Respondent Validation for Non-ID Processing in the 2020 Decennial Census

Contact: Dan McMorrow — dmcmorrow@mitre.org

November 2015

JSR-15-Task-015

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 983-6997

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* November 2015	**2. REPORT TYPE** Technical	**3. DATES COVERED** *(From - To)*

4. TITLE AND SUBTITLE

Respondent Validation for Non-ID Processing in the 2020 Decennial Census

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
1315JA01

5e. TASK NUMBER
SS

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
JASON Program Office
7515 Colshire Drive
McLean, Virginia 22102

PERFORMING ORGANIZATION REPORT NUMBER

JSR-15-Task-015

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

ADC for Geographic Operations
Decennial Census Management Division
US Census Bureau
Washington, DC 20233

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

JASON's study charge is stated as follows: The Census Bureau seeks expert advice to develop methodologies to validate respondents are who they say they are when responding to online questionnaires as well as methodologies to detect and combat fraud. Ensuring that we count every person, once (but only once) and in the right location is critical to the success of the 2020 Census and the reapportionment of the House of Representatives.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Mr. K. Evan Moffett
a. REPORT UNCL	b. ABSTRACT UNCL	c. THIS PAGE UNCL	UL		**19b. TELEPHONE NUMBER** *(include area code)* 301-763-9406

CONTENTS

EXECUTIVE SUMMARY

The 2020 Decennial Census will be the first to use the Internet as the primary (and preferred) means by which data are collected from households. As currently planned, households will receive by U.S. Postal Service a postcard that provides an Internet link as a text URL and as a scannable QR code. The postcard will also have a printed, human-readable unique ID that links the postal delivery address to an entry in the Master Address File (MAF). Respondents will be asked to enter this identifier on the web form accessed by the URL. Although not strictly within its charge, JASON offers two suggestions for increasing the utility of the QR codes (see recommendations below).

Use of the Internet is not just a means for reducing cost. It also offers new opportunities for reaching previously undercounted populations and thus for increasing the accuracy of the Census. However, reliance on the Internet will expose the Census to new forms of attack. The most likely attacks will be hacker attempts to expose confidential information and/or to publicly discredit Census' operational and cyber integrity. Cybersecurity as such is not within the scope of this JASON study, but the point must be made that system design for effective cybersecurity is of paramount importance. At the same time, fallback processes must be in place for segments of the population who might fail to be counted because of computer-illiteracy or computer-hostility.

Non-ID forms (those received on the Internet without a census-provided ID) pose certain challenges. At first blush, non-ID processing should be no different in principle from what was done in the 2010 census: If a non-ID form's respondent-supplied address could be matched to a MAF entry, either automatically or with clerical intervention, then, subject to enumerator follow-up in cases of incompleteness, duplication, etc., it was accepted. Different in 2020, however, will be the Internet's lower barrier of entry to the submission of fraudulent or mischievous forms, and the existence of a "hacker mentality" that affects all things Internet. This suggests that some level of additional validation of non-ID forms is prudent, even if it goes beyond what was done in 2010. The goal is not to significantly change the rules for being counted, but only to avoid unfortunate surprises associated with the transition to the Internet.

Several distinguishable types of fraud against the census must be considered, including: hacking the Census for fun or bragging rights; social media attempts to discredit the Census and reduce cooperation; mimicry of the Census forms or apps for purposes including phishing; city or district-level attempts to change population numbers or distributions; large scale attempts to affect apportionment of the House of Representatives; individual mischief and anti-government protest. Surveying this overall threat scope, JASON finds that fraudulent non-ID returns are likely to play, at most, only a small role. With proper preparation by the Census Bureau, JASON

believes that attempts to significantly manipulate the Census by the fraudulent use of non-ID returns will be readily detectable and mitigatable.

Indeed, the greater danger in some respects is that the process of validating non-ID respondents will result in their experiencing substantially higher barriers to being counted, compared to ID respondents. Such barriers would negate the Internet's prospective ability to better reach undercounted populations. A non-ID respondent is by definition a willing respondent, because he/she has taken the trouble to seek out the Census web page; such respondents deserve to be counted.

The main text of this report uses the concept of a ROC (receiver operating characteristic) curve to illustrate the tradeoff between undercounts and erroneous counts. Without some countervailing attention, there is a danger that non-ID processing will end up on the "wrong side of the ROC curve", leading to unacceptable break-off rates and hence undercounts. In particular, JASON finds that "identity validation" as practiced in ordinary commerce is likely not a useful model for the Census, because its ratio of penalties for false positives to those of false negatives puts it in a vastly different regime from the Census.

The most important type of validation for a non-ID return is evidence supporting where a respondent *lives* (here called "Class A"). Evidence supporting who a respondent *is* adds value only to the extent that it supports the overall legitimacy of the return and thus, indirectly, the address assertion (here called "Class B"). Administrative records are useful, but they are intrinsically Class B: They allow better and more cost effective matching to the MAF, but they don't directly tie the actual web response to the claimed address. By contrast, something as simple as address uniqueness (by the end of the response period, no other returned form has claimed the same MAF entry) does provide Class A evidence, since in most scenarios it is difficult for a malfeasor to avoid creating duplications.

Mobile devices, especially if utilized via native apps (not just browser interfaces), offer important opportunities for increasing the accuracy of the Census. SMS (texting) is a universal protocol on mobile phones, and can be immediately responded to with a URL link to the census form. However, a census app has significant advantages over a web-form based dialog, even one that is mobile-optimized. From an app, respondents can be offered options that may reduce the need for a Census enumerator to visit their address, or to find them at home. Such options, which may be attractive to some respondents, include one-time use of the phone's geolocation service, uploaded photos of their house or apartment front door, and informal validation by well-known, trusted individuals in their neighborhood. The main text describes how data flows could be arranged in the latter case so as to be compatible with both user preferences and Title 13 (the statute governing census confidentiality).

JASON believes that the Census Bureau's authority to utilize imputation on individual forms, as upheld in *Utah v. Evans* (see Section 1.3), could be used to increase the accuracy of the census.

In the main text we describe a protocol by which responses that are incomplete or questionable to any discernible degree could be sorted into "pigeonholes" by objective criteria tied to well-accepted legal standards of evidence. Statistical evidence could then be used to optimally impute validity or invalidity to each pigeonhole, the criterion for optimality being the accuracy of the overall count. All returns in pigeonholes imputed valid would be counted, while all returns in pigeonholes imputed invalid would be rejected. JASON believes that a carefully constructed protocol could readily meet the three-pronged test in *Utah v. Evans*, namely "nature of the enterprise" (data collected from full population), "methodology" (deterministic, not sampling), and "immediate objective" (impute validity or non-validity to each separate return).

Although there was insufficient time to study the matter in detail, JASON expresses some concern that record linkage practice and matching algorithms seem narrowly grounded and may be out of date.

JASON offers twenty-one specific recommendations, with background described in greater detail in the main text:

Regarding ID processing:

1. QR codes on mail-out postcards should include the geocoded Census ID, so that respondent entry errors can be avoided when the QR codes are used.
2. QR codes should be sparse and quasi-random to inhibit mischievous code guessing and some kinds of fraud.

Regarding administrative records and identity validation:

3. Use of administrative records in non-ID processing should be directed towards (i) improved matching to the MAF, and (ii) statistical detection of systematic fraud; and not towards individual identity validation *per se*.
4. The possible use of commercial identity validation services must be weighed against potential undesirable outcomes. Break-off rates (with accurate denominators) should be accurately measured in tests, before approving use of these services.
5. Census should evaluate whether its quantitative understanding of the ("ROC-curve like") tradeoff between coverage and accuracy is sufficient for optimizing its policy decisions.

Regarding appropriate levels of validation for non-ID responses:

6. Non-ID forms that are matchable to valid MAF records—and with no other form claiming the same MAF record—should be accepted as valid.
7. Unusual bursts of duplicated matches (especially late in the response period), or of matches to MAF records coded as unoccupied, should be monitored as indicators of possible systematic fraud.
8. Regional (e.g., city-scale) geolocation of incoming IP addresses through public data should be done routinely and used to detect remote systematic fraud against the region or city.

9. Census should seek cooperation with the largest Internet Service Providers (ISPs) to enable a “locate me” option for respondents, yielding accuracy at the individual service-address level.

Regarding mobile devices:

10. Incoming SMS (text) should be widely publicized as a respondent point of entry to the 2020 Census.
11. Text response should be immediate and include: (i) link to the Census app installation page, (ii) smart-phone tappable “long” URL that preserves context (user text address for possible follow-up), and (iii) a simple, default URL for keyboard entry.
12. Census App users should be provided the option (appropriately framed) to use their phone’s location services to validate their address.
13. Census App users should be provided the option of taking and uploading a location photo (“housie”) to help validate their address.

Regarding fraud detection and mitigation generally:

14. Census should develop plans for the timely detection and mitigation of Internet-based fraud during the 2020 Census, utilizing space and time correlations, national and local situational awareness monitoring, data from the returned forms, and metadata from the form collection process.
15. Threat/fraud detection and response should be part of all tests leading to 2020 and there should be concurrent red-team operations to explore plausible challenges to these methods.

Regarding better matching algorithms:

16. Census should seek to improve its record-linkage and matching algorithms by the use of more modern techniques, including those of machine learning.
17. Census should seek to engage a broader academic community in its problems, especially the computer science community. There is still time to do this before the 2020 Census.
18. Census should publish non-Title 13 test datasets that can be used by the broader academic community to develop new methods of potential utility to the census.
19. Within the Census Bureau, attention should be given to allocating sufficient personnel and resources to applied research activities in support of the 2020 Census.

Regarding possible new directions:

20. Census might consider using third-party assistive services (e.g., address completion, or vetted unsworn volunteers) to help validate non-ID returns, engineering the data flows so that this mode of validation, with respondent

permission, occurs securely, but prior to the respondent's data coming under Title 13 protection.

21. Census should investigate the feasibility of categorizing returns by standards of evidence that map to their probability of being fraudulent, and then imputing validity deterministically (not by sampling) to a statistically optimized set of categories. Such a protocol would likely reduce the cost of follow-ups while improving accuracy.

1. INTRODUCTION

1.1 Background

Article I, Section 2 of the United States Constitution directs that a state's representation in the House of Representatives be determined by an "actual enumeration" of all persons in the state, and that such an enumeration be performed at least every ten years: thus the *decennial* census. The term *persons* is different from the term *citizens* (as used elsewhere in the Constitution) and is less restrictive. The goal of the 2020 Decennial Census is thus to count every person, exactly once, on April 1, 2020, by the geographical location within the U.S. where they "live and sleep most of the time" (or a similar formulation). The total number of people thus counted is expected to be about 335 million.

Over time, the decennial census has acquired, by statute and judicial order, uses in addition to the apportionment of the House of Representatives, many of which require that persons be geolocated with precision better than the state in which they are counted. The Warren Court's "one man, one vote" decision (Reynolds v. Sims, 377 U.S. 533, 1964), for example, constrained by population how the boundaries of state legislative districts can be drawn, while the Voting Rights Act of 1965 (PL 89-110) added constraints based on the racial composition of districts—implying a Constitutional need for the census to collect demographic data beyond simple counts.

Unlike counts in many other countries, the U.S. census is therefore a location-centric count of the U.S. population. A person who cannot be geolocated to where they "live and sleep most of the time" is a person who cannot be fully counted for many of the purposes of the decennial census.

The census is governed by Title 13 of the United States Code. In particular, 13 USC §9 mandates the confidentiality of private information such as names, addresses (including GPS coordinates), Social Security Numbers, and telephone numbers. While such confidentiality is essential to the purpose of the census, it also imposes limitations on how the Bureau may utilize services of third parties; this is further discussed in Sections 5.4 and 5.6.

1.2 Geographical Tagging

For the 1810 decennial census, Congress required that the count be performed by "actual inquiry at every dwelling-house, or of the head of every family within each district, and not otherwise". Similar requirements were in place for the next 14 decennial censuses. The 1960 decennial was the first mail-out census. Residents in every known housing unit received a census form in the mail and were asked to complete it. They were not allowed to mail the form back, however, because the law still required enumerators to "visit personally each dwelling house in his subdivision" in order to obtain "every item of information" (68 Stat. 2015). Thus the forms were collected door-to-door by enumerators. Apart from satisfying a statutory requirement, this process can be viewed as the “1960 method” by which completed census forms—lists of persons, essentially—were geotagged to a location in a manner deemed sufficiently reliable.

In 1964, Congress repealed the part of the Census Act that required in all cases personal visits by enumerators (78 Stat. 737). This change enabled the 1970 census to make use of the U.S. Mail not just to mail out census forms but also to receive back completed forms. The mail-back response rate was 78%. Enumerator visits were limited to follow-ups of cases in which forms were not returned, responses were defective, or housing units did not receive mail.

Thus it was the 1970 census that first raised the question of how to match a mail-returned form to a geolocation. The solution adopted was to code every census form delivered by mail with an identifier uniquely linked to its delivery postal address. Although this approach resulted in a somewhat weaker chain of custody than the previous system—forms might be delivered to the wrong address, or residents in principle could mischievously exchange forms—this geotagging method was continued without controversy in the next four decennial censuses, including the most recent 2010 census.

The adoption of this system—geotagging by documentary link to postal delivery address—also provided the impetus to standardize a national register of addresses based on all U.S. Postal Service delivery routes, augmented by other data sources so as to include all housing units. Where necessary, these addresses were checked by Census field employees. First used for the 1970 census, this register evolved to today’s Master Address File (MAF), geocoordinated by the companion Topologically Integrated Geographic Encoding and Referencing (TIGER) system. At a high level, a decennial census population count can be described as (i) a validation of the MAF/TIGER data base, followed by (ii) the assignment of a positive integer (or zero) to each MAF/TIGER entry that most accurately represents its number of inhabitants.

In the 1980 census, forms were mailed to households encompassing about 95% of the population, with a 75% response rate. In the years following, while census forms were mailed or hand delivered to 99% of the population, response rates began to fall. In the 1990 census, the mail-back response rate fell to 65%, necessitating a correspondingly greater number of costly non-response follow-ups (NRFUs). An aggressive paid advertising campaign in 2000 attempted to counter the possibility of a further drop and a 67% mail-back rate was achieved. However, the 2010 mail-back response rate again fell, to about 63%.

1.3 Sampling vs. Imputation

Another important piece of background to this JASON study is the special meaning given to the terms “sampling” and “imputation” in the context of the U.S. Census. In 1976, Congress in effect mandated the use of statistical sampling—largely as a cost-saving measure—for some purposes, amending the Census Act (13 USC §195) to read, in part:

> Except for the determination of population for purposes of apportionment of Representatives in Congress among the several States, the Secretary shall, if he considers it feasible, authorize the use of the statistical method known as "sampling" in carrying out the provisions of this title.

Notwithstanding the "except for…" language, the plan announced for the 2000 census proposed to use statistical sampling in two ways. First, for NRFUs, only a statistical sample of non-responding households would be visited for follow-up, enough to allow a statistically accurate estimate of their contribution to the count. This policy would substantially reduce costs. Second, a sample of 750,000 households (from both mail returns and personal visits) would be closely scrutinized for quality control purposes, resulting in statistically robust estimates of missed and duplicated persons. These estimates would be used to adjust, that is, "correct", the census counts. This process, it was believed, would help to alleviate the known problem of undercounts of some populations.

In short order, lawsuits seeking to forbid this use of statistical sampling were filed by four counties of Virginia, residents of 13 states, and the House of Representatives itself. In January, 1999, the U.S. Supreme Court ruled (Department of Commerce et al. v. United States House of Representatives et al., 98-404) that the Census Act prohibited the proposed use of statistical sampling. Holding that the statutory language was clear, the Court avoided ruling on whether, by its very nature, statistical sampling intrinsically conflicts with the phrase "actual enumeration" in the Constitution. It is unlikely that Congress will change the statute in the foreseeable future, so the Court's prohibitions against sampling are effectively permanent.

Another issue, with a different outcome, also arose from the 2000 census. When the number of occupants in a housing unit (a MAF entry) cannot be directly determined, despite one or more visits by an enumerator and when the unit appears occupied, yet no one answers the door—then the enumator will "impute" (colloquially, estimate) the number of occupants in the unit. This number is then counted in the census, along with the vastly larger number of better-determined occupancy counts. More technically, three levels of imputation may occur: status imputation (whether a unit exists), occupancy imputation (whether it is occupied), and household size imputation. Imputation is based on the observable characteristics of a housing unit in the context of similar, nearby housing units; it may in some cases include information supplied by neighbors. Separate from this "count imputation" is the imputation of missing responses to individual census questions, which can be either via other information provided ("assignment"), inferred from the responses of another person in the same household ("allocation"), or, failing these options, the "substitution" of data from a representative but different household nearby.

In the 2000 census, about 0.4% of the U.S. population was imputed. In some states, the rate of imputation deviated from this average, with consequences that may have affected the outcome of the apportionment of the House of Representatives. Utah, which added only 0.2% of its population through imputation, lost one seat to North Carolina by only 80 people, leading it to sue to invalidate the use of imputation on the grounds that it was a type of statistical sampling.

In Utah vs. Evans (01-714), however, the U.S. Supreme Court upheld the validity of imputation, relying in part on a picturesque metaphor worth quoting in full because it will inform parts of the discussion of non-ID processing, below.

> Imagine a librarian who wishes to determine the total number of books in a library. If the librarian finds a statistically sound way to select a sample (*e. g.,* the books contained on every 10th shelf) and if the librarian then uses a statistically sound method of extrapolating from the part to the whole (*e. g.,* multiplying by 10), then the librarian has determined the total number of books by using the statistical method known as "sampling." If, however, the librarian simply tries to count every book one by one, the librarian has not used sampling. Nor does the latter process suddenly become "sampling" simply because the librarian, finding empty shelf spaces, "imputes" to that empty shelf space the number of books (currently in use) that likely filled them—not even if the librarian goes about the imputation process in a rather technical way, say, by measuring the size of nearby books and dividing the length of each empty shelf space by a number representing the average size of nearby books on the same shelf.

The Court distinguished sampling from imputation by three tests that, while they may sound peculiar to statisticians and other scientists, are now the law of the land. 1. Sampling collects data from only part of the population, while imputation collects or imputes data for the full population (termed *nature of the enterprise*). 2. Sampling uses random "sample selection techniques", while imputation doesn't (termed *methodology*). 3. The objective of sampling is to extrapolate results to the full population, while the objective of imputation is to fill in only single records at a time (termed *immediate objective*).

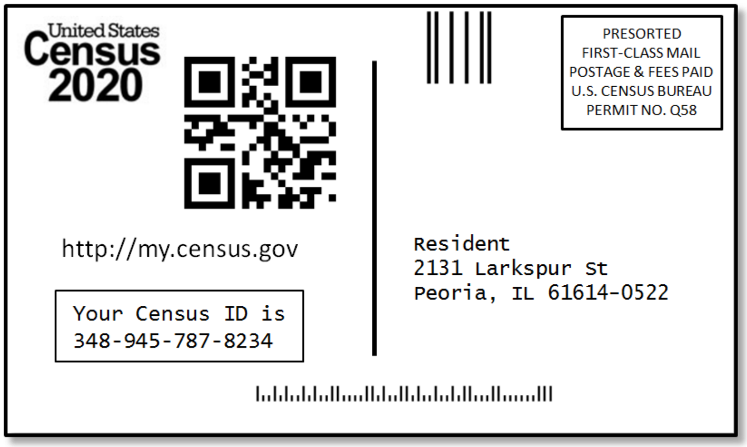
The above language, validating the use of imputation "even if [the imputer] goes about the imputation process in a rather technical way", is relevant to this study.

2. THE 2020 CENSUS PLAN AND CHARGE TO JASON

2.1 Baseline Plan for the 2020 Census

As briefed to JASON, the 2020 census will be the first census to use the Internet as the primary (and preferred) means by which data are collected from households.[1] Instead of receiving a full census form in the mail, households will receive—still by U.S. Mail—a postcard that provides an Internet link both as a text URL and as a scannable QR code (see Figure 1) that can be read by most mobile devices. The postcard will also have a printed, human-readable unique identifier (or ID, e.g., a long number or alphanumeric) that can be entered by the respondent onto an electronic or paper census form.

The unique identifier on the postcard plays exactly the same role as the unique identifier on the mailed census forms in the censuses of 2010 and before: it is a documentary link geotagging the postcard to a postal delivery address and thus to an entry in the MAF. In the intended base case, householders with Internet access (on a computer or a mobile device) will be directed to a web form that is the Internet equivalent of a paper census form. The respondent will be asked to enter their unique ID onto the form. From that point, the process is not different conceptually from mail-returned census forms, except that it can be made more user-friendly in a number of ways enabled by a computer screen, and with a cost of processing that is potentially much lower than processing paper forms.



United States
Census
2020

http://my.census.gov

Your Census ID is
348-945-787-8234

PRESORTED
FIRST-CLASS MAIL
POSTAGE & FEES PAID
U.S. CENSUS BUREAU
PERMIT NO. Q58

Resident
2131 Larkspur St
Peoria, IL 61614-0522

Figure 1. Notional version of postcard to be mailed to MAF postal delivery addresses, highlighting three elements: the census URL, the unique ID, and the scannable QR code.

[1] The 2000 census was actually the first to allow responses by Internet, with 70,000 households choosing to use this option.

Although not strictly within the study charge, JASON offers suggestions about this part of the process. First, JASON encourages the use of thoughtfully designed QR codes. Many respondents will want to use mobile devices, and most mobile devices are able to scan these codes and bring up a web page. Quite modest “Version 3 (29 x 29)” QR codes encode up to 35 alphanumeric characters with the highest level of error correction (see Figure 2). This information is more than enough to encode both the form URL and the unique ID. In other words, the QR code can just as easily encode the string “http://my.census.gov” as the string “http://my.census.gov/?3489457878234”. Neither is visible to the user, but the latter can lead directly to a census form in which the unique ID has been automatically filled in, or even better, where the MAF address is filled in for confirmation and the user never need enter (or mis-enter) the unique ID directly at all.

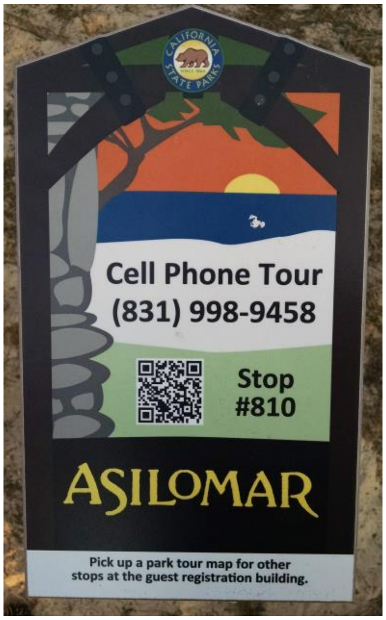

CALIFORNIA STATE PARKS
Cell Phone Tour
(831) 998-9458
Stop
#810
ASILOMAR
Pick up a park tour map for other
stops at the guest registration building.

Figure 2. QR codes can contain both a URL and ancillary data. The QR code above encodes the string “http://myoncell.mobi/8319989458/810?qr=qr”, uniquely identifying a sign at one particular geographical location. (Try it!)

Second, JASON suggests that unique IDs on the postcard (both human readable and within the QR code) should be designed so that valid entries are a sparse, and not reverse-engineerable, subset of all possible entries that a human might enter. If properly designed, this strategy can both allow for some degree of error detection and correction, and also make it difficult for a malicious user to concoct an ID that is valid. There are standard coding techniques available for accomplishing this algorithmically; or (e.g., conceptually) one could simply choose a subset of

10^9 cryptographically random values from the population of 10^{13} thirteen-digit entries and store this subset in a table. User entries can be compared in real time to the table to determine validity and connect to a MAF record. This lookup would be no harder than the MAF lookup that is already contemplated.

2.2 Non-ID Processing

The term *non-ID processing* refers to the fate of census forms that are received without a valid geotagging census ID. The lack of a valid census ID can occur for a variety of reasons, the most important being efforts to encourage people who did not receive (or have lost) the mailed postcards to respond by other means. The immediate goal of non-ID processing is to determine, preferably by an automated process, whether addresses submitted by respondents should be identified with MAF records—in which case, they can become part of the census count, or, alternatively, whether they represent valid addresses that should be added to the MAF.

Non-ID processing is not new. In the 2010 census, 2.9 million non-ID forms were submitted to automated processing. Of these, 1.3 million contained a respondent-supplied address, but not a census-supplied ID. The "Be Counted" program encouraged respondents to pick up blank census forms at locations across the country, complete them, and mail them in—without census-supplied IDs. Of these 1.3 million cases 53% were successfully matched by automated processing to geocoded MAF entries. Another 1.6 million returns contained address information supplied by census enumerators that was not in the MAF. These could be resolved by additions to the MAF, followed by redelivery of new, now ID coded, forms. Of these, 5% were successfully geocoded but not matched to the MAF. Cases failing the automated process were sent for much more expensive manual processing; about three-quarters of these were eventually geocoded and thus counted.

Successful processing of non-ID forms is desirable because it is considered an effective method for reducing undercounts, notably of populations that are hard to reach reliably by U.S. Mail, including underserved, low-income, or so-called "thin-file" residents. Such populations are geographically distributed non-uniformly, implying that successful non-ID processing will increase the accuracy of the census for many of its intended uses, ranging from apportionment to local allocation of resources.

2.3 Charge to JASON

The Census Bureau projects that 62% of self-responders in 2020 will elect to answer the Census questionnaire using the Internet. Of these, an unknown number, but certainly millions, will be non-ID responses. With this in mind, JASON's study charge is stated as follows:

> The Census Bureau seeks expert advice to develop methodologies to validate respondents are who they say they are when responding to online questionnaires as well as methodologies to

> detect and combat fraud. Ensuring that we count every person, once (but only once) and in the right location is critical to the success of the 2020 Census and the reapportionment of the House of Representatives.

This charge was developed through discussions between JASON and Census Bureau personnel (see Appendix A). These discussions aided JASON's understanding of the term "validate" in the above charge. The goal of validation is to ensure that use of the new, web-based technologies does not open the doors to systematic fraud that might significantly reduce the accuracy of the census. The desired outcome is that the cost-saving use of the Internet, along with new efforts that the Internet may enable to reach undercounted populations, should produce a 2020 census no less accurate (and, if possible, more accurate) than the 2010 census. Identity validation is a tool for reducing undercounts while paying due attention to the prevention of systematic fraud and avoiding the possibility of unexpected technology surprises in the transition to an Internet-based census.

2.4 Attitudes towards Government Collection of Data

The 2020 Census will take place against a background of long-term increasing distrust of government, and also increased awareness of the government's collection of data on the population. Illustrating the consistency of these trends over long periods, Figure 3 shows the frequency of occurrence in the Google Ngram corpus from 1950 to 2000 of the overtly normative phrase "mistrust of government" (combined with also "don't trust the government"), and also the possibly neutral phrases "census form" and "government databases".

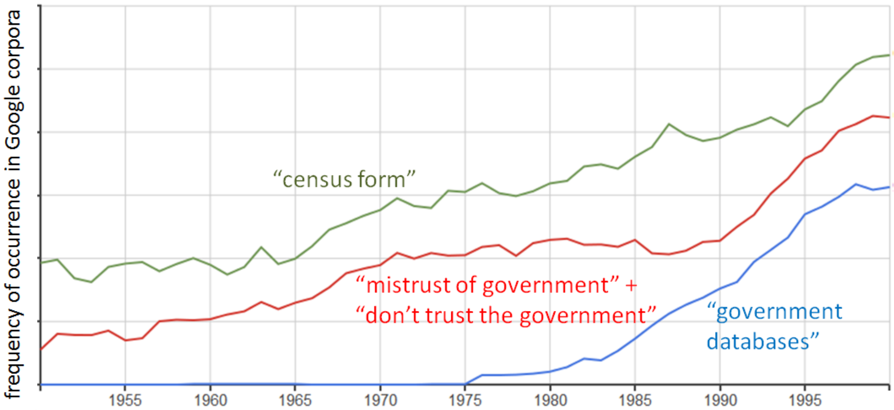


Figure 3. Google Ngram results (1950-2000) for some search terms that may be indicative of distrust of government.

While this uncontrolled sample (all books and printed media scanned by Google Books, arranged by their indicated copyright date) may not be entirely reliable, it is suggestive of the effort that will need to be made in the 2020 Census if all segments of the population, some with deep mistrust of government, are to be accurately counted.

3. DETECTION AND MITIGATION OF FRAUD

Fraudulent census responses can be defined as responses that aim to deceive or mislead. While non-organized fraud from random individuals (e.g., pet names listed as family members) is unlikely to have any significant impact on the outcomes of the U.S. Census, the same is not true for organized or large-scale fraud. It is also worth noting that, occasionally, small numbers of census responses determine the loss or gain of seats in the U.S. House of Representatives. For example, in the 2000 U.S. Census, Utah fell only 80 persons short of gaining a congressional seat, which was instead allocated to North Carolina.

Some of the new census practices proposed above provide new opportunities to create fraudulent responses. It is therefore useful to describe a taxonomy of potential types of fraud and to identify ways to detect and mitigate fraudulent responses.

3.1 Types of Fraud

History suggests it is impossible to accurately predict all possible types of fraud in a massively large-scale process, but JASON believes that most examples of fraud in the 2020 Census will fit one or more of the following classifications, not all of which are mutually exclusive:

1) The most dangerous trouble-makers for an Internet-based census may prove to be hackers, breaking into the system for fun or bragging rights, rather than any political or economic motivation. The biggest single danger is the discovery of a *scalable* exploit, since word of it will spread in the hacker community within hours, so that reaction by Census must be practiced and swift. A serious invasion by hackers could damage the credibility of the census, even if it does not affect its accuracy. Cybersecurity of the whole enterprise is therefore of the highest priority. The more complicated the system, the more opportunities there will be for hackers to disrupt and undermine it. Simplicity, with a strong, explicit security model, is likely to be a virtue.

2) Social media campaigns aimed at creating widespread mistrust in the census process are possible threats, resulting in a lower response rate in targeted geographic regions or social networks. The increasing importance of social media as a source of news and opinion will make this type of approach increasingly dangerous to the integrity of the census. It is also possible that simple misinformation, rather than intentional deception, could lead to a viral piece of social media that negatively impacts response rates or accuracy. Note that this threat occurs completely outside of the cyber domain of the census itself.

3) Mimicry of census apps, websites, or paper forms by third parties. The purpose may be criminal, as pretending to be a census worker, caller, or website, to phish for social security numbers, bank account information, and so forth. Alternatively, the purpose may be to affect the

accuracy of the census. People may be lured to false websites, or may receive email or other communication indicating that they have been counted when in fact they have not been.

4) City or district-level attempts to change population numbers or distributions, for example by adding people to housing units, including vacant and non-existent ones. “Be counted” campaigns organized with good intent by employers or civic leaders could go viral in unexpected ways that might encourage fraudulent filings on a significant scale. Even pre-Internet, organized campaigns to hand out census forms on the street, without safeguards to ensure the legitimacy or non-duplicative nature of the responses, have historically occurred. The Internet will make this easier.

5) National scale attempts to manipulate the apportionment of the House of Representatives, by selective efforts to increase the response of legitimate returns in some areas (in itself a lawful activity) while discrediting and reducing counts in other areas (an activity involving fraud). A region flooded with fraudulent returns or beset with operational problems might be forced to tighten standards or reduce follow-up visits to the point of inducing an undercount (see discussion of ROC curves in Section 4.3).

6) Individual mischief, for example, a response from Seymour Butts of 6 E. Psycho Path. This type of fraud is generally easy to detect and historically of small enough scale to not significantly impact census outcomes.

3.2 Detecting and Reacting to Fraud

Sections 4 and 5, below, touch on many specifics that will be useful in the detection and mitigation of fraud, but some preliminary remarks are offered here.

In general terms, a system for detecting fraud has sensors (algorithmic or human) and defined processes for what actions should be taken in response to various kinds of sensor detections. A system that is too insensitive, with not enough sensors or high barriers to useful responses, will leave real fraud undetected or unmitigated. A system that is too readily triggered , with sensors of questionable validity or ill-defined procedures for response, will raise too many false alarms. Depending on the politics of the situation, false alarms can be more harmful than undetected fraud. False alarms can be exploited by trouble-makers to destroy public confidence in the census.

Approaches to mitigating fraud require its detection on a timescale compatible with amelioration. The most likely source of extensive fraud, especially in the era of the online census, is at least partially automated. For example, a program pretending to be a person could repeatedly visit the census web site or use the census mobile app to submit a form containing misinformation. The nature of website or mobile app usage can inform methods to detect this type of repetitive electronic fraud. When a browser connects to a web site, the web site has available the IP address of the browser, plus header information that the browser sends. This header contains information useful to a web server, such as what browser it is, what kind of responses it accepts, and what fonts it knows about. For example, the web site panopticlick.eff.org will estimate the uniqueness of this information. Pilot experiments conducted by JASON suggest that desktop or mobile browser headers provide ~20 bits of distinguishable information, meaning that browsers chosen at random have only about a one in a million chance of sending the same header information.

The IP address, the exact time of the interaction, and the browser header information together provide the basis for detecting some kinds of extensive fraud. The most unsophisticated type of repetitive electronic fraud would send many different interactions in short succession from the same browser and IP address and could be easily detected based on this pattern. Sessions from libraries and other common service points would use the same IP address and browser, but would be spaced out over time, and further constrained by operating hours, if those are known. Although the browser header information is under the control of the fraudster, he/she has to fake the information in a way that avoids detection. The use of a browser that is too unusual would raise suspicion, especially if it comes from many different IP addresses in the same geographic area. The use of too many different browsers from the same IP address should also raise red flags. The use of a seemingly identical browser from different IP addresses also would be suspicious because, on the whole, different browsers do not send the same header information. Further, there are other ways of fingerprinting browsers. One of these, for HTML5 browsers, is canvas fingerprinting (see, for example, https://www.browserleaks.com/canvas).

Ideally, the Census Bureau would create a set of rules to decide whether to accept or reject interactions based on IP address, time, and browser header information. It is not necessary to be certain of these rules in advance. Information can be processed the day after it is collected, or week by week, to determine which statistical patterns appear to be anomalous and worth further investigation as potential fraud.

If there is a census app (see Section 5.5.2, below), it might be implemented so that its back end looks like a browser to the Census web site, or might connect to a different Census site and use a different protocol to transfer data. In either case, the app should transmit data containing the same kind of open information about its environment that a mobile browser would send. This practice would facilitate the identification of fraudulent misuses of the mobile app.

Fraud perpetrated through social media (for example, category 5 above), should be easy to detect—with appropriate monitoring— because, by their nature, social media campaigns are designed to spread quickly and freely among a population. Real-time monitoring of popular social media sources is an important activity so as to identify fraud. It should be at least partially automatable by searching for keywords and social network patterns suggestive of coordinated attempts to perpetrate census fraud. Analogously, web-crawling and "app store crawling" for fraudulent websites and apps that share cosmetic characteristics with the bona fide census website or app should be performed in a continuous, automated manner.

It is also possible to detect fraud through analyses that are not dependent on examining electronic footprints. For example, as further described in Section 5.2, below, a geographically localized and unexpectedly high number of non-unique returns, or a burst of non-unique returns near the end of the response period, can point to systematic fraud.

The Census Bureau should establish expected process behavior and anomaly detection for every proposed (and existing) census practice that is vulnerable to fraud, which automatically includes any online or mobile methodology. Exercises including “red teams” trying to perpetrate fraud and evade detection should also become part of the process of developing and implementing new practices, including the online and mobile interfaces discussed above. Red teams should be employed in all tests leading up to the 2020 Decennial Census. It is important to note that cybersecurity experts, especially those involved in the preparation of the 2020 Decennial’s defensive measures, are *not* effective red team participants, for two reasons: (i) Their participation in the specific cybersecurity defense inevitably means that they will not recognize their own attackable “blind spots”. (ii) Effective red team attackers have a different mentality and training than do cybersecurity experts. Red team expertise exists in government (e.g., NSA), in non-governmental Federally Funded Research and Development Centers (FFRDCs, e.g., Sandia National Laboratories), and in industry.

Combatting census fraud can take place either before or after the fraud is committed. A savvy social engineering campaign that explains the purpose and importance of the census, and that helps build a culture of trust about using census forms, websites, and apps, can decrease the likelihood that some of the types of fraud described above have been initiated or will be successful. Of course, JASON anticipates that some fraudulent activities, including large-scale, organized fraud for political gain, is unlikely to be thwarted by such a campaign, and therefore post-fraud countermeasures must also be prepared.

Legal practices that are presumably already planned to halt census fraud, including trademarking the census logo and seeking injunctions to rapidly shut down fraudulent websites or apps pretending to be bona fide census respondent portals, are obvious post-detection countermeasures. But JASON also suggests that census practices be adjusted on a locality-by-locality, day-by-day basis, in response to the detection of fraud. For example, the Census Bureau should consider temporarily strengthening the threshold for legitimacy needed to trigger the acceptance of website or app submissions if systematic fraud that would be rejected by these higher thresholds has been detected. Of course, the negative impact of fraud must be weighed against the negative impact of altering these thresholds, as described by the ROC and tradeoff discussions below in Sections 4.3 and 4.4.

4. IDENTITY AND ADMINISTRATIVE RECORDS

4.1 Address Matching and Geocoding is the Goal

An identity is validated by linking it with some degree of assurance to a previously established chain of identity (see Appendix B). In a related manner, a completed census form can be viewed as asserting links between one address (more specifically, a housing unit) and the identity of one or more persons, the inhabitants of that unit. Figure 4 illustrates the concept for both ID and non-ID cases. When the form also includes a census-provided ID, it is linked to a MAF entry. Non-ID returned forms are formally identical, except that they lack the pre-existing link to a MAF address.

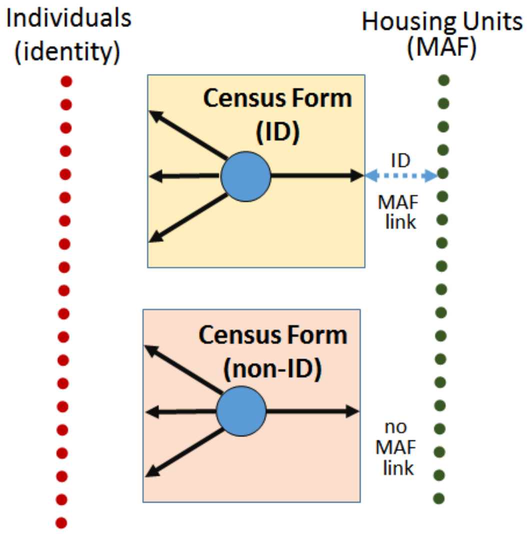


Figure 4. A returned census form asserts links between people (nodes on the left) and an address (node on the right). Forms with census-supplied IDs have validated links to the Master Address File (MAF), while non-ID forms have no such link.

When a form's assertion as to an address (arrow pointing right in Figure 4) is readily matchable to a MAF record that is in turn geocodable, then—absent any indicators that it might be fraudulent—it can be accepted for census tabulation. For some fraction of non-ID forms received, however, the address match, or the validity of the return, may be unclear. For these, it is generally understood that there needs to be some additional corroborating evidence that the form was in fact returned from the indicated address—enough evidence at least to detect, and thus deter, the exploitation of non-ID forms for systematic fraud. It is important to recognize that, with regard to mischief that is *not* systematically different between ID and non-ID forms—

for example, listing the family pet as an inhabitant—the two types of forms need not be treated differently.

A form's assertions as to the inhabitants can be regarded as pointers to the verifiable identities of individuals (arrows pointing left in Figure 4). By contrast with addresses, *no* U.S. census has attempted the wholesale validation of these links, for example, by connecting the form's assertions to verified identities such as Social Security numbers or birth records. Nor is it plausible that the 2020 Census would attempt to do this, especially given the increasing sensitivity of the public to government databases, as well as the Constitutional mandate to count all resident persons, including those who may be difficult to document. Thus, in the Figure, links to the left and links to the right are not subject to equivalent verification.

Legitimate non-ID returns signify not a failure of the respondent, but a failure of the Census Bureau to have reached a willing respondent with an ID-coded form. It follows that the purpose of identity validation in non-ID processing (in the Figure, connecting arrows to the left) is to impute a MAF connection on the right. While this sounds like an obvious point, it has operational consequences. In particular, it implies that non-ID processing, taken as a system, should not be biased in such a way as to reject forms whose asserted identities (arrows to the left in the Figure) are *no worse* than what is deemed acceptable for ID-coded forms. One might refer to this undesired outcome by the phrase *non-ID bias burden.* The minimization of non-ID bias burden should be a design goal for non-ID processing.

4.2 Use of Administrative Records

Administrative records, or "adrecs", are databases available to the Census Bureau that link names and addresses (and sometimes other information about identity) independently from the returned ID and non-ID census forms. By law, Census has access to IRS records, including Forms 1040, 1040a and 1040-EZ (in principle all with income over a low minimum), W-2 (in principle all wage earners), 1099s (in principle all pensioners, recipients of government aid, dividends, interest and certain other payments), and W-4s (which include declared numbers of dependents). Depending on interagency negotiations, Census *may* have access to records from other government agencies, and it may buy administrative records on the open market.

Adrecs can be regarded as links "around the back" of the bipartite graph that was shown in Figure 4. Figure 5 illustrates this perspective. An adrec may not be uniquely matchable to an individual on a non-ID census form by name alone (e.g., the case of "John Smith"), nor to an address alone (e.g., when its address encompasses more than one housing unit). But the comparison of its data, taken together, to the data on the census form, also taken together, may provide a unique match, from which a MAF link can be corroborated or inferred *ab initio*.

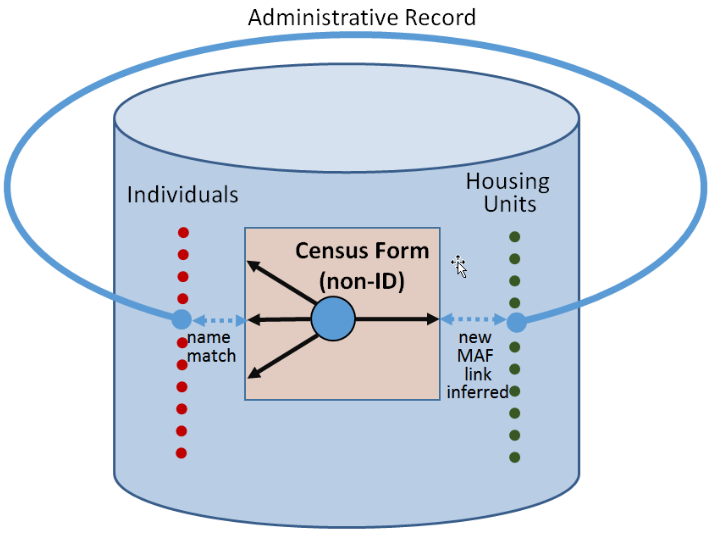


Figure 5. Administrative records such as tax forms can be pictured as providing additional, confirmatory links between names and addresses, “behind the back” of the census form.

Any inhabitant (any leftward arrow in the Figure) can serve as the reference person for this use, not just the responding householder. The purpose of the adrec match is not to verify the identity of any particular person, but only to provide the necessary degree of additional assurance that the address pointed to by the census form is validly MAF-linked. It is worth noting that an adrec match, by itself, does not directly corroborate the address from which a non-ID census form actually originated. For example, a family who has moved recently, but (for whatever reason) returns a non-ID form pointing to their previous address, may well be corroborated by adrecs that are not completely up to date.

The value of adrecs, other than for enabling matches to the MAF that could otherwise not be made at all, is thus less direct: an adrec match increases confidence in the legitimacy of a returned non-ID form. Evidence *of any kind* that a returned non-ID form is legitimate (i.e., that it was completed in good faith by a householder) is *weak* evidence that all the data in the form, especially address, are accurate.

The value of the weak evidence, such as that provided by adrecs, is not that it perfectly validates any particular address assignment, but rather that it can, in bulk, detect (and thus deter) systematic fraud. Adrec corroboration is statistical, with both false positives (the mischievous family who has moved) and false negatives (no corroborating record for a legitimate return). The discussion below considers how variable patterns of adrec matches in space and time may be used to identify systematic fraud.

4.3 Commercial Identity Validation Services and ROC Curves

The use of the Internet for collecting census data—especially the ability to process data in real-time so that real-time interactions with the respondent become possible—opens up many useful opportunities. It is tempting to think that the use of commercial identity validation services is one such opportunity. As a means of linking a respondent to an address, he or she might seamlessly be transferred online to a commercial entity such as a credit bureau (Equifax, Experian, TransUnion), a bankcard consortium (Visa, Mastercard), or a comprehensive social media company (Google, Microsoft, Facebook), where the respondent's identity might be verified and an address returned. Unfortunately, JASON is highly skeptical that system designs of this kind will be useful for the census, first, because they intrinsically carry a large non-ID bias burden (as defined above), making it harder for historically undercounted populations to be counted; and second, because the requirements of the census intrinsically put them in a regime that is uncharted and unreliable territory for commercial identity-services providers, as will now be explained.

Identity validation is a type of binary classification: Either a person *is* who he says he is, or else he *isn't*. All binary classifiers involve a tradeoff between false positives and false negatives. This tradeoff is often formalized as a ROC curve.[2] Figure 6 shows a notional ROC curve (not real data) as it might apply to the census. The horizontal axis shows the number of bad forms accepted, as a percentage of all forms, ideally close to zero. The vertical axis represents the percentage of legitimate forms accepted, ideally close to 100%. The ideal census therefore is described by a 100% vertical value, and a 0% horizontal value in Figure 6.

[2] ROC is technically an acronym for “receiver operating characteristic”, which is mainly an historical artifact but still applies literally to radar and other physical systems.

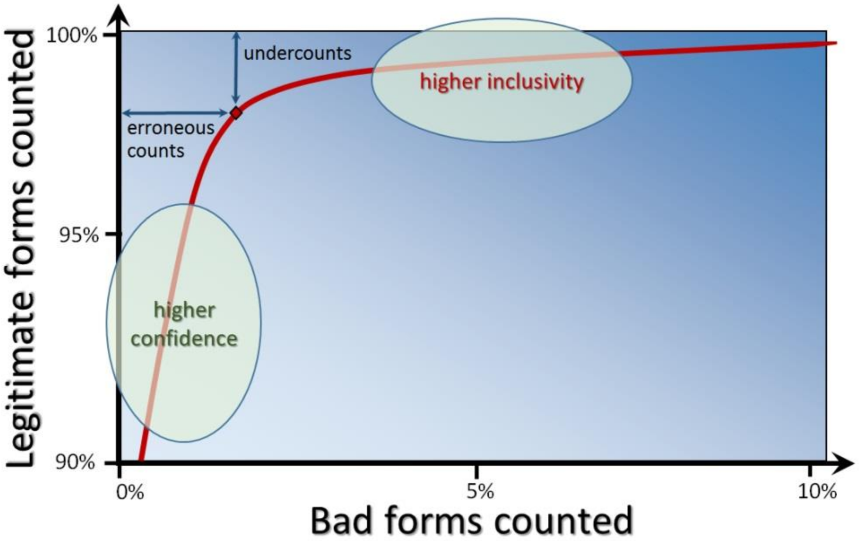


Figure 6. ROC tradeoff curve. Any one-parameter procedure for deciding which non-ID forms to accept and which to reject defines an ROC; the user may choose any point on that curve by choice of that parameter, exchanging confidence (fewer erroneous counts) for inclusivity (fewer valid missed counts).

A given procedure for processing non-ID forms implies a point on the ROC curve. Each point on the curve represents a number of undercounts (false negatives) and a number of erroneous counts (false positives). A change in process that rejects more questionable forms moves the point to the left along the curve, towards higher confidence in each counted form. That change ordinarily will also reject additional valid forms and increase the undercount, lowering the position on the curve in Figure 6. A change that accepts more legitimate forms moves the point the opposite way, towards higher inclusivity of legitmate forms, but usually with a greater number of erroneouscounts.

In the case of almost all commercial applications of identity verification, the cost of a false positive (for example, allowing the hacker access to empty a victim's bank account) is very much larger than the cost of a false negative (for example, telling the customer to call the bank and provide additional verification before the account can be accessed). As a result of this asymmetry, commercial identity verification adopts practices described by the far left of the ROC curve. This regime is the region of the curve in which commercial providers are competent. Here, they can be expected to understand the absolute and relative utility of the various techniques at their disposal, such as two-factor authorization, knowledge-based authentication, and so forth. This example is also one in which customers will tolerate a significant amount of respondent burden, as they genuinely need to access their bank accounts and recognize the importance of verifying their identities before access can be granted.

By contrast, the goals of the census are best served by practices in the opposite regime of the ROC curve. Respondent burden is a primary concern. It may take very little in the way of delays on the web form, or intrusiveness or time burden of questions asked (e.g., "what was the amount of your last mortgage payment?") to cause a respondent to break off a session and contribute to an undercount. Furthermore, the regrets of letting through a small number of bad forms—especially if they not part of any concerted scheme to skew census results—may be much smaller than the regrets of systematically undercounting a known elusive population. Thus, census's needs drive the optimal tradeoff towards the right on the ROC curve.

It is best practice in identity authentication to distinguish four levels of assurance. The *lowest* of these, termed "LOA 1", corresponds roughly to the familiar username/password protocol, connecting a user with what is usually only a self-asserted identity, that is, asserted at the time that the username/password was first registered. This is not relevant to the census because the returned non-ID forms are themselves self-asserted identities. Higher levels of assurance, however, require correspondingly more complex and intrusive identification protocols. For these higher LOAs, both suspicion of government intrusiveness and respondent burden are likely to drive high break-off rates (i.e., individuals with intent to respond but who become discouraged and stop before completion) and therefore significant undercounts.

JASON thus believes that commercially available identity validation services would prove to be of very limited value in the 2020 Census and does not recommend their use. This judgment could, if necessary, be validated or disproved empirically by testing well before 2020. Such testing, however, would need to include measurement of the "denominator" in the break-off rate, that is, the number of unique individuals with a *first intent* to fill out the census questionnaire. JASON is unsure how such a measurement could be made because it would have to account for word-of-mouth spreading of a negative reputation—for example, that the non-ID form is too difficult for respondents to get through—to people who would then never even present at the web interface.

4.4 Comparison to Legal Standards of Evidence

Legal standards of evidence furnish a different lens through which one can view the census non-ID problem and suggest an optimal point on the ROC curve. Government, which has the power to inflict "harms" of varying degrees on its citizens, is required to meet greater burdens of proof before inflicting greater harms. Table 1 gives a skeleton summary of some required standards of evidence that are recognized in various aspects of U.S. law.

Table 1.

	Burden-of-Proof for government to inflict "harm" on a citizen	Examples of "harm"	Proposed analogy to census
low	"reasonable suspicion"	brief police stop	failing to count a (random) individual
	"reasonable to believe"	vehicle search	
	"probable cause"	arrest, indictment	
	"some credible evidence"	temporary injunction	
	"substantial evidence"	adverse administrative ruling	systematic undercount of an identifiable sub-population
	"preponderance of the evidence" "more probable than not"	civil damages, permanent injunction	
	"clear and convincing evidence"	*habeas corpus*, child custody	
high	"beyond a reasonable doubt" "no plausible reason to believe otherwise"	criminal conviction	

A resident of the U.S. who submits a legitimate non-ID census form is "harmed" if that form is not counted in the census. Therefore, the government (Census Bureau) should meet some burden of proof before rejecting a seemingly well-intentioned form. What that standard should be depends on how great the harm of not being counted is considered to be. One can argue that little or no harm is inflicted by failing to count a single, random, individual, because errors of this type have little or no statistical effect on census results and they are evenly distributed across all groups of interest. In contrast, the *systematic* undercount of an identifiable population of individuals can inflict a significant harm on all the members of the group, for example depriving them of political representation or government assistance. This is indicated somewhat notionally in the last column of Table 1, where systematic undercounting is placed at around the same level as an adverse ruling by a government agency or the imposition of civil damages. By this standard, such forms should be rejected only when it is "more probable than not" that they are invalid. This level of test is identified with the point on the ROC curve with a 45 degree slope (roughly the point shown in Figure 6).

However, because the consequences of random rejected forms are so different from those of systematic undercounts of some populations, one might additionally argue that there is no "universally correct" tradeoff point on the ROC curve as regards non-ID returns. By this way of thinking, the tradeoff between undercounts and erroneous counts should depend on (i) the degree to which undercounts may fall disproportionately on an identifiable subpopulation (in which case the tradeoff point in Figure 6 should move to the right), counterbalanced by (ii) the degree to which erroneous counts can be shown statistically to represent attempts at systematic fraud (in which case the tradeoff point in Figure 6 should move to the left). Both (i) and (ii) will likely vary geographically, and for attempts at fraud likely also in time as participation in a fraudulent scheme peaks and falls off.

It is important, therefore, to develop methods for estimating both undercounts and fraud attempts, as a function of space and time, both in real time during the census and also in postprocessing before decisions on which non-ID returns to count are finalized.

Section 6 will discuss how a hierarchy of standards of evidence, like the one discussed here, might be employed in an allowable imputation technique that could increase the accuracy of the census.

4.5 Consider the Poor SSN!

JASON appreciates that, although it might legally be allowed to do so, the Census Bureau is highly resistant to the idea of asking anyone for their Social Security or Tax Identification number. It should be noted, however, that the public has become accustomed to the use of "the last four digits of your social security number" as confirmatory evidence of identity, although these common uses lie in the private sector and do not raise the specter of government data tracking.

JASON suggests that the Census Bureau study the tradeoffs of implementing a policy such as the following:

> If you would like to enter the last *two digits* of your social security number, please do so here. This will help us to ensure that no one is able to use your name and address in filing a fraudulent census form.

Even this small amount of voluntary specificity, combined with matching to adrecs, especially IRS forms, would be likely to catch the great majority of attempts at naïve identity theft on census forms.

5. OTHER FORMS OF CORROBORATING EVIDENCE

5.1 Purpose of Corroboration and Classes of Corroborating Evidence

Administrative records are a form of corroborating evidence. In general, corroborating evidence has three distinct uses. First, it can be used to assist in matching and geocoding otherwise unmatchable non-ID returns, yielding fewer undercounts. Second, used statistically, corroboration may detect systematic fraud attempts and reveal their specific signatures, applicable to a particular geographical regions and spans of time. Such signatures can then be used to identify, for follow-up, individual returns that are potentially fraudulent. Third, related to the first two, the use of multiple forms of corroboration increases overall confidence in the accuracy and integrity of the census. In this regard, corroboration—even when not strictly necessary from an operational perspective—is a kind of due diligence.

For a non-ID census form, there are two classes of corroborating evidence. Evidence may directly support the assertion that the address provided by the householder is actually where he or she lives. One can refer to this kind of evidence as "Class A". Alternatively, evidence may support the legitimacy of the form in general, thus supporting the address assertion, but only indirectly. This kind of evidence can be referred to as "Class B". Class A evidence is more precious and harder to come by than Class B evidence.

The census-provided, geotagged ID is Class A evidence, because that ID was delivered by a U.S. Postal Service worker to a known address with high confidence. The evidence provided by administrative records is more ambiguous. For example, although IRS Form 1040 asks the taxpayer to enter a "home address", people with transient lifestyles often enter the address of a parent, relative, or friend who collects their mail. Administrative-record evidence will not by itself identify cases where a respondent has, perhaps for convenience and not out of duplicity, responded on a non-ID census questionnaire in similar manner. Administrative records have evidentiary value, but, as defined here, the evidence lies in Class B.

5.2 Address-uniqueness

Address-uniqueness is an important property. When two conditions hold: (i) the address asserted by the respondent on a non-ID return is sufficiently complete that it can be matched unambiguously to a MAF record and geocoded; and (ii) *no other* return (ID or non-ID) claims that same MAF record; then a return is generally includable in the census tabulation. When a MAF record is claimed by more than one return, then there is a reconciliation process, first automatically by the co-called Primary Selection Algorithm and then (where necessary) by clerical labor and possible follow-up visit.

Address-uniqueness is Class A evidence. It directly supports the hypothesis that the non-ID form claiming the MAF address has been provided honestly. Mathematically, one would say that it is the result of the accumulation of a large number of Bayes factors, one factor from every respondent *not* claiming that address. Although no single factor is more than very weak evidence, their large number accumulates to significant evidence in support of the non-ID form.

While address-uniqueness can be determined only at the end of the response period, *non-uniqueness* occurs whenever a second form is assigned to a particular MAF record. Non-uniqueness can be an important indicator of fraud or other systematic problems. For this reason, workflows and processing should be designed so that instances of non-uniqueness are immediately flagged and analyzed statistically in near-real time, soon enough to trigger compensatory responses in the field.

To avoid creating non-unique returns, a malefactor would need to know either a significant number of unoccupied units that are coded in the MAF as being occupied, or (nearing the end of the response period) the response status of a significant set of MAF records. Imperfections in the attempts of such an individual would produce the easily detected signal of a population of MAF records with two responses, often one ID and the other non-ID.   A geographically localized and unexpectedly high number of non-unique returns, or a burst of non-unique returns near the end of the response period, is thus a signature of systematic fraud. Scrutiny of the non-ID duplicates would presumably reveal common characteristics that would identify the fraud.

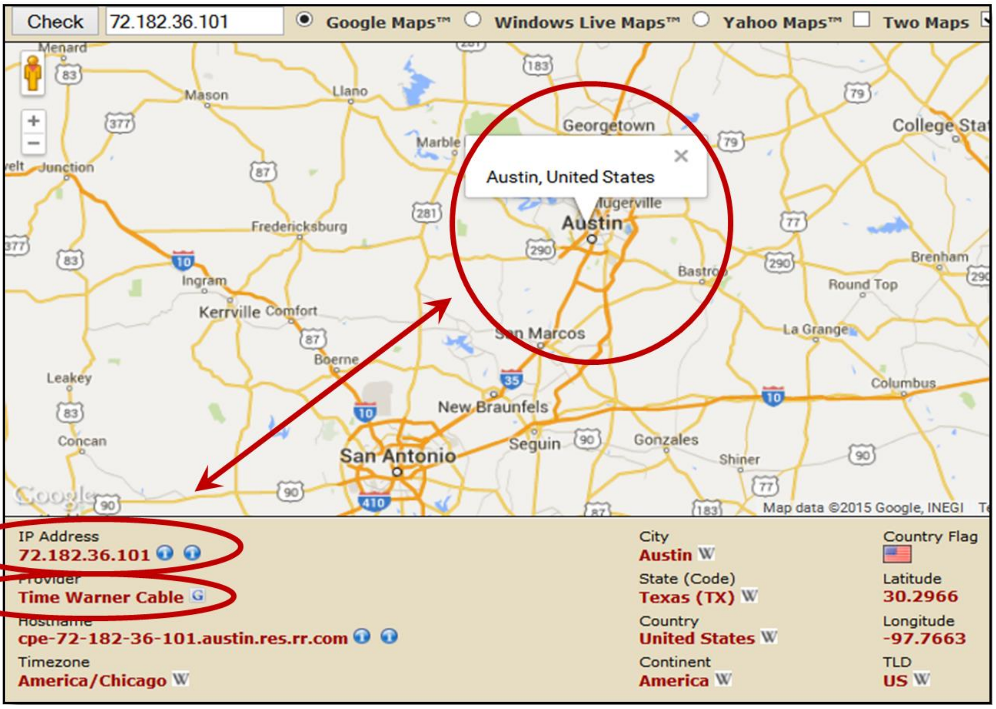
5.3 Match to the 2010 Census

Non-ID responses in the 2020 Census may match, in both name and address, records in the 2010 Census. If the 2010 record was an ID response (with census-provided geotag), or if it was a record directly verified by an enumerator, then a matching 2020 non-ID response should be considered to have Class A evidence supporting it, with the presumption that it should be counted.

It might be argued that this counting rule opens the census to a "graveyard vote". It seems implausible, however, that this could be exploited systematically without being easily detected, as above, by the duplication of non-ID responses and the MAF records of ID returns.

5.4 Internet localization

All Internet responses automatically arrive with IP addresses—their packets' "return address". Many non-ID respondents will have Internet access in their homes and will be responding through that access. In such cases, the arriving IP address will usually, although not always, be physically associated with the respondent's geographical address. If Census can learn that association, and if the geographical address matches the address provided by the respondent on the non-ID response, it is strong Class A validating evidence.



72.182.36.101

IP Address
72.182.36.101

Provider
Time Warner Cable

Hostname
cpe-72-182-36-101.austin.res.rr.com

Timezone
America/Chicago

City
Austin

State (Code)
Texas (TX)

Country
United States

Continent
America

Country Flag

Latitude
30.2966

Longitude
-97.7663

TLD
US

Figure 7. Example of publicly available data that links an IP address (as would be automatically associated with a returned census form) to an Internet Service Provider (here, Time Warner) and a city (here, Austin).

Publicly available databases associate IP address ranges to geographical location, but they do so only with typical resolution as large as a city, sometimes a state (see Figure 7). Public data also associate every IP address to an Internet Service Provider. By itself, city-level localization may be useful for some purposes. For example, an organized national campaign to submit fraudulent census forms would likely produce an easily statistically detectable signature of excess responses whose city of origin (as determined by IP address) did not match their claimed residential addresses.

Much more useful, however, would be the mapping from IP address to geographical address ("service address"). This is known in detail, at any given time, only by the respondent's Internet Service Provider (ISP, see Figure 8).

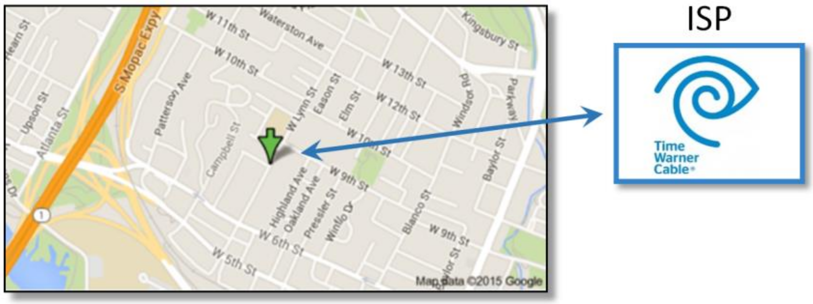


Figure 8. The Internet Service Provider (ISP) knows, at any given time, the exact mapping of an IP address to a geographical “service address”.

There are thousands of small ISPs, but the biggest ones (Comcast, AT&T, Time-Warner, Verizon, etc.) account for the large majority of residential Internet connections. It would be worth considerable effort to convince as many as possible of the large ISPs that a contractual relationship with the 2020 Census is in the interest of their customers and themselves. Because the protection of individual privacy is a strong driver, an implementation scenario for this approach might be structured as follows:

- A respondent initiates a non-ID form on the web.
- Census identifies by IP range that the respondent is, for example, a Comcast subscriber, hence knows what message to display. Census attaches the IP address to its record of the form being submitted as a temporary additional identifier.
- The respondent enters a residential address.
- The page where the address is entered has two action buttons. One says: “I am a Comcast subscriber and request Comcast to validate the above address for the Census. This will make my return easier to complete and helps ensure that it is properly counted.” Text explains that Comcast will see *only the address exactly as entered*, not any other information from the census form, not even the name that the respondent has entered on the form. The other, alternative action says, “I prefer to send nothing to Comcast. Send the above address only to the U.S. Census.”

When the responder chooses the Comcast action link, the following things happen:

- In real time, an SSL-encrypted message containing the text of the entered address (and nothing else) goes to Comcast. This message goes directly from the user’s machine to Comcast, without going through Census.
- In real time, Comcast records the address string and notes the timestamp and the IP address from which it was received.
- The respondent finishes completing the census form. There is no need to wait for further action by Comcast.

- On its own schedule (possibly by batch) Comcast associates the IP address with its own service address record and compares it (possibly using an algorithm supplied to it by Census) to the address string supplied by the respondent.
- Comcast periodically transmits securely to Census a set of records, each containing (i) an IP address and timestamp, and (ii) an address string exactly as entered by the respondent (i.e., not Comcast's own service address record), together with a Yes/No indicator of whether there was an address match.
- Census matches the IP address from Comcast to the IP address that it previously associated with a non-ID form and thus attaches to that record the submitted address and the verification result. It can now delete the IP address from the record.

This proposed implementation, shown schematically in Figure 9, is structured so that no Title 13 information ever leaves Census, and no Comcast customer records ever leave Comcast, hence the requirement that only the address string exactly as entered by the respondent be transmitted unidirectionally from respondent to Comcast to Census.

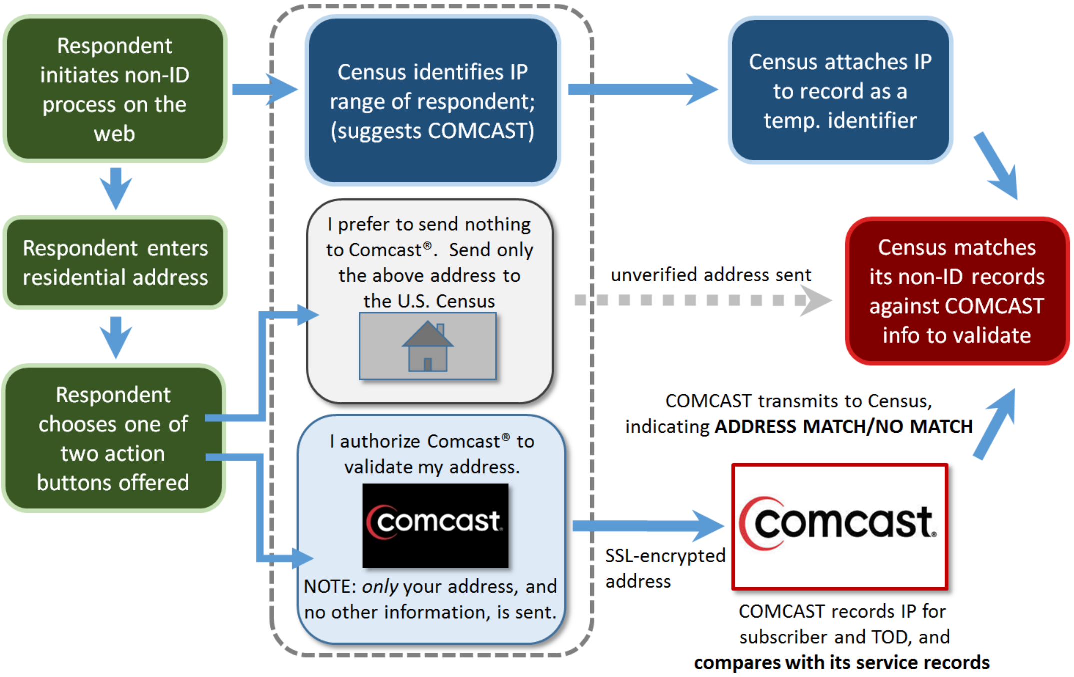

Respondent initiates non-ID process on the web
Respondent enters residential address
Respondent chooses one of two action buttons offered
Census identifies IP range of respondent; (suggests COMCAST)
I prefer to send nothing to Comcast®. Send only the above address to the U.S. Census
I authorize Comcast® to validate my address.
comcast
NOTE: *only* your address, and no other information, is sent.
Census attaches IP to record as a temp. identifier
unverified address sent
Census matches its non-ID records against COMCAST info to validate
COMCAST transmits to Census, indicating **ADDRESS MATCH/NO MATCH**
SSL-encrypted address
comcast
COMCAST records IP for subscriber and TOD, and **compares with its service records**

Figure 9. Data flow for using a cooperating ISP to validate the physical address associated with an IP address, structured so as to be consistent with Title 13 and customer confidentiality requirements.

5.5 Mobile Device Scenarios

The prospect of communications via mobile devices is one of the most exciting new features of the 2020 Census. A minimal use, for smart devices with Internet access, is to use the mobile device simply as a web browser, showing pages that are optimized for its screen size and type of device. Such use will result in both ID and non-ID responses, depending on whether the respondent has a mailed ID postcard. Equally interesting, however, are interactions between respondent and Census that utilize other features of mobile devices. Four such features are discussed below: SMS (Short Message Service, i.e., "texting"), GPS geolocation, cameras and picture uploading, and the ability to download and install apps.

5.5.1 Short Message Service

SMS is one of the oldest add-ons to voice in mobile telephony, dating back to the 1980s. SMS texts travel not via the Internet, but directly on the cellular infrastructure. As a result, almost all cell phones, both "smart" and "dumb", are capable of sending and receiving SMS texts.

In a possible scenario, LeBron James looks into the television camera at half-time, holds up a cell phone, and says: "I've been counted in the Census. If *you* haven't, text 'Count Me' to 7777." The result is some tens of thousands of "Count Me" text messages received. The message itself is irrelevant. What is relevant is that every text comes with a text-back (usually also a voice call-back) number, and that text-backs can be automatically processed in real time.

When text messages contain web URLs, smartphones convert them to live links. Thus the simplest response to each text would be to text back the link to the (mobile optimized) census web form. If the phone is dumb, the user will be able to refer to the link later, when he/she has access to the Internet. For smartphones (which will be the vast majority by 2020), the user can touch the link and immediately be directed to the form, hopefully producing a complete response. Such responses will, however, all be non-ID.

5.5.2 Apps

In addition to, or instead of, the link to the form, a texted URL might point to the installation page of an app, typically in either the Apple Store or Android Play Store. ("Thanks for responding! Be Counted! Download the app [link], or just go to the form [link] right now.") The availability of a 2020 Census app can also be publicized independently of the texting scenario. The app can be appropriately publicized as a simple and convenient way to complete one's obligation to be counted.

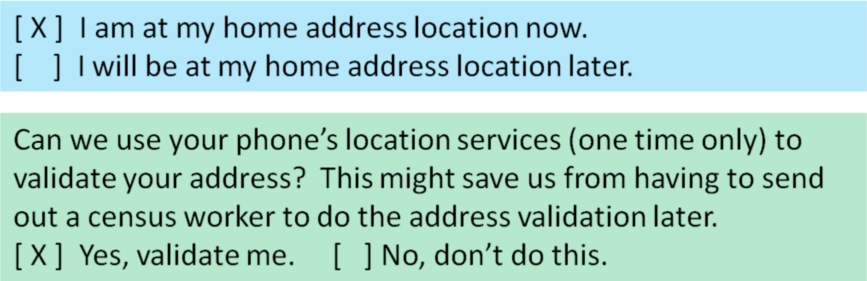
Interacting with a potential responder by means of an app, not just a browser, has several advantages:

- Browser-independence, hence no glitches caused by browser settings.
- Seamless utilization of third-party services. For example, address-completion suggestions (like the familiar drop-down lists in Google Maps) could be shown without using Census (Title 13) information.
- Ease of maintaining context between partial sessions, with no need for repeated logins.
- Association of returns to the unique IDs of mobile devices, perhaps to limit each device to a single return, or at least to flag for scrutiny cases of multiple returns from the same device
- Interaction with other features of the mobile device, such as GPS and camera (see below).
- Faster (e.g., fewer lags) and richer (e.g., more game-like) user experience. Game-like features can be targeted at younger people to raise Census awareness.
- Easier integration with other social media. With one click a respondent can urge everyone on his/her contact list to also “Be Counted”, perhaps also distributing to these contacts a one-click link to install the app.
- Can implement advanced security models (not just SSL on transmissions)

5.5.3 GPS and WiFi geolocation

All phones sold in the U.S. have GPS capability. In non-smart phones, the capability is accessible only for 911 Emergency use. In smart phones (nearly all phones by 2020) GPS can be accessed by apps. Android and Apple phones are also able to determine location without GPS in many cases, using proprietary databases that link the unique identifiers of WiFi base stations to their geographical locations as observed by “war driving” (electronic canvasses of neighborhoods from mobile vehicles).

After a respondent enters an address, a Census app might ask if the respondent is at that address now and, if so, ask for permission to upload his or her current geolocation (see Figure 10). If the respondent is not then at the census address, he/she could be asked to upload the geolocation at a later time, when they are at home, via the app and without a lengthy login process (because the app would maintain context in a cryptographically secure manner).



[X] I am at my home address location now.
[ ] I will be at my home address location later.

Can we use your phone's location services (one time only) to validate your address? This might save us from having to send out a census worker to do the address validation later.
[X] Yes, validate me. [ ] No, don't do this.

Figure 10. Notional text in Census App that requests user permission to utilize cell phone geolocation capabilities for address validation.

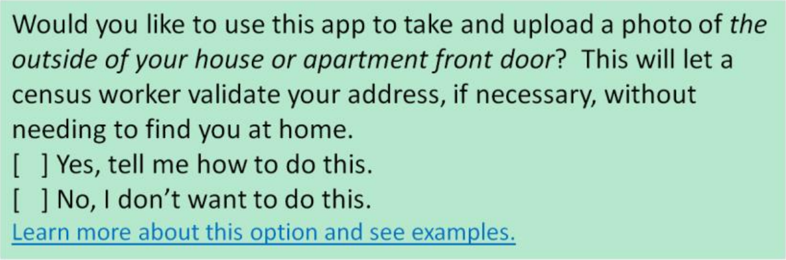
GPS or WiFi geolocation provides Class A confirmation for non-ID returns because it links the respondent to the asserted physical location (or to a location very nearby). JASON thinks that this link is comparable in strength to the link provided by the U.S. Postal Service for the traditional ID return. The two links are not exactly the same. Although a census-provided ID is delivered to a specific mailbox with higher spatial resolution than GPS and has some protection against theft by the "sanctity of the mails", there is no guarantee that the person who ultimately files a census form with that ID is, or ever was, at the physical mailbox location. GPS has lower resolution, but it confirms a true physical location of the person actually filling out the form (via the app).

JASON sees no significant risk that non-ID returns that include a matching GPS location and are uniquely associated with a single mobile device could be used effectively for any systematic fraud.

Every cell phone knows the unique identifier of the cell phone tower to which it is currently connected, and this information is generally available to apps on the phone. While this geographical resolution is not as good as GPS or WiFi (it may span more than one census tract, for example), it might be used as weaker, confirmatory evidence of location.

5.5.4 Camera

Failing to obtain a GPS or WiFi geolocation, the Census app might ask the respondent to take a picture of his/her house or apartment as seen from the street. This could be compared in manual, office-based follow-up to pictures from Google Streetview and similar services. Or, a picture of a respondent's front door might be requested, to "help us find your residence for a follow-up if we need to". (See Figure 11.) The enumerator on a follow-up visit might then only have to verify the existence of the door, not actually find the responder at home. This could reduce the cost of follow-up visits. The link between the door and the respondent will have been provided by the link between the app and the uploaded photograph.



Would you like to use this app to take and upload a photo of *the outside of your house or apartment front door*? This will let a census worker validate your address, if necessary, without needing to find you at home.
[ ] Yes, tell me how to do this.
[ ] No, I don't want to do this.
Learn more about this option and see examples.

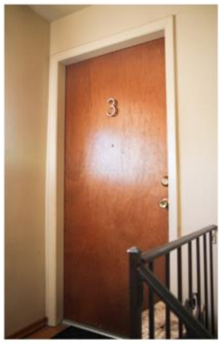


Figure 11. (Left) Notional text in the Census App that invites the respondent to upload a "housie". (Right) Notional example of such a photo, which could be used by the enumerator to verify an address without finding the residents at home.

Electric meters might be another useful target for cell-phone photography (Figure 12). As one potential voluntary alternative, a respondent could be invited to upload a picture of the electric meter associated with his/her residence. Electric meters have prominent unique identifiers, often both alphanumeric and barcoded, used by the power company to identify the subscriber. With cooperation from the utility company, these identifiers could be automatically read from the uploaded photo and matched to a service address. Even without such cooperation, a photograph of the meter could be matched by a census enumerator to the actual meter, which is often in an accessible location to facilitate old-style meter reading. This would enable validation of the geolocation of the respondent (at the time the meter photo was taken) without the necessity of direct contact with the enumerator.

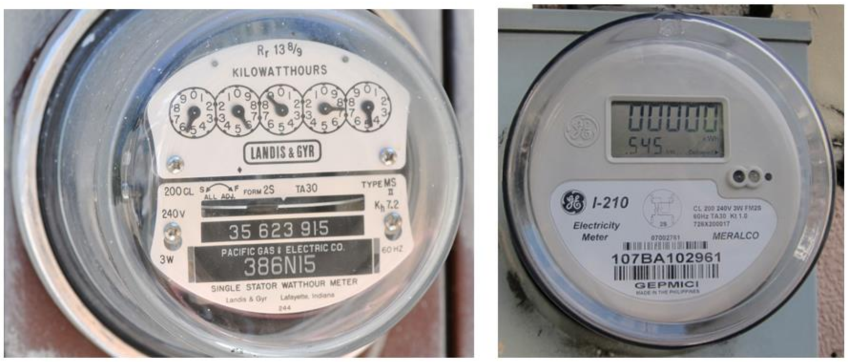


Figure 12. Analog and digital residential electric meters. Unique identifiers on the meters establish their geolocation, while meter readings establish the approximate date of a photograph.

With cooperation of the utility, the meter reading itself could be used to provide additional evidence because it can be translated to an approximate date—interpolating between billing meter readings—of the photograph. That date should correspond approximately to the date when the photo was uploaded.

5.6 Unsworn Validators

The Census Bureau distinguishes between people with access to Title 13 census data, who are "sworn", and everybody else. Census enumerators are sworn, and their determination that Person X is resident at MAF address Y is taken, for the most part, as dispositive.

JASON suggests that, aided by technology, "access to Title 13 data" and "probity in confirming addresses" might be teased apart as two separable functions. Stably resident citizens with leadership credentials and reputations for honesty exist in most communities and neighborhoods. Such people could be locally identified and engaged, not as full-time sworn enumerators during the period of intense canvassing, but as part-time "validators" during a longer period spanning the census. Validators might include social service workers, U.S. Postal Service workers whose routes are in the neighborhood, election poll watcher volunteers, librarians, clergy and lay religious leaders, community organizers, and so forth. Validators would not be sworn in the sense of Title 13, but they might be asked to attest under penalty of perjury that information that they provide is correct to the best of their knowledge.

Technology would provide efficient means for validating, and also protect and channel data flows, so that Title 13 is respected. One implemented scenario (see Figure 13) might be as follows:

- Respondent enters his/her name and address on an app. The data is stored locally in the app, not uploaded, so it does not yet come under Title 13.
- The app displays a list of names of possible validators (based on geographical proximity as determined by a third party) and asks the respondent if he/she knows any of these people and consents to sending his/her name and address (only) to that validator. ("Your permission can help us be sure that you are counted correctly in the Census.")
- If the respondent chooses one or more validators, his/her name and address are sent, not through the Census Bureau, but directly by the SMS channel, to a "validator's app" on the validator's mobile device. (Apps have the capability of intercepting SMS messages for custom processing and display.)
- The respondent need not wait for a response, but completes the form, which is then sent to Census, where it is protected by Title 13.
- At a later time of his/her choosing, the validator, using the app, communicates an opinion to Census, perhaps on a scale ranging from "personally known to me at this address", through "not known to me but entirely plausible" to "likely to be fraudulent".

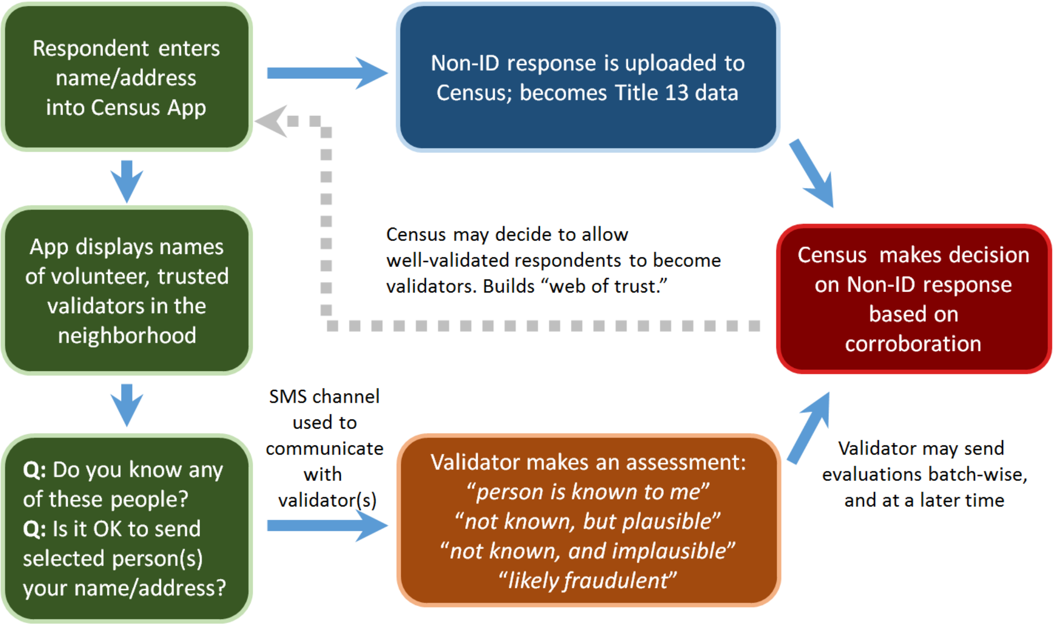

Respondent enters name/address into Census App
Non-ID response is uploaded to Census; becomes Title 13 data
App displays names of volunteer, trusted validators in the neighborhood
Census may decide to allow well-validated respondents to become validators. Builds "web of trust."
Census makes decision on Non-ID response based on corroboration
SMS channel used to communicate with validator(s)
Q: Do you know any of these people?
Q: Is it OK to send selected person(s) your name/address?
Validator makes an assessment:
"person is known to me"
"not known, but plausible"
"not known, and implausible"
"likely fraudulent"
Validator may send evaluations batch-wise, and at a later time

Figure 13. Data flow for the use of unsworn validators. With respondent permission, data is routed by the app from the respondent directly to a chosen validator,
without its becoming Title 13.

A related concept would be to allow anyone who (i) installs the Census app, (ii) submits an ID return or a sufficiently vetted non-ID return, and (iii) expresses an interest in doing so, to serve as a validator, in effect building up "webs of trust" from ID and non-ID respondents. Such webs can be constructed automatically, without expensive on-the-ground recruiting, from the returns data. Criteria could be established as to which webs of trust, based on their size and their collective degree of independent vetting, should be taken as sufficient validation of non-ID returns. This might be viewed as a better-grounded and more rigorous version of imputation compared to simply asking a neighbor.

6. PROBABILISTIC ENUMERATION WITHOUT SAMPLING OR CORRECTIONS

In a hypothetical world in which all U.S. residents were Bayesian statisticians, the census would be quite different. Census results would be recognized as estimated quantities and statistical estimators would be chosen so as to have an agreed-upon set of tradeoffs among desirable properties (e.g., “unbiased”, “minimum variance”, “efficient”, “consistent”). The effect of a single census return on the reported count would be more nuanced than simply incrementing a ledger count by one. It might have the effect of incrementing a count by rather less than one (e.g., if the return has a substantial probability of being fraudulent), or by greater than one (e.g., if it is representative of populations known to be undersampled). Because a return might probabilistically shed light on the validity of other returns (making individual returns not statistically independent), census estimates would not be simple sums over returns.

In the real world, it is a matter of law, as well as a matter of practice, that decennial census results be sums of individual returns, each accepted return contributing exactly the value one to its appropriate count. Much is gained by this simplicity, but some opportunities for increasing the accuracy of the census are thereby lost. JASON suggests that, with careful attention to the law of the land, not all such opportunities to benefit from Bayesian principles are lost and that non-ID processing represents a possible point of application.

A simple example of what one might wish to do is to estimate the number of valid non-ID returns (in each census block, say) by summing the estimated probability that each is valid over all returns. Doing exactly that is likely not permissible by the tests laid out by the Supreme Court. However, we might instead proceed as follows:

- Based on the degree to which it can be validated, assign each non-ID return to a recognized category of standard of evidence (see Table 1 for a list of categories), ranging from “reasonable suspicion” that it might be uncountable (e.g., duplicate, fraudulent, or insufficiently complete) to that it is uncountable “beyond reasonable doubt”. Also assign to each return a score indicating its position within each group (e.g., “top decile of the ‘substantial evidence’ group”).
- For each evidence category and subgroup, estimate over the population (by sampling and secondary validation, or any other statistically valid procedure) the probability that a return *in that subgroup* is uncountable. In other words, quantify the meaning of the group assignments.
- For each evidence category, determine the number of deciles (from the most valid end) that, *if fully enumerated*, will most closely approximate the actual number of valid returns within that evidence category (see Figure 14). For example, the evidence category “more

probable than not" might have its top 5 deciles counted; the evidence category that is fraudulent "beyond a reasonable doubt" might have none of its deciles counted.

- Enumerate those returns in the accepted deciles of their evidence category. Reject those returns in the lower, not accepted, deciles.

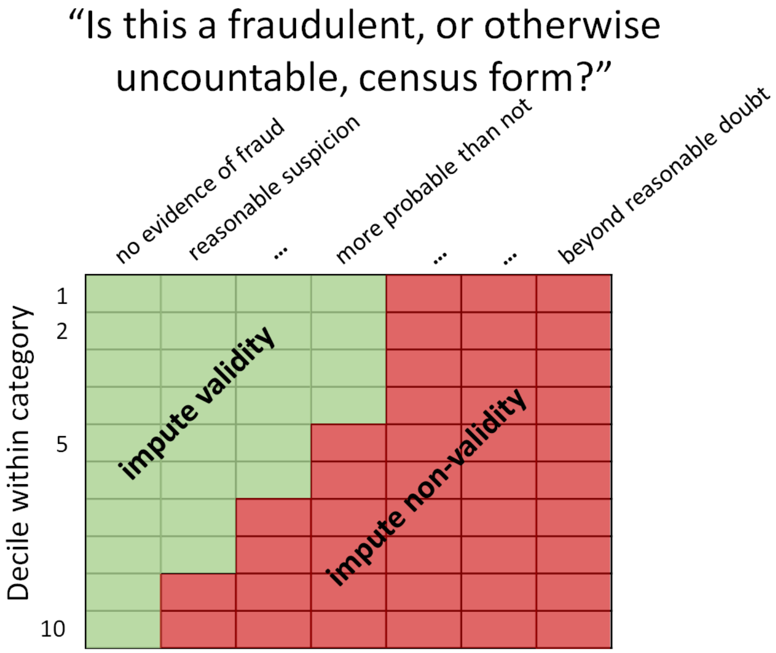


Figure 14. Census forms grouped into "pigeonholes" by objective criteria that are indicative of their validity can then be counted, or not, based on an imputation of validity or non-validity. No "sampling" is involved.

While somewhat baroque, this procedure is designed to produce statistically valid results while satisfying the constraints laid out in Department of Commerce v. United States House of Representatives and Utah v. Evans (see Section 1.3 above), for the following reasons:

1. Data are collected from the full population, not a random sample ("nature of the enterprise test").
2. The decision to accept any particular return is deterministic, not random ("methodology test"). Returns are assigned by objective criteria to recognized standard-of-evidence categories. Within each category, more-likely-valid returns are counted before less-likely-valid ones.

3. The process ultimately imputes validity or non-validity to each single record and enumerates the valid ones ("immediate objective test"). In the words of Utah v. Evans, it is allowable imputation even if it "goes about the imputation process in a rather technical way".

The sharp reader might wonder what is the need to stratify into standard-of-evidence categories at all? Why not just order all non-ID returns by their probability of being valid, and then accept (from the top of the list) a statistically determined appropriate number? There are three answers to this question.

First, the vocabulary of standards of evidence (equivalently, burdens of proof) is an established means that the law uses to deal with probabilistic events, while numerically expressed probabilities are not. Using standards of evidence as the top-level classification is likely to be grounded in firmer legal precedent and is easier to explain without numbers.

Second, the use of categories helps to distinguish logically using the side calculation (e.g., "what fraction of non-ID returns with 'substantial evidence' of fraud are in fact fraudulent?") from being the calculation of a (forbidden) sampling probability assigned to a particular return. A calculation over a whole category is more clearly a step in an imputation process, like the librarian "measuring the size of nearby books and dividing the length of each empty shelf space by a number representing the average size of nearby books on the same shelf" (Utah v. Evans, see Section 1.3 above).

Third, the rule that deciles within each category are included in order from greater to lesser validity has the objective benefit of enriching the enumerated sample with more valid returns. This approach does not affect the number of returns counted, but it will increase the statistical accuracy of conclusions based on other census questions asked, justifying that the counting order is at no stage arbitrary.

Many variations of this suggested process are possible. The point is that imputation procedures can be quantitatively designed to maximize the statistical accuracy of the count. Such procedures might greatly reduce the need for on-the-ground NRFUs by, in effect, increasing the accuracy of imputation to the point where NRFUs (other than those required to validate the imputation procedure) add no further value.

7. IMMEDIATE NEED FOR APPLIED RESEARCH

Although JASON was not briefed in detail on what resources are currently being applied to research and development (as distinct from operational planning) in support of the 2020 Census, JASON was left with a strong impression that such resources are inadequate and could usefully be increased. A good example would be the development of matching algorithms, which are crucial for making the match between the address string entered by a non-ID respondent and the canonical address as recorded in the MAF. To complete this match may require linking records from multiple files including IRS, Social Security, and the MAF. Record linkage research and methods adopted by Census tend to be grounded in the foundational work of Felligi and Sutner[3] which is based on probabilistic matching and conditional independence of comparison attribute fields.[4] JASON learned, anecdotally, that implementations such as BigMatch are holdovers from bygone eras in computing. [5]

It may be possible to improve the performance of string matching by using more modern techniques. Machine learning algorithms implementing multilayer neural nets, as well as various kinds of clustering techniques, deserve attention, as do n-grams and other index-based approaches. Census is aware of research across several fields in record linkage, data cleaning, and object identification.[6] However, the more modern techniques have not yet been sufficiently explored to be adopted by Census.

Also deserving some consideration are more global formulations of the matching problem, where one searches for the best matches not one at a time, but for the best overall set of all matches (for example in three-digit zipcode regions). Weighted bipartite matching problems (e.g., MAF nodes vs. respondent-string nodes) are well known to computer scientists, with such classical results as the “Hungarian algorithm” for the weighted assignment problem and the Gale–Shapley algorithm for the so-called “stable marriage problem”. While these do not apply directly to the

[3] Fellegi, Ivan P., and Alan B. Sunter. "A theory for record linkage." *Journal of the American Statistical Association* 64.328 (1969): 1183-1210.

[4] Winkler, William E. "Matching and record linkage." *Business survey methods* 1 (1995): 355-384.

[5] Yancey, William E. "BigMatch: a program for extracting probable matches from a large file for record linkage." *Computing* 1 (2002): 1-8.

[6] Winkler, William E. "Overview of record linkage and current research directions." *Bureau of the Census*. 2006.

needs of the Census, they suggest that discussions with computer scientists, as well as statisticians, may prove fruitful. Because time is already short for useful academic engagement, such interactions should be directed at very specific problems, such as address matching. On a longer timescale, a broader research program might be contemplated.

It would also be useful to have sample datasets, freely distributable to academic researchers without Title 13 restrictions. This would enable large numbers of groups to develop matching algorithms on an informal basis. Such test datasets might be constructed by Census from commercially purchased mailing lists, using only their address data. A releasable test problem might be to match two unrelated such lists and to compare with answers produced by the Census Bureau internally (e.g., matching both to the MAF, but not releasing any MAF data). JASON notes that, somewhat similarly, the U.S. Postal Service offers a Coding Accuracy Support System (CASS) that includes a public Stage 1 CASS file of about 150,000 addresses, each in both a standardized "correct" form and as a string, possibly garbled, that is actually seen "in the wild". USPS also offers Stage 2 services in which it "grades" submissions from commercial mailing-list companies on their matching accuracy. As an analogy, enormous progress on handwriting recognition resulted in the 1990s as the result of USPS making available a large sample of scanned, handwritten digits.[7] Competition on test problems with public data is a recognized mode by which computer science advances.

Within the Census Bureau, resources (both people and computational resources) might usefully be directed towards applied research. Again anecdotally, JASON has a sense that the talent for this already exists within the Bureau, but that the appropriate people and computer resources are now assigned to operational tasks. JASON recommends a new look at increasing resources for applied research in support of the 2020 Census by Census Bureau personnel.

[7] *A Database for Handwritten Text Recognition Research*, J. J. Hull, IEEE PAMI 16(5) 550-554, 1994.

8. FINDINGS AND RECOMMENDATIONS

JASON's findings and recommendations follow from the above discussion.

8.1 Findings

Use of the Internet Generally

1. The primary use of the Internet is not just a means for reducing cost; it also offers new opportunities for reaching previously undercounted populations and thus for increasing the accuracy of the Census.

2. Reliance on the Internet will expose the Census to new forms of attack, the most likely being hacker attempts to expose confidential information and/or to publicly discredit Census' operational and cyber integrity. System design for cybersecurity is paramount.

3. Fraudulent non-ID returns are likely to be only a small part of the overall threat.

4. With proper preparation by the Census Bureau, attempts to manipulate the Census by the fraudulent use of non-ID returns will most likely be readily manageable.

Validation of Questionaires

5. Willing respondents have a right to be counted. There is a danger that the process of validating non-ID respondents will result in their experiencing a substantially higher barrier to being counted compared to ID respondents.

6. Without some countervailing attention, there is a danger that non-ID processing will end up on the wrong side of the trade-off between accuracy and inclusiveness ("the ROC curve"), leading to an unacceptable break-off rate and hence undercount.

7. Evidence supporting where a respondent *lives* (here called "Class A") is the most valuable. Evidence supporting who a respondent *is* adds value only to the extent that it supports the overall legitimacy of the return and thus, indirectly, the address assertion (here called "Class B").

8. Class B evidence, such as Administrative Records, can add value when used statistically as part of a system for monitoring, detecting, and deterring fraud.

Data collection, data flows, and data processing

9. Mobile devices, especially if utilized via native apps (not just browser interfaces), offer important opportunities for increasing the accuracy of the Census.

10. Engineering of data flows from respondents may enable the use of third-party services or vetted, unsworn volunteers without violating Title 13. However, care must be taken to ensure security and appropriate data handling.

11. A structured approach to the imputation of validity or non-validity based on deterministic assignments ("pigeonholes") that are informed by statistical "side calculations" might allow increased accuracy with fewer NRFUs and decreased cost, without the use of sampling as defined by Utah v. Evans.

12. Record linkage practice and matching algorithms seem narrowly grounded and may be out of date.

8.2 Recommendations

ID processing:

1. QR codes on mail-out postcards should include the geocoded Census ID, so that respondent entry errors can be avoided when the QR codes are used.
2. QR codes should be sparse and quasi-random to inhibit mischievous code guessing and some kinds of fraud.

Administrative records and identity validation:

3. Use of administrative records in non-ID processing should be directed towards (i) improved matching to the MAF, and (ii) statistical detection of systematic fraud; and not towards individual identity validation *per se*.
4. The possible use of commercial identity validation services must be weighed against potential undesirable outcomes. Break-off rates (with accurate denominators) should be accurately measured in tests, before approving use of these services.
5. Census should evaluate whether its quantitative understanding of the ("ROC-curve") tradeoff between coverage and accuracy is sufficient for optimizing its policy decisions.

Appropriate levels of validation for non-ID responses:

6. Non-ID forms that are matchable to valid MAF records—and with no other form claiming the same MAF record—should be accepted as valid.
7. Unusual bursts of duplicated matches (especially late in the response period), or of matches to MAF records coded as unoccupied, should be monitored as indicators of possible systematic fraud.

8. Regional (e.g., city-scale) geolocation of incoming IP addresses through public data should be done routinely and used to detect remote systematic fraud against the region or city.
9. Census should seek cooperation with the largest Internet Service Providers (ISPs) to enable a “locate me” option for respondents, yielding accuracy at the individual service-address level.

Mobile devices:

10. Incoming SMS (text) should be widely publicized as a respondent point of entry to the 2020 Census.
11. Text response should be immediate and include: (i) link to the Census app installation page, (ii) smart-phone tappable “long” URL that preserves context (user text address for possible follow-up), and (iii) a simple, default URL for keyboard entry.
12. Census App users should be provided the option (appropriately framed) to use their phone’s location services to validate their address.
13. Census App users should be provided the option of taking and uploading a location photo (“housie”) to help validate their address.

Fraud detection and mitigation generally:

14. Census should develop plans for the timely detection and mitigation of Internet-based fraud during the 2020 Census, utilizing space and time correlations, national and local situational awareness monitoring, data from the returned forms, and metadata from the form collection process.
15. Threat/fraud detection and response should be part of all tests leading to 2020 and there should be concurrent red-team operations to explore plausible challenges to these methods.

Better matching algorithms:

16. Census should seek to improve its record-linkage and matching algorithms by the use of more modern techniques, including those of machine learning.
17. Census should seek to engage a broader academic community in its problems, especially the computer science community. There is still time to do this before the 2020 Census.
18. Census should publish non-Title 13 test datasets that can be used by the broader academic community to develop new methods of potential utility to the census.
19. Within the Census Bureau, attention should be given to allocating sufficient personnel and resources to applied research activities in support of the 2020 Census.

Possible new directions:

20. Census might consider using third-party assistive services (e.g., address completion, or vetted, unsworn volunteers) to help validate non-ID returns, engineering the data flows so that this mode of validation, with respondent permission, occurs securely, but prior to the respondent's data coming under Title 13 protection.
21. Census should investigate the feasibility of categorizing returns by standards of evidence that map to their probability of being fraudulent, and then imputing validity deterministically (not by sampling) to a statistically optimized set of categories. Such a protocol would likely reduce the cost of follow-ups while improving accuracy.

APPENDIX A: List of Briefers and Topics Briefed

Briefings were held June 24-25, 2015, at the JASON facility in San Diego, CA.

Evan Moffett (Census Bureau, Decennial Census Management Division) served as the principal Point of Contact to JASON and briefed on “2020 Census Overview”.

Adreana Able (Census Bureau, Decennial Statistical Studies Division) briefed on “Census Matching Process Overview”.

Stuart Irby and Andrea Johnson (Census Bureau, Geography Division) briefed on “Master Address File and Address Matching”).

Christa Jones (Census Bureau, Special Assistant to the Director) briefed on “Census Title 13 Background”

Jennifer Kerber (GSA) and Mike Garcia (NIST) briefed by teleconference on “Connect.gov”.

Frank McPhillips (Census Bureau, Decennial Census Management Division) briefed on “Census Bureau Non-ID Processing Business Process Overview”.

Amy O’Hara (Census Bureau, Center for Administrative Records Research and Applications) briefed on “Maximizing the Use of External Data”.

Meagan Tydings (Census Bureau, Decennial Census Management Division) briefed on “Respondent Validation Request for Information Results”.

Rick Knowles (MITRE Corporation) provided an additional briefing on some of MITRE’s work (unrelated to JASON) for the Census Bureau.

APPENDIX B. Identity Validation "Chains" to Previous History

In the context of today's Internet, "identity validation" can mean several different things. For example,

- A customer registers a credit card with a merchant for repeated future use. The merchant needs enough information to validate, through a bank, not just the existence of the card, but the identity of the customer supplying the information. The customer is asked to provide address and phone number ("knowledge-based authentication").
- That same customer uses the registered card at a later time. The merchant now needs to verify only that the customer is the same person as the one who first registered the credit card, minimally by username and password ("shared secret").

Note that neither of these examples establishes who a customer "really" is. Instead, the protocols establish (to some degree of assurance) that a present interaction is the authorized continuation of a previous history of interactions—adding a link, as it were, to a "chain" of previous identity history. The purpose of the identity verification in the first example is to link the present customer to a chain that contains the event of his/her opening a bank or credit card account. That in turn, links him to a chain that presumably included his/her supplying to the bank a tax ID number and some form of physical identification. The purpose of the identity verification in the second example is simply to link to the identity established by the first.

Chaining in this manner inhibits fraud because, and to the degree that, it is made hard for a malefactor to link a valid, existing chain of identity (i.e., identity theft) and also hard for him or her to counterfeit a credible false chain in its entirety (i.e., false identity creation).

APPENDIX C: Another View of Census Tradeoffs

The ROC curve in Figure 6 describes just one tradeoff relevant to census practices, that of inclusivity versus accuracy in non-ID processing. A general framework for identifying other tradeoffs follows. Census practices alter the distribution of census response outcomes among these six main possibilities:

1) The response is legitimate (with or without errors) and is correctly associated with the respondent's address. This situation is akin to a "true positive" and contributes to an accurate count.

2) The response is not legitimate (for example, it is fraudulent, mischievously created, or so full of errors that it is beyond any hope of being associated with the respondent's address) and is not counted. This situation is akin to a "true negative" and contributes to an accurate count by avoiding situation #3.

3) The response is not legitimate, but is nevertheless incorrectly associated with an address and is counted. This situation is akin to a "false positive" and contributes to an inaccurate count by overcounting.

4) The response is legitimate, but is incorrectly matched with an address that is not the respondent's. This situation will likely contribute to an inaccurate count, depending on the similarities between the individuals actually living at the falsely matched address and the respondent.

5) The response is legitimate, but is not counted. This situation is akin to a "false negative" and contributes to an inaccurate count by undercounting.

6) No response is received from an individual who should be counted by census policy. This situation is a different form of a "false negative" and contributes to an inaccurate count by undercounting.

An ideal census maximizes outcomes 1 and 2 and minimizes outcomes 3–6. Each practice of the U.S. Census Bureau, including those designed to increase the response rate and those designed to match non-ID responses with individuals or addresses, will affect the distribution of outcomes into these six possible outcomes. JASON emphasizes that methodologies that increase the true positive rate, or that decrease the false positive or false negative rates, may actually *decrease* the overall accuracy of the census by altering this distribution. Therefore, every candidate methodology must be carefully evaluated in light of its impact on this outcome distribution. Such an evaluation can reveal the tradeoffs that are inherent to most, if not all, proposed census practices.

The tradeoff between non-ID processing coverage versus identification accuracy described above can be reanalyzed in light of this framework. In some cases, methods that offer broad coverage of potential non-ID responses are less accurate than methods that offer more narrow coverage. A method that may be less perfect in associating a non-ID response with an identity or address (and therefore only moderately increases the frequency of desirable outcomes 1 and 2) may nevertheless result in a more accurate census than a more perfect method *if* it is applicable to a larger percentage of non-ID responses (thereby minimizing outcome 6).

A second example that can be illuminated by this framework is the tradeoff between respondent burden and identification accuracy. Requesting additional information that facilitates identification can increase the frequency of desirable outcomes 1 and 2, and decrease the frequency of undesirable outcomes 3, 4, and 5 *among the responses received*. But requesting additional information also increases respondent burden, which in turn will increase the no-response outcome 6. Thus, the optimal balance between respondent burden and matching accuracy is achieved when any further increase in respondent burden decreases overall census accuracy (however it be defined; we offer no definition but note that Census accuracy is distinct from matching accuracy) by increasing the positive effect of outcomes 1 and 2 and decreasing the negative effect of outcomes 3, 4, and 5 *less than* it increases the negative effect of outcome 6; and when any further decrease in respondent burden conversely causes improvements in outcome 6 less than it causes harm by changing the frequencies of outcomes 1–5.

This framework suggests that the Census Bureau should therefore determine or estimate—for example, through model studies—the extent to which each proposed practice alters these six outcomes, then use these estimates to inform the implementation or rejection of these proposals. JASON also emphasizes that optimal practices resulting from this analysis are likely to change depending on human nature and other circumstances that are, in part, beyond Census’s control, including the perceived trustworthiness of the census mobile app or form. Therefore, Census should be aware that the optimal practices reflecting these tradeoffs may vary from region to region within the U.S., and even with time within each region.