, undocumented individuals' propensity to respond even though households with undocumented immigrants will have a markedly lowered willingness to respond to a census with the CQ wherever they live. This finding implies that varying concentrations of non-citizens will result in somewhat deeper pockets of undercount in communities with relatively more non-citizen Latino households.

Within the San Joaquin Valley, Census 2020 undercount resulting from adding the citizenship question to the census would result in inequitable allocation of funding and political representation. The smaller rural communities with higher proportions of Latino, predominantly farmworker, immigrants would be disproportionately undercounted. There would be parallel inequities in other regions within California and across the nation.[33]

Approximations of patterns of the aggregate undercount resulting from undercount of Latino first- and second-generation immigrants in any sub-state geographic region can, therefore, be developed based on ACS or Census Bureau Planning Database data. These ACS-based estimates of non-citizens in any geographic area can then be adjusted to incorporate estimates of the proportions of Latino legal resident and undocumented immigrants within their Latino non-citizen population using analyses produced by the Center for Migration Studies of New York.[34]

---

31 The LRS (low-response score) is an updated version of the initial "hard to count" (HTC) score developed by Antonio Bruce, J. Gregory Robinson, and Monique Sanders. See Chandra Erdman and Nancy Bates, "The Low-Response Score (LRS): A Metric To Locate, Predict, and Manage Hard-to-Survey Populations," Public Opinion Quarterly, Vol. 81, #1, Spring, 2017.

32 Neighborhood concentrations of non-citizens (and low-response scores) in the areas where individual respondents lived were estimated using Census Planning Database data drawn from the ACS.

33 ACS data provides a readily-accessible way to generate rough estimates of undercount resulting from Latino immigrant non-response in different communities, counties and states since it includes information on numbers of non-citizens and naturalized citizens down to the census tract level. For example, the national distribution of the population of about 4.7 million farmworkers and dependents has been estimated using two independent methodologies (e.g. Kissam and Williams 2013 estimates for the National Legal Aid and Defenders' Association, and Aguirre International's, "Locating Migrant and Seasonal Farmworkers," report to Population Division, Census Bureau, 2007 developed as a resource for guiding outreach/census promotion in Census 2010.) Although there are regional variations in the proportion of farmworker households who are Mexican or Central American immigrants, three-quarters (73%) of U.S. farmworkers are Latino immigrants. For detailed demographic characteristics, see Table 1, NAWS National Demographics at https://www.doleta.gov/naws/pages/research/data-tables.cfm

34 The CMSNY estimates of numbers of undocumented immigrants in any geographic area are reliable to the PUMA level (in the San Joaquin Valley, therefore, available for each county). These estimates, developed by demographer Robert Warren, rely on a methodology that has evolved over the past several decades. Versions of the analysis are used also by the Pew Hispanic Institute, the Migration Policy Institute, and the Center for the Study of Immigrant Integration at the University of Southern California. The SJVCRP has used the CMSNY estimates to project the region-wide extent of undercount stemming from undercount of undocumented Latino immigrants.

It is not easy, at this point, to definitively determine, based on San Joaquin Valley Census Research Project survey data, how neighborhood effects might affect willingness to participate in a proxy interview. This analysis is not feasible simply because the overall willingness to participate in proxy interviews is so low already (8%) and because the characteristics of households contacted by enumerators as potential proxy interviewees is not known. Nonetheless, we believe that proxy interviews are likely to be more difficult to secure in neighborhoods densely settled by undocumented immigrants than in other areas.

Given these considerations, projecting geographic patterns of differential undercount is feasible—examining the distribution of the affected population (Latino immigrants) and the concentration of Latino non-citizens in any area. This analysis can be relatively easily be done using ACS data.[35]

The Projection of Statewide Impact in California

Because the cascade model provides an estimate of undercount linked to the immigration status and naturalization status of Latino immigrant heads of households, we believe that it is useful to at least consider what the implications would be for statewide undercount— since the Latino first- and second-generation immigrants make up about two-thirds of the Hispanic population in California.

Since about 39% of Californians are of Hispanic origin and about two-thirds of California Hispanics are first- or second-generation immigrants, about 26% of California's population of 40 million are first- or second-generation Hispanics. Therefore, there are slightly more than 10 million Californians in this undercounted population.

If we assume—based on the predicted 11.7% undercount of Latino immigrants in the San Joaquin Valley and acknowledge uncertainties in the model coefficients— that the statewide undercount of Hispanic first- and second-generation immigrants would likely fall into a range of 9%-13%, this level of undercount would result in an aggregate census undercount of 900,000 to 1.3 million Californians.

The eventual aggregate statewide differential undercount in California in Census 2020 would, of course, be even higher due to undercount of other Hispanic-headed households, undercount of African-American households, and undercount of American Indian households. It is likely that

there would also be a significant undercount of Asian-headed households, although the Census Bureau's PES-based analysis could not detect it. The San Joaquin Valley Census Research Project will publish in the next two months an estimate of non-response among non-Latino immigrants, primarily those of Asian origin.[36] We expect that it will be lower than that observed among the Latino first- and second-generation immigrants, but higher than would be expected from PES-based official reports of undercount.

At the level of undercount implied by the cascade model, California would be very likely to lose at least one Congressional seat just from Hispanic immigrant undercount, and more if one also considers the undercount among other hard-to-count U.S.-born and foreign-born populations in California. The corresponding fiscal loss from the Latino immigrant undercount alone would likely range from $970 million to $1.5 billion per year during the decade from 2021-2030, unless there were to be statistical adjustment for the purpose of allocating federal funding.

---

35 The impact that density of Latino immigrant settlement has in transforming non-response into undercount is probably attenuated by the fact that, on the one hand, denser settlement appears to play a modest role in decreasing propensity to respond to a census that includes the citizenship question while, on the other hand, errors stemming from systematic bias in hot-deck imputation are greater in less-densely settled neighborhoods where non-Latino households are more likely to become donors for non-responding households where proxy interviews and administrative records have failed to yield information. Modeling scenarios should be developed to explore these countervailing factors.

36 The Census Bureau's reliance on the overly-broad characterization of a range of ethnic groups as Asian is likely to have played a large role in its inability to detect undercount in this population. Given the interactions of multiple factors in determining undercount in a particular group, it is unlikely that standard coverage measurement methodology (the DSE used in PES-based estimates of undercount) would have discerned some deep pockets of undercount in different ethnic groups among Asians. For example, the Census Bureau's ethnographic research in 1990 showed significant undercount of Cambodians in Long Beach. Pamela Bunte and Rebecca Joseph reported that only 212 out of 229 Cambodians in their alternative enumeration area were enumerated (a 7.4% undercount) and that 26 who were enumerated were not identified as Cambodian. It is hoped that the SJVCRP survey samples of Hmong and Cambodian immigrants will be large enough to generate a reasonable estimate of likely undercount in these groups.

Conclusions and Implications

In communities with concentrations of Latino immigrants and low levels of self-response, increased Census Bureau reliance on untested procedures to compensate for household non-response threatens to transform the decennial census from an empirical data-gathering statistical endeavor into an exercise in unreliable statistical imputation.

In particular, efforts to rely on often-unavailable administrative records to compensate for non-response, where even apparent matching records will often be out-of-date and incomplete, is extremely problematic. Moreover high-quality matching records will not be available at the rate the Census Bureau has implied and, when available, will be systematically biased to inaccurately represent the actual numbers of people residing in a housing unit where no one responded.

A second major cause of undercount is reliance on hot-deck imputation in lieu of actual data collection. This will result in unreliable census counts and population profiles in the San Joaquin Valley region, and in other regions like it, with the extent of resulting undercount varying in relation to housing conditions in each area and socioeconomic profile of the local community.

A third major cause of undercount stems from omissions of hidden and/or unconventional housing units from the Census Bureau's Master Address File. This, too, will vary in relation to local housing condition, local economy and the quality of housing records maintained by each municipality or county.

Incorporating San Joaquin Valley Census Research Project survey findings into the cascade model framework for estimating census undercount based on patterns of non-response shows that the Department of Commerce's decision to add the citizenship question seriously degrades census accuracy for the San Joaquin Valley region, and quite probably for other regions of California and the U.S.

The cascade model developed in the San Joaquin Valley Census Research Project has practical utility for census planning because it provides a basis for going beyond current reliance on the LRS indicator of expected level of non-response/response in a census tract to include other factors that contribute to eventual undercount in operational planning and targeted census promotion.

The analysis of the dynamics through which non-response translates into undercount, based on the cascade model, underscores how important it will be for census accuracy to include attention not only to messaging and outreach focused on improving self-response, but also to give careful attention to strategies for overcoming operational causes of census undercount.

Such efforts might, for example, include the following:

- **To improve the number of hidden housing units included in census count:** Collaboration between local organizations in immigrant-dense and other low-income neighborhoods and Census Bureau address listing teams in targeting Summer 2019 "in field" address canvassing and successfully identifying the sorts of low-visibility housing units that are missed in diverse neighborhoods (e.g. converted garages in some, "back houses" in others).

- **To reduce partial household undercount in complex households:** Amplified messaging focused on the safety and importance of including non-family members in complex households on a census form. Omission of "extra persons" is not usually due to forgetting; it is intentional.

- **To reduce partial household undercount in complex households:** Partnerships to send "community navigators" out with Census Bureau enumerators to help persuade reluctant census respondents in complex households to include everyone in the household.

- **To reduce partial undercount in complex households and total household undercount in hidden housing units:** Provide additional support to assist community organizations in setting up and establishing online response centers and/or roving mobile virtual online assistance response teams to provide easily accessible help to low-literate respondents and to promote and facilitate NID response for those who did not receive an invitation or mailed census form.

- **To improve NRFU "direct interview" completion rate:** Advice to Census Bureau NRFU supervisors regarding optimal scheduling to secure response from households where most adults work.

- **To improve NRFU "direct interview" completion rate:** Census promotion focusing not simply on self-response, but also on the safety of responding to enumerators who come to visit in the course of NRFU.

- **To improve NRFU "direct interview" completion rate:** Targeted hiring of local naturalized citizens, second-generation immigrants, and local employment-authorized non-citizens with experience in community outreach enthusiastic about census response in efforts as promotore/as encouraging their neighbors who are concerned about confidentiality to respond to enumerators.

The cascade model suggests that "Get out the Count" strategies focused only on messaging and outreach and exclusively on self-response are misguided. To be effective, messaging efforts will also need to include explicit attention to getting households to respond to enumerators and helping those with digital or print literacy problems.

However, the San Joaquin Valley Census Research Project survey data, coupled with data from the Latino focus groups, suggests that efforts to increase willingness to participate in proxy interviews will be futile. Because there is an underlying resistance to providing strangers with other households' personal information and because the resistance is amplified by inclusion of the citizenship question, it is unlikely this aspect of census response behavior can be significantly changed.

Past experience does not provide sound guidance for planning census operational efforts in 2020, because outright refusals have been relatively infrequent in the past. It should be assumed they will be much higher in 2020—if the citizenship question is included and quite possibly even if it is not included (due to residual apprehension about the purpose of the census). Self-response rates will decrease dramatically, but the cascade model analysis supports experts' concerns that the Census Bureau's efforts during NRFU will not be able to compensate for greatly reduced self-response.

The Census Bureau should revise its overly-optimistic assumptions regarding likely success during operations conducted at successive stages of NRFU—in securing direct enumerator interviews, in securing proxy interviews, and in recourse to administrative records as a source of accurate information on household size. Revised estimates should be used in planning NRFU staffing in order to avoid serious disruption of data collection protocols.

It is likely that Census Bureau current estimates of non-response will not predict the serious decrease in self-response in the Latino immigrant neighborhoods. The Census Bureau should revise its current estimates of NRFU workload for these areas.

Strategic planning for Census 2020 should include vigorous efforts to not only increase respondent motivation to participate in the census, but also efforts to overcome at least some of the operational barriers in the census process that contribute to undercount. This would require institutional flexibility to craft effective partnerships between local and state census stakeholders—including local government and community-based organizations— willing to work collaboratively with the Census Bureau toward achieving an accurate decennial census in 2020.

However, given the many uncertainties involved in undertaking this sort of collaboration, which would inevitably require an unlikely Census Bureau commitment to innovation, it is unlikely that even strategic and vigorous efforts to ameliorate the impacts of the proposal to add the CQ will be successful enough to avert a serious differential undercount.

It would also be wise for California and other states with higher-than-average concentrations of Latino (and other) immigrants to prepare carefully for independent high-quality research to measure Census 2020 coverage and determine patterns of differential undercount. Such preparation would require a firm state commitment to engage in independent census coverage measurement using all the methodological tools at its disposal— demographic analysis (DA), adaptations of the triple-system estimation used in the Los Angeles TARO in 1986, coupled with ethnographic research and documentation of deep pockets of undercount, pioneered by the Census Bureau in its Census 1990 Alternative Enumeration initiative.

Appendix A: Cascade Model Estimating San Joaquin Valley Latino First- and Second-generation Undercount

The table describing the overall cascade model of undercount below has two sections. The first section reports the coefficients relating to projected errors from reliance on administrative records, under-reporting in complex households, and systematic differences between the size of households likely to respond and those unlikely to respond. The second section on the following page has assumptions and calculations regarding the numbers of households enumerated at each stage in the census process and, finally, the estimated contributions made by partial household undercount, reliance on administrative records, hot-deck imputation, and the incomplete Census Bureau address list to aggregate undercount.

Calculation and Estimation of Model Coefficients Based on HH Size	Undocumented Immigrant	Legal Resident	Naturalized Citizen	U.S.-Born Citizen
Av. HH size—responding complex HH's	5.09	5.09	5.09	5.09
Av. extra persons—responding complex HH's	2	1	1	1
Av. size –responding HH's	4.12	4.17	3.44	4.6
Av. size—non-responding HH's	4.61	4.32	4.06	5.32
Estimated % loss in HH size for complex HH that do report. Under-reporting of 80% for UNDOC, 60% for LPR, NATZ and USCIT. Av. 2 extra persons in UNDOC responding HH's and 1 extra person in LPR, NATZ, USCIT HH's	-0.31	-0.12	-0.12	-0.12
Estimated % loss in HH size due to error in size of responding HH's enumerated via administrative records. Loss of 1.5 PPH in undocumented HHs and 1 PPH in others.	-0.36	-0.24	-0.29	-0.21
Estimated # of persons loss per HH Imputed with responding Latino HH as donor	-0.49	-0.15	0.62	-0.69
Estimated # of persons loss in HH size—imputed w/ av. SJV HH size (3.24 persons)	-1.37	-0.9	-0.2	-2.11
Estimated % loss per HH Imputed w/ responding Latino HH as donor (PPH/HH size)	-0.11	-0.035	0.34	-0.13
Estimated % loss per HH—imputed w/ average SJV HH size (PPH/HH size)	-0.2973	-0.208	-0.202	-0.394

Model Components in Cascade	Undocumented Immigrant	Legal Resident	Naturalized Citizen	U.S.-Born Citizen
Universe	100%	100%	100%	100%
Housing Units In Sampling Frame (MAF) Available to Enumerate	95%	97%	97%	97%
SJVCR Self-response rate	20%	58%	65%	43%
Enumerated via Self-response (% in sampling frame X self-response rate	19.00%	56.3%	63.0%	41.7%
SJVCR Enumerator Response Rate-Direct Interviews	20%	58%	65%	43%
Enumerated via Direct Enumerator Interview (remaining HH's not enumerated by self-response X Enumerator Response Rate)	15.20%	23.63%	22.07%	23.77%
SJVCR Proxy Interview enumeration rate	8%	8%	8%	8%
Enumerated via proxy Interview (remaining HH's not enumerated by either self-response or direct enumerator interview	4.86%	1.37%	0.95%	2.52%
Assumptions re availability of "high quality" matching administrative record	30%	60%	70%	80%
HH's "Enumerated" via Admin records (% available records X not enumerated via self-response or direct enumerator interview or proxy interview)	16.78%	9.45%	7.65%	23.20%
HH's "Enumerated" via Hot-deck imputation (remainder of HH's not enumerated in earlier stages of NRFU)	44.02%	6.30%	3.28%	5.80%
Assumption- 20% of complex HH's are actually housing units missing from MAF and 80% are bona fide complex HH	18%	22%	12%	16%
Undercount from erroneous enumeration due to partial HH undercount in responding complex HH's (% complex HH's X assumed 80% non-reporting X "extra" non-family members in responding undocumented HH's), 60% in LPR, NATZ, and US-born citizen HHs	-1.11%	-1.48%	-0.92%	-0.81%
Undercount from erroneous enumeration due to out-of-date and/or inaccurate Admin. Records (Est. loss of 1 person/HH X % of HH's enumerated via admin record)	-6.11%	-2.27%	-2.22%	-4.99%
Undercount from erroneous enumeration due to bias in hot-deck imputation (SJVCR-based size of non-responding HH's vs. responding HH's and average SJV HH size assuming 50% "donor" HH's are responding Latino and 50% are "average" SJV HH's)	-8.88%	-0.77%	0.23%	-1.52%
Undercount due to total household omission from sampling frame (bad MAF)	-5%	-3%	-3%	-3%
Cumulative Undercount (Sum from multiple causes of undercount)	-21.10%	-7.51%	-5.91%	-10.32%

Appendix B: Estimate of Distribution of First- and Second-generation Latino Immigrants as % of SJV Population and Resulting Undercount

Sub-population of Latino first- and second-generation Immigrants	Undercount rate for survey sub-sample	Population share of SJV region-ACS 2017 and CMSNY	Weighting factor for sub-populations	Impact on aggregate SJV regional undercount (% undercount for group X % of total population in region)
Undocumented Immigrants	-21.1%	8.5%	0.24	-1.7945%
Legal Residents	-7.5%	5.3%	0.15	-0.3980%
Naturalized Citizens	-5.9%	6.2%	0.18	-0.3664%
U.S.-born second-generation	-10.3%	14.8%	0.43	-1.5279%
First- and second-generation Latino Immigrants	-11.7%	34.8%	—	—
Latino Immigrant Contribution to SJV Total Undercount				-4.0858%

Selected References

Aguirre Division, JBS International, "Identifying High Concentrations of Migrant and Seasonal Farmworkers," Final Report to Census Bureau, 2007.

Bhaskar, Renuka & Fernández, Leticia & Porter, Sonya. (2015). Assimilation and Coverage of the Foreign-Born Population in Administrative Records. CARRA Working Paper Series #2015-02. Center for Administrative Records Research and Applications U.S. Census Bureau, April 21, 2015.

Bhaskar, Renuka, et al, "Assimilation and coverage of the foreign-born population in administrative records," *Statistical Journal of the International Association of Official Statistics*, June 2018.

Census Bureau, *2020 Census Operational Plan: A New Design for the 2first- Century (v. 3),* September 2017.

Census Bureau, "Census Detailed Operational Plan for: 12. Internet Self-Response Operation (ISR: A New Design for the 21st Century), Decennial Census Management Division, Version 1.0, August 22, 2018.

Compton, Elizabeth, and Bently, Michael, "2010 Census Contact Strategy Non-Response Followup (NRFU) Experiment, Decennial Census Division, U.S. Census Bureau, February 2012.

Erdman, Chandra, and Bates, Nancy, "The Low-Response Score (LRS): A Metric To Locate, Predict, and Manage Hard-to-Survey Populations," *Public Opinion Quarterly*, Vol. 81, #1, Spring 2017.

Fein D.J. "The social sources of census omission: Racial and ethnic differences in omission rates in recent censuses," Ph.D. dissertation, Princeton University, 1989.

Fein D.J, West K.K. "The Sources of Census Undercount: Findings from the 1986 Los Angeles Test Census," *Survey Methodology*, 1988 Dec.

Government Accountability Office, *Lessons Learned for Locating and Counting Migrant and Seasonal Farmworkers*. 2003.

Government Accountability Office, "Additional Actions Could Strengthen Field Data Collection Efforts," GAO-17-191 January 2017.

Government Accountability Office, "2020 Census: Actions Needed to Address Challenges to Enumerating Hard-to-Count Groups," July 2018.

Government Accountability Office, "2020 Census: Additional Steps Needed to Finalize Readiness for Peak Operations," GAO Report 19-140, December 2018.

Keller, A., "Imputation Research for the 2020 Census" https://www.census.gov/content/dam/Census/library/working-papers/2015/dec/DSSD-WP2015-03.pdf

Kissam E., Herrera E., Nakamoto J. "Hispanics' Response to Census Forms and Procedures." Aguirre International, Final Report to Population Division, U.S. Census Bureau, 1993.

Kissam E. "Differential undercount of Mexican immigrant families in the U.S.," *Statistical Journal of the International Association of Official Statistics.* 2017.

Kissam, E., Quezada, C. and Intili, J. "Community-Based Canvassing to Improve the Census Bureau's Master Address File: California's Experience in LUCA 2018," *Statistical Journal of the International Association of Official Statistics*, Vol. 35, December 2018.

Kissam, E., and Jacobs, I. (2004). "Practical research strategies for Mexican indigenous communities in California seeking to assert their own identity. In J. Fox & G. Rivera-Salgado (Eds.), Indigenous Mexican migrants in the United States. San Diego, CA: Center for U.S. Mexican Studies, University of California.

Jennifer Reichert, "Findings from the 2016 and 2017 Census Tests," presentation to State Data Centers Annual Meeting, April 6, 2018.

Schwede L "Complex Households and Relationships in the Decennial Census and in Ethnographic Studies of Six Race/Ethnic Groups: Final Report", Statistical Research Division, U.S. Census Bureau 2003 Aug. https://www.census. gov/pred/www/rpts/Complex%20Households%20Final%20Report.pdf.

Trevelyan, E., et al., Characteristics of the U.S. Population by Generational Status: 2013. *Current Population Survey Reports*, P23-214. US Census Bureau, November 2016.

Turner K, Guzman L, Wildsmith E, Scott M. "The Complex and Varied Households of Low-Income Hispanic Children" National Center for Research on Hispanic Families and Children, 2015 Jan.

West, Kirsten, "Causes of Coverage Error: What Can We Learn from the Respondent," paper presented at the annual meeting of the Demographic Association, October 1985.

West K, Fein D.J. "Census Undercount: An Historical and Contemporary Sociological Issue," *Sociological Inquiry,* Vol. 60 #2, 1990.

West, K. K. and J. G. Robinson. "What Do We Know about the Undercount of Children?" U.S. Census Bureau Population Division, Working Paper No. 39, August 1999.