Census 2000 Evaluation K.1.b
March 12, 2003

Evaluation of the Quality of the Data Capture System and the Impact of the Data Capture Mode on the Data Quality

FINAL REPORT

This evaluation study reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

Joseph Conklin
Decennial Statistical Studies Division

U S C E N S U S B U R E A U
Helping You Make Informed Decisions

CONTENTS

CONTENTS

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

EXECUTIVE SUMMARY

The purpose of evaluation K.1.B is to see how well the reading of census forms can be delegated to automated data capture and imaging technology. We examine the performance of the technology during Census 2000.

The raw data for this evaluation consist of a sample of 768,000 short forms and 768,000 long forms distributed among these types:

- Mailout/Mailback short and long form,
- Enumerator short and long form, and
- Update/leave short and long form.

The enumerator and update/leave forms include Puerto Rico and continental U.S. versions. The mailout/mailback forms include both English and Spanish versions. We used the following methods to collect and analyze the data. The collection method involved the following:

- run the sample of forms through the Census 2000 data capture system,
- key the entire sample after Census 2000 using Key From Image,
- match the Key From Image content with that captured by the automated technology in Census 2000,
- evaluate the content and determine the most likely intent of the respondent,
- determine whether the automated technology, Key From Image, or both correctly captured the content from the paper,
- determine whether the automated technology, Key From Image, or both captured the intended response, and
- create a file of the fields where the methods disagree on content.

The data went through a two stage filtering process. The Key From Image operators entered what they thought was on the scanned image. Then an independent group of analysts looked at the content from KFI and from the automated technology and compared them against what they judged to be the most likely intent of the respondent. They determined intent based on a set of rules they had been trained on.

Throughout this evaluation we present results and comments based on the analysis of data capture errors. The automated technology is prone to any one of the following errors:

- failure to read a field on the form,
- picking up content that is not really there, as in trying to interpret a stray mark,
- incorrectly capturing the content on the paper, and
- correctly capturing what the respondent wrote but this is not what the respondent intended..

KFI is also subject to the same errors.

There is more than one way to miss a respondent's intention:

- in the case of check-box responses, the automated technology or KFI might report a box other than the one chosen by the respondent, and
- in the case of write-in responses, the automated technology or KFI might miss characters or add characters not provided by the respondent.

Picking up the wrong check-box is a hard match error. We determine hard match errors by placing the content read by the automated technology or by KFI against what the clerical evaluators judged was the true response. We compare the two check-box by check-box to see if they are identical. The check-box contents must be identical to be considered a match.

Missing characters or dropping or adding characters can lead to soft match errors. We determine soft match errors by comparing the write-in content read by the automated technology or by KFI against what the clerical evaluators judged was the true response. The comparison is also character by character. The write-in contents do not have to be identical to be considered a match. The divergence between the contents is scored using a soft match algorithm. A soft match error occurs when the divergence score exceeds a threshold.

The method for analysis was to take the judgements of the people assessing the intent of the respondent and then to

- classify the fields on the forms into thirteen separate categories,
- classify fields as to whether the automated technology or Key From Image captured the intent correctly,
- to tabulate the frequency at which the intent was not correctly captured, and
- to break out for the incorrect cases the reasons why.

- classify a check-box field as to whether it is a hard match error,
- classify a write-in field as to whether it is a soft match error,
- calculate the overall hard match and soft match error rates by form and field,
- test for statistically significant relationships between error rates and factors such as form and field category,
- identify error rates for specific fields that are high enough to be considered outliers, and
- show whether the overall error rate for a specific group of fields is high enough to be considered an outlier.

When evaluating the performance of the automated data capture and imaging technology in Census 2000, we ideally wish to answer two basic questions:

- does it accurately record the contents of a field, and
- does it accurately record what the respondent (directly or through an enumerator) meant?

Content can differ from intent. This can happen for reasons such as stray marks being read as characters or if the respondent writes poorly. The standard for Key From Paper is to capture content with no more than a 2 percent error rate. Our answer to the first question is as follows.

- The performance of the automated technology depends on whether its character recognition algorithm determines the content is clear enough to process.
- If the automated technology determines the content of a write-in field is clear, it processes it with a typical error rate of 1.0 percent to 1.1 percent.
- If the automated technology determines the content of a check-box field is clear, it processes it with a typical error rate of 1.2 percent to 1.5 percent.
- If the automated technology rejects content as unclear, the typical error rate after remedial keying by human operators is 4.8 percent to 5.3 percent.

We can summarize our answer through these confidence intervals for the median nonblank error rate, averaged over all fields. They are constructed to support multiple pairwise comparisons with 90 percent confidence.

Optical Character Recognition mode of data capture (technology thinks content is good	1.007 percent to 1.128 percent confidence interval for the soft match error rate

and processes it as a write-in field)

Optical Mark Recognition mode of data capture (technology thinks content is good and processes it as a check-box field)	1.185 percent to 1.495 percent confidence interval for hard match error rate
Key From Image mode of data capture (technology thinks content is not good and sends it to keying)	4.781 percent to 5.319 percent confidence interval for theerror rate for check-box and write-in fields combined.

The intervals do not overlap. We conclude with 90 percent confidence the modes are all significantly different from one another. The Key From Image mode tends to deal with content particularly hard for human or machine to interpret. Its error rate is not necessarily a poor reflection on the automated technology.

The error rates reflect effects of multiple sources such as the following:

- the hardware design of the automated technology
- the design of the software used by the automated technology
- the complexity of the editing rules used in this evaluation's keying operation
- general typing errors in this evaluation's keying operation
- collection of our data before all in process Census 2000 QA checks were complete
- color choices for some fields that made it harder for the automated technology to work.

Unfortunately, the design of our data collection did not allow us to determine the contribution of these various causes to the overall error rates. The error rates shown in K.1.B should be considered conservative upper limits for the true rates attributable solely to the hardware and software configuration of the automated technology.

The error rates for OCR and OMR are significantly below the target for KFP by a considerable margin. Although good news, this performance is after the automated technology recognizes and accepts content. Not all content is accepted. In the case of the write-in fields in our data, only 24,857,562 of 31,523,300 were accepted. The rest were sent to KFI. The percent accepted was 78.9. Although the automated technology brought increased speed and efficiency to Census 2000 processing, considerable human resources were still required to handle the many millions of write-in fields that posed a problem for it.

We now turn to the questions in the study plan for this evaluation.

- Is there a statistically significant difference in data quality by field, form, Census 2000 regional census center, data capture center or race categories?

- Is there a statistically significant difference in data quality between Optical Mark Recognition, Optical Character Recognition, and Manual Inspecting and Keying?

- Does Key From Image affect our ability to capture intent at the risk of a higher soft match error rate?

- Are some fields sent unnecessarily to Key from Image more often than others?

- Across modes of data capture, what are the reasons for not capturing the intended response?

Here are our answers. At several points in the following, we refer to "fields filled out for multiple persons on a form." These are fields like name, age, and sex which appear more than once on a decennial census form. They are repeated so information can be recorded for every member of a household. For other fields, we use the phrase "fields filled out for only one person on a form."

The statements about statistical significance frequently refer to form type and field category. Form type means one of the long or short forms included in our sample of raw data. Field category means one of thirteen categories into which the fields on the various forms were classified for analysis purposes. The specific form types in our raw data consisted of

- Short Form, Mailout/Mailback (d1),

- Short Form, Enumerator (d1e),

- Short Form, Enumerator, Puerto Rico (d1er),

- Short Form, Mailout/Mailback, Spanish (d1s),

- Short Form, Update/Leave (d1u),

- Short Form, Update/Leave, Puerto Rico (d1ur),

- Long Form, Mailout/Mailback (d2),

- Long Form, Enumerator (d2e),

- Long Form, Enumerator, Puerto (d2er),

- Long Form, Mailout/Mailback, Spanish (d2s),

- Long Form, Update/Leave (d2u), and

- Long Form, Update/Leave, Puerto Rico (d2ur).

The thirteen categories used to classify the fields for analysis were

- Coverage (Household coverage questions on enumerator form),
- Form Management (Contact data, persons added or canceled on enumerator form),
- POP–Demographic (Age, marital status, ancestry, and similar demographic data),
- POP–Disability (Existence and extent of personal disability of household members),
- POP–Education (Educational attainment of household members),
- POP–Ethnic (Ethnic data of household members, including Hispanic origin),
- POP–Income (Income characteristics of household members),
- POP–Military (Military service characteristics of household members),
- POP–Name (First, middle, and last names of household members),
- POP–Occupation (Occupational characteristics of household members),
- POP–Race (Racial data of household members),
- Residential Profile (Features, expenses, age and similar data of residential structure), and
- Special Housing (Special Place, Usual Home Elsewhere, and related designations).

Is there a statistically significant difference in the percentage of erroneously captured fields by form?

- **Respondent-returned forms have statistically significantly higher nonblank hard or soft match error rates for ethnic, name, and race fields compared to enumerator-returned forms**.
- **Although enumerator-returned forms have lower soft match error rates for name related fields compared to respondent-returned forms, the rates for name related fields are higher compared to rates for other fields on forms returned by enumerators**.

Is there a statistically significant difference in the percentage of erroneously captured fields by field?

- **For fields filled out for only one person on a form, the hard or soft match error rate is significantly affected depending on the specific field being considered; form type or field category do not have a significant influence**.

- **For fields filled out for multiple persons on a form, the soft match error rate is significantly affected by form type and field category.**

Is there a statistically significant difference in the percentage of erroneously captured fields by Census 2000 regional census center?

- **Census 2000 regional census center is a significant influence on the hard or soft match error rate.**

- **The soft match error rate for name related fields in Census 2000 regional census centers 22, 23, 27, 29 and 32, centers covering areas of traditional immigrant concentration in Florida, Los Angles, and New York City is significantly higher compared to other regional census centers.**

Is there a statistically significant difference in the percentage of erroneously captured fields by data capture center?

- **For fields that are filled out for only one person on a form, the largest significant factor affecting the nonblank error rate is form. There is a significant secondary contribution from field category. The structure of the data set did not allow us to test field for significance.**

- **For fields that are filled out for multiple persons on a form, the largest significant factor affecting the nonblank error rate is field category. There is a significant secondary contribution from form. The structure of the data set did not allow us to test field and person number for significance.**

- **Although not outliers in all four data capture centers, the categories Form Management and POP–Name have the highest nonblank error rates in all.**

Is there a statistically significant difference in the percentage of erroneously captured fields by race?

- **The race response has a statistically significant effect on the nonblank error rate. Within our limited data set for race, we are not able to find individual error rates that are outliers. The effect of race may be part of other significant factors time did not permit us to include in our models. It would be helpful to include other factors with race in a future evaluation.**

Is there a statistically significant difference in the percentage of erroneously captured fields between Optical Character Recognition, Optical Mark Recognition, and Key From Image, the modes of data capture?

- **For fields filled out for only one person on a form, the error rate is not significantly affected by data capture mode.**

- **For fields filled out for multiple persons on a form, the specific field being considered and the data capture mode interact to significantly affect the error rate.**

- **As can be seen in the confidence intervals stated above, for all fields, Optical Character Recognition has the lowest error rate, followed by Optical Mark Recognition, and then Key From Image. All three rates are statistically different.**

Does Key From Image improve our ability to capture intent at the risk of a higher soft match error rate?

- **When content is sent to Key From Image, we do not capture respondent intent better at the expense of a higher soft match error rate.**

- **For fields filled out for only one person on a form, there is not a statistically significant relationship between the impact of Key From Image and the soft match error rate.**

- **For fields filled out for multiple persons on a form, there is a significant relationship between Key From Image impact and the soft match error rate, but it changes depending on what specific field is being considered.**

Are some fields sent unnecessarily to Key From Image more often than others?

- **Compared to other fields, name related fields are more likely to go to Key From Image unnecessarily, particularly for the middle initials of higher numbered**

persons in the household.

- **For fields filled out for only one person on a form, the redundancy rate is significantly affected depending on the specific field category being considered**.

- **For fields filled out for multiple persons on a form, the redundancy rate is significantly affected depending on the specific form and field category being considered.**

Across modes of data capture, what are the reasons for not capturing the intended response?

- **The most frequent ways we fail to capture the intended response are**

  Extra check-box--the output from the automated technology output shows more check-boxes marked than are on the scanned image,

  Missing characters, the output from the automated technology has fewer characters than the scanned image, and

  Wrong character, the output from the automated technology and the scanned image have the same number of characters, but output from the technology disagrees with the image in one or more characters.

- **The most common reasons our clerical evaluators found for these problems are**

  Poor handwriting--the respondent's handwriting makes one letter look like another, but one can tell what the respondent meant,

  No reason found--the response is written clearly and there is nothing to suggest why it was not captured correctly, and

  Rules not followed, the rules used during the KFI after Census 2000 processing in an attempt to edit the content on the fly were not followed.

The preceding results support strategic and tactical comments about the future of automated data capture and imaging technology in the decennial census. At the strategic level, the future role of the automated technology reduces to two possibilities.

- The automated technology has a supporting role in decennial census processing. It is used to rapidly complete the clear and easy responses. Traditional methods claim the majority of resources for especially difficult responses.

- The automated technology has a dominant role in decennial census processing. Census forms are dramatically streamlined and redesigned to eliminate the long form's vast

sea of handwritten responses requiring interpretation.

Which role it will have depends on whether we retain the long form. As long as we gather huge numbers of write-in responses in the decennial census, a supporting role is far more likely. At the tactical level, several possible research questions exist for tests leading up to the 2006 Census test.

- Should the Census Bureau expand efforts to make certain groups of fields
  easier for respondents to understand and fill out?

- Do the outlier error rates for the long form Puerto Rico update leave form
  suggest challenges to the automated technology that require increased attention?

- Do the outlier error rates for name related fields on the

  English language enumerator short form,
  Spanish language mailout/mailback short form,
  English language enumerator long form, and
  English language update leave long form for Puerto Rico

  suggest challenges to the automated technology that require increased attention?

- Is the disproportionately higher number of outlier error rates on the English language
  mailout/mailback long form an issue?

- Is it necessary to explain why the nonblank error rate for name related fields occupies one
  of the top two positions in all four data capture centers?

- Is the especially high nonblank error rate for name related fields in Census 2000 regional
  census center of traditional immigrant concentration something that requires more
  investigation?

- Should certain fields sent automatically to KFI be allowed to go through the automated technology for processing?

- If the present long form data collection process is retained for the 2010 census in
  is it worthwhile to improve the quality performance of the automated technology?

1. BACKGROUND

In the discharge of its Constitutional and statutory obligations, the Census Bureau takes the paper responses from the decennial census and converts them to electronic files that are stored on computers. In this way, the files are readily edited, tabulated, and analyzed. One medium for converting responses to stored electronic files is Key From Paper (KFP). In KFP, keying is done directly from the census form.

Because the Census Bureau employs a wide array of forms to enumerate the population, the success of KFP or any other medium depends on complex procedures and tight controls. While these procedures and controls operate on many levels of detail, at the most basic level there are two essential challenges.

The responses to a form can be indicated by checking a box or by writing an answer in the spaces provided for this purpose. The first challenge consists of distinguishing the check-box and write-in responses and accurately transcribing the contents of each.

All the varieties of forms reduce to two basic types: short and long. Most households receive the short form. It asks for information on household size and on the gender, race, and Hispanic origin characteristics of the members. The long form asks for this and for additional information on income, education, occupation, and other characteristics. Separate processes are needed to handle each type of form. The second challenge consists of matching the type of form to the right process.

Automated data capture and imaging technology has tremendous potential to increase accuracy, efficiency, and speed beyond the capabilities of the traditional media. This technology was part of the 1995 Census Test. It worked well enough to be part of the Census 2000 Dress Rehearsal. Its performance in the Census 2000 Dress Rehearsal was covered in, H3: Quality of the Data Capture System, an evaluation issued in July 1999. That evaluation reported the overall percentage of erroneously captured check-box fields was 0.81 percent. The corresponding percentage for write-in fields was 3.01 percent. Several recommendations for the next application of the technology were accepted:

- modify the definition of an error for write-in text responses to include only significant deviation from what is present on the form, as long as it does not impact the usage of the data,
- include more content edits as a way of improving the data capture quality,
- add a check-out function to ensure that data are captured for all scanned forms,
- and use the Data Capture Audit and Resolution process during Census 2000.

The contractor developing this technology for the Census Bureau continued to refine it after the Census 2000 Dress Rehearsal. The evaluation issued in 1999 anticipated a need to once again evaluate its use in light of these refinements. With the conclusion of Census 2000, we now have the data to carry this evaluation to the next stage of currency and depth.

Evaluation K1.B, Evaluation of the Quality of the Data Capture System and the Impact of the Data Capture Mode on the Data Quality, presents the next detailed stage in our understanding of what automated data capture and imaging technology means for data quality in the decennial census. The study plan for this evaluation was issued in December 2000 and encompasses these questions.

- Is there a statistically significant difference in the percentage of erroneously captured fields by field, form, Census 2000 regional census center, data capture center, or race categories?

- Is there a statistically significant difference in the percentage of erroneously captured fields between optical mark recognition, or OMR mode, optical character recognition, or OCR mode, and fields resolved by manual inspecting and keying, or KFI mode?

- Does KFI improve our ability to capture intent at the risk of a higher soft match error rate?

- Are some fields sent unnecessarily to KFI more often than others?

- Across modes of data capture, what are the reasons for not capturing the intended response?

The methods used to answer these questions, with the subsequent results and conclusions, appear in subsequent sections. For definitions of common or special terms in this section, see the glossary in Appendix M.

2. METHODS

2.1 Collecting the Raw Data to Measure the Quality of Data Capture

The method for collecting the raw data worked as follows:

- determine the forms to be included,
- determine the number of each form to sample,
- collect the required types and numbers of forms after Census 2000 processing,
- and have keying personnel at Jeffersonville, IN, record the form content by KFI.

Following this work, clerical evaluators at Jeffersonville, IN,

- matched the KFI content with that captured by the automated technology in Census 2000,
- studied the content and judged what was the most likely intent of the respondent using the rules they were trained on,
- determined whether KFI or the automated technology correctly captured the content on the paper, and
- determined whether the KFI content or the content captured by the automated technology was the intended response, and
- if the content captured by the automated technology was determined to be in error, they made a determination as to the reason for the incorrect value.

The final phase required the coordinated effort of an outside contractor and personnel from the Decennial Systems and Contracts Management Office (DSCMO) to create a file of the fields where the clerical evaluators determined the automated technology and KFI disagree on the content.

The raw data for this evaluation consist of 768,000 short forms and 768,000 long forms distributed among these types:

- Mailout/Mailback short and long form,
- Enumerator short and long form, and
- Update/leave short and long form.

The enumerator and update/leave forms include Puerto Rico and continental U.S. versions. The mailout/mailback forms include both English and Spanish versions. Four forms included for sampling were later dropped. It turned out either they did not go to automated capture or they were of too low a volume to justify the effort needed to match them. A list of the forms ultimately included in the sample can be found in Appendix A. The KFI and matching operations were concluded by the end of 2001. The finished files were delivered for analysis in the first quarter of 2002.

2.2 The Varieties of Data Capture Errors

Throughout this evaluation we present results and comments based on the analysis of data capture errors. At first thought, "What is a data capture error?", is a simple question. Depending on the context, several possible answers exist.

In the later sections of this evaluation, we will identify in context exactly what we mean by a data capture error. For purposes of general understanding, we summarize the various possibilities.

The automated technology is prone to any one of the following errors:

- failure to read a field on the form,
- picking up content that is not really there, as in trying to interpret a stray mark,
- incorrectly capturing the content on the paper, and
- correctly capturing what the respondent wrote but this is not what the respondent intended..

KFI is also subject to the same errors.

There is more than one way to miss a respondent's intention:

- in the case of check-box responses, the automated technology or KFI might report a box other than the one chosen by the respondent, and
- in the case of write-in responses, the automated technology or KFI might miss characters or add characters not provided by the respondent.

Picking up the wrong check-box is a hard match error. We determine hard match errors by placing the content read by the automated technology or by KFI against what the clerical evaluators judged was the true response. These are the evaluators mentioned in section 2.1. We compare the two check-box by check-box to see if they are identical. The check-box contents must be identical to be considered a match.

Missing characters or dropping or adding characters can lead to soft match errors. We determine soft match errors by comparing the write-in content read by the automated technology or by KFI against what the clerical evaluators judged was the true response. The comparison is also character by character. The write-in contents do not have to be identical to be considered a match. The divergence between the contents is scored using a soft match algorithm. A soft match error occurs when the divergence score exceeds a threshold. Pseudocode for the soft match algorithm appears in Appendix G.

This evaluation is mainly, but not exclusively, focused on hard match and soft match errors.

2.3 General Comments About Data Editing Methods

Before generating the results and recommendations of this evaluation, we first edited the raw data. We did this to unduplicate the data and to separate them into logical portions for analysis.

The raw data consist of two groups of files. One group has a separate file for each of the twelve Census 2000 regional census centers. These twelve files hold all the contents originally read by the automated data capture and imaging technology. There are a total of 69,701,287 records in the twelve files, each record corresponding to a field on a Census 2000 form.

The second group is a stand alone file that holds all the data from the first set where the automated technology and KFI disagree on the contents of a field. There are 1,725,518 records, each record also corresponding to a Census 2000 field on an individual form.

We were prepared to use the combination of form, field, and Census ID number in a data record as a unique key. However, examination of the raw data showed records where combinations of these variables were repeated among records. Two possible ways duplicates can enter the raw data are

- for the same form to be run through the automated technology more than once by mistake, and

- for two or more Census 2000 enumerators to return forms for the same Census ID that are inadvertently processed as if they were distinct households.

Unfortunately, the limits of time did not allow us to verify whether these two possibilities or some others were the actual reasons for the duplicates.

Our policy for handling duplicate records was to retain the one with the most completed fields. If two or more duplicate records had the same number of completed fields, we randomly selected one to retain. The file consisting of 1,725,518 disagreements between the automated technology and KFI reduced to 1,715,967 after unduplication.

After unduplication, we initially broke the file further into one set of 1,049,256 records we were able to match successfully against the twelve regional census center files mentioned above. The residual set of 666,711 records are those we were not able to match. Near the end of writing the initial draft of this evaluation, we discovered the reason why they did not match. The details can be found in section 3, the limits section. For the final draft, we are able to analyze the file of disagreements between methods as a single data set using all 1,715,967 unduplicated records.

We next summarize how we analyzed the data, leaving more detailed descriptions to the results section of this evaluation. The highlights of the results can be found in section 4.1. For definitions of common or special terms in this section, see the glossary in Appendix M.

2.4 General Comments About the Data Analysis Methods

The general strategy for analysis is to take what is judged to be a respondent's intent and then to

- classify the fields on the forms into thirteen separate categories,
- classify the fields as to whether the automated technology or KFI captured the intent correctly,
- to tabulate the frequency at which the intent was not correctly captured,
- to break out for the incorrect cases the reasons why,
- classify a check-box field as to whether it is a hard match error,
- classify a write-in field as to whether it is a soft match error,
- calculate the overall hard match and soft match error rates by form and field,
- test for statistically significant relationships among error rates and factors such as form and field category,
- identify error rates for specific fields that are high enough to be considered outliers, and
- show whether the overall error rate for a specific group of fields is high enough to be considered an outlier.

2.5 Applying the Quality Assurance Procedures

We applied quality assurance throughout the creation of this report. They encompassed how we determined evaluation methods, created specifications for project procedures and software, designed and reviewed computer systems, developed clerical and computer procedures, analyzed data and prepared this report.

3. LIMITS

3.1 Raw Data are Not a Random Representative Sample of the U.S. Population

Some Census 2000 personnel have used the raw data from this evaluation for their own special queries. We are aware of analysis to understand trends in responses to some of the personal disability questions on the long form. We are also aware of analysis to understand patterns in the Hispanic origin write-ins. After this evaluation, we will issue an evaluation examining exclusively the industry and occupation fields.

All users of the data in this evaluation should not treat them as if they are a random, representative sample of the U.S. population. Although we strove to include the more frequently occurring forms, a representative sample of the population was not a goal of the data collection plan.

3.2 Failure to Obtain All Data Originally Planned

The road from form collection to data capture to KFI to matching and to assessment for respondent intent had some bumps. Setting up the network server to support KFI took two and one-half weeks longer than expected. Loading the form data to the server was planned for March 2001 but was not completed until July 2001. Some of the CD-ROMs holding the form data for KFI became corrupted. As a result, approximately 10 percent of the data had to go to KFI a second time.

The computer program to perform the matching took three weeks longer than expected to complete and test. We relied on internal Census Bureau resources for matching. Obtaining all the data required adhering to a tight schedule before these resources were needed for urgent Census 2000 processing activities. We discovered a separate matching program was needed for each of the twelve forms. This introduced more delays which made adhering to the schedule impractical.

Also, for various reasons, we were unable to provide in one installment all the form data that needed matching. Some of the long form data arrived after the matching for these forms had started. Additional time was needed after this happened to figure out how to align the new data with what had already been matched.

The net result was we lost the chance to match the 10 percent of the data that went through KFI twice. The experience pointed to the desirability of placing a project of this scope and complexity under the responsibility of a single contractor. We paid a price by attempting to accomplish ourselves certain things we were not in the best position to perform.

How does the failure to match 10 percent of the data affects this evaluation? We believe results are not significantly affected. We conclude this for two reasons. First, the problems we

encountered occurred after processing by the automated technology. It does not change how it captures data depending on how well we perform KFI or matching afterwards.

Second, our understanding of how CD-ROMs are corrupted makes it more likely than not the unmatched data were randomly distributed between forms, Census 2000 regional census centers, and all other relevant factors conducive to distortion by clustering. Unfortunately, time constraints have prevented us from reviewing our documents in a manner to establish this position beyond a reasonable doubt.

In sum, we have an extra, unanticipated reason for treating the results of this evaluation as provisional. However, they still hold some meaning and value for understanding the implications of the automated technology for data quality.

3.3 Resolution of 666,711 Records Not Matched to the Twelve Regional Census Center Files

In mid-2002, we worked with our contractor to find out why we did not match 666,711 records. We discovered our February 2002 request to the contractor to exclude from the twelve regional census center files the records existing in the file of disagreements between methods. That was why they could not be matched. In February 2002, we hoped to combine all the files during analysis. Excluding the records prevents duplicated data from contaminating the analysis.

We found computer memory limits made combining files impossible. Solving this problem and working out the analysis of the data took four months. By then we had forgotten our February 2002 request. We should not have been able to match any records, but for reasons still unknown, we were able to match some. This proved harder to explain than matching none. With what we know now, the 666,711 records can be included as valid cases. We do so in this final draft.

3.4 Subjectivity in Interpreting the Most Likely Intent of the Respondent

The data for this evaluation are the product of a two stage filtering process. The KFI operators entered what they thought was on the scanned image. Then an independent group of analysts looked at the content from both methods and compared them against what they judged to be the most likely intent of the respondent.

We do not have an absolute standard of correct content to measure against. When responses are written outside of boxes, crossed out on a page, squeezed so that more than one letter appears in a single write-in box, and so on, then judging intent is difficult and the possibility for subjective error is the greatest. Also, judging the intent of the respondent is a subjective activity in and of itself. Fortunately, we believe there are enough correctly judged cases to support a good approximate understanding of how the data quality of the automated technology compares to that of the benchmark method, KFI. We now turn to building that understanding. For definitions of common or special terms in this section, see the glossary in Appendix M.

3.5 Data Reflect Multiple Sources of Error Beyond Those Attributable to System Design

From section 3.4 it is clear the data on which evaluation K.1.B are based are not pure in the sense of reflecting errors that arise solely from the hardware and software design of the automated data capture and imaging technology. As with many complex projects, several compromises were made in the course of implementing the technology that affected the nature of the data available from our data collection process. The compromises induced additional limitations that are worthy of separate mention. We summarize these here and strongly encourage readers to keep them in mind when perusing this report.

When the keyers reproduced the contents of our QA sample after Census 2000 processing, they were asked to key and edit at the same time. The rules for the keyers required them to edit the content if any one of a large number of special circumstances arose. One example of an editing rule is one that said to key in a string of 8's if certain fields were blank. Other rules required keyers to adjust the formatting of certain numeric values supplied by respondents. These cases were counted as errors if our analysts concluded the resulting content did not properly capture the respondent's intent.

It proved difficult in many cases for the keyers to keep the built up habit of exact reproduction from clashing with the editing rules. In the course of implementing data capture, the editing rule set was modified in an attempt to lessen this problem. The data for evaluation K.1.B were collected after this modified rule set was put in place. Even after modification ample opportunity for confusion remained. Obviously, errors caused by the keyers' confusion with this rule set are not the fault of how the technology was designed. In this evaluation what we are counting as an error is whether our analysts thought what was captured during the census differed from the respondent's intent. It is possible, therefore, that a census keyer's product was correct under the requirements of the automated technology but incorrect in this evaluation.

The processes for Census 2000 forms included several in-stream quality checks to maximize the probability of correctly recording the responses. We could have collected our data at any point in Census 2000 processing. The point we thought was the most practical choice turned out to be where some but not all of these quality checks were completed. It is likely some of the errors in our data would have been removed if they had gone through the entire battery of checks. To the extent this happened, we are left with a certain number of errors that should not be charged to the design and implementation of the automated technology.

The outline color for check-box fields on the Census 2000 forms was black. While intended to make the forms more readable to the human eye, it made it harder for the automated technology to detect the degree of contrast necessary to trigger recognition of a character. Characters lost or

garbled as a result of inadequate contrast therefore are a function of form design rather than the design of the automated technology. Besides the issue with the black background color, other aspects of form design made it harder for the automated technology to perform optimally.

Unfortunately, we are not able to separate these various effects from our data. As a result, we probably have a picture of the automated technology's performance that while useful is somewhat harsher than what a purer data set would reveal. The error rates shown in K.1.B should be considered conservative upper limits for the true rates attributable solely to the hardware and software configuration of the automated.

4. RESULTS

4.1 Contents of This Section (Highlights of Results)

In this section, we place the highlights of the results. We believe readers will more easily understand the logic underlying our suggestions for possible future research if they can find the highlights of the results in one place. This section should also serve those readers needing only a summary view of the results.

At several points in this section, we refer to "fields filled out for multiple persons on a form." These are fields like name, age, and sex which appear more than once on a decennial census form. They are repeated so information can be recorded for every member of a household. For all other fields, we use the phrase "fields filled out for only one person on a form."

We have framed the highlights as answers to questions readers may have about the quality of automated data capture and imaging technology. The questions form the section titles. For definitions of common or special terms in this section, see the glossary in Appendix M.

4.1.1 How do the soft and hard match error rates compare for the modes of capture?

We begin by describing how we determine hard and soft match errors. We compare the Census 2000 context value against the evaluation truth value. The context value is the characters returned by the automated technology after special editing. The editing removes extra characters inserted by the automated technology that are needed to execute its program. The evaluation truth value is the content that was judged to be the most likely intent of the respondent. This judgement was performed by the clerical evaluators in Jeffersonville, IN, mentioned in section 2.1.

For check-box fields, we compare the context value to the evaluation truth value check-box by check-box. If the sequence of marked and unmarked check-boxes fails to match exactly, the context value is a hard match error. We do not compare check-box fields that are trailing blanks.

For write-in fields, we take all the characters in the context value and the evaluation truth value and count how many times each appears. Then we pass this information to the soft match algorithm to score the degree to which context and truth diverge. If the returned score exceeds a threshold, the context value is soft match error case. The algorithm does not count trailing blanks in the scoring.

To compare hard and soft match error rates by mode of data capture, we display Table One. Table One contains approximate 96.5 percent confidence intervals for the median nonblank error rates. These are combined rates for hard and soft match errors, averaged across all forms and fields, and broken out by capture mode. The reason for 96.5 percent confidence intervals is in Appendix E. This way we have 90 percent confidence about how the modes compare.

If the confidence intervals for a pair of modes overlap, we conclude the median error rates are not significantly different. None of the confidence intervals overlaps with the other two. We conclude the error rates by mode are all significantly different from each other. OCR is the lowest. KFI is the highest. Since KFI occurred for fields the automated technology considered too hard to read, we are not surprised to see it associated with a significantly higher rate for hard and soft match errors.

Table 1. Approximate 96.5 Percent Confidence Intervals for Median Nonblank Error Rates By Data Capture Mode, Consolidating Hard and Soft Match Errors Across All Fields and Forms

Data Capture Mode	Lower Confidence Interval Bound	Upper Confidence Interval Bound
KFI	4.781%	5.319%
OCR	1.007%	1.128%
OMR	1.185%	1.495%

4.1.2 How do the above error rates compare to the Census 2000 Dress Rehearsal?

Our source for the Census 2000 Dress Rehearsal error rates is evaluation H3: Quality of the Data Capture System, issued in July 1999. It reported the overall error rate for check-box fields was 0.81 percent, with a standard error of 0.04 percent. The overall error rate for write-in fields was 3.01 percent, with a standard error of 0.05 percent.

Unfortunately, our error rates are not directly comparable to the Census 2000 Dress Rehearsal for four reasons:

- the raw data were restricted to forms mailed back by respondents,
- the raw data were restricted to short forms,
- the raw data were not broken out by mode of data capture, and
- the automated technology was still being designed before and immediately after the Census 2000 dress rehearsal.

We can compute error rates restricting ourselves to the same forms as were used in evaluation H3. Even after this, to achieve a nearly direct comparison, we must blend the KFI error rate with the OCR and OMR error rates to duplicate evaluation H3's failure to break out by data capture mode. We do not believe this exercise is worth the effort involved. Evaluation H3 does say the Census Bureau's maximum threshold for errors under the traditional data capture methods is 2.0 percent. The performance of the automated technology in Census 2000, as reflected in the OCR and OMR error rates, is significantly better than 2.0 percent by a considerable margin. We consider this insight the most valuable of any we can draw from comparisons to evaluation H3.

Although the error rates are not directly comparable, the spread between the OMR and OCR rates in H3 and the corresponding rates in K.1.B is large enough to deserve some comment. In

fact, in

K.1.B, the overall OCR error rate is lower than the overall OMR error rate, the exact opposite of the results in K.1.B. An answer is suggested by studying the different ways we can misinterpret respondent intent. A full discussion of misinterpretation and misinterpretation rates is in section 4.11. For purposes of discussion here, we note the misinterpretation rate correlates with the hard or soft match error rate. The behavior of the former sheds light on the latter.

The OMR misinterpretation data show 90% of the cases are for "extra check boxes." This type of misinterpretation occurs when the automated technology shows more boxes checked than actually occur on the form. The length of the captured content is longer than the content on the paper. This makes it impossible to meet the character for character correspondence requirement which avoids a hard match error.

The OCR misinterpretation data show 86% of the cases are for "wrong character." This type of misinterpretation occurs when the automated technology preserves the length of the content but alters one or more characters. As explained in section 4.1.1, the error measure we use for write-in fields is the soft match error rate. The soft match error condition has a looser criterion compared to the one for hard match error. The automated technology can alter some of the characters in the content, but as long as the alternation preserves the length and does not violate the threshold in the soft match algorithm, it is possible to avoid a soft match error.

We conclude the OCR median error rate is benefitting from a relatively more charitable criterion for error, and this explains the reversal in magnitude between OCR and OMR compared to the 1998 Dress Rehearsal. This more charitable criterion was adopted after then.

4.1.3 How do the hard and soft match nonblank error rates compare for Respondent-Returned vs. Enumerator-Returned Forms?

As we can see from section 4.2, the two groups are statistically equal for fields in the Housing Profile, POP–Demographic, POP–Disability, POP–Education, POP–Income, POP–Military, and POP–Occupation categories. The automated technology performs better for enumerator-returned forms in the POP-Ethnic, POP–Name, and POP–Race categories. Although not the source for the majority of the data in Census 2000, it is helpful the enumerator- returned forms show lower error rates for the critical variables of ethnicity and race.

4.1.4 What forms have particularly high hard or soft match nonblank error rates?

As we can see from section 4.3, high outliers appear in the field category POP-Name for forms

- d1e, the English enumerator short form,
- d1s, the Spanish mailout/mailback short form,
- d2e, the English enumerator long form, and
- d2ur, the English update/leave long form.

After averaging across all data capture modes and fields, the form with the most high or very high outliers is d2ur. The capture of name and ethnicity fields on this form is a challenge for the automated technology.

4.1.5 What can we say about the association between form, field, and field category and the hard or soft match nonblank error rates?

These factors are nested. The individual fields nest within the categories, and the categories nest within the forms. In terms of the variation in the nonblank error rate, it is possible to have a significant contribution by the individual fields. There may be a significant marginal contribution of field category above and beyond the individual fields, and a like possibility exists for the marginal contribution of form beyond field category.

As we can see in section 4.4, for fields that are filled out for only one person on a form, the only significant factor affecting the nonblank error rate is field. There is no significant contribution of form or field category. In the other words, differences in the nonblank error rate are driven more by which field one chooses to look at. The choice of form or field category is not a significant influence.

Section 4.4 also shows for fields that are filled out for multiple persons on a form, the largest significant factor affecting the nonblank error rate is field category. The structure of the raw data did not allow us to estimate the contribution of field.

4.1.6 In addition to the factors in the above question, what can we say about the impact of person number for fields that have them?

The structure of the raw data does not allow us to estimate the effect of person number on the variation in the nonblank error rate. Another way to assess the impact of person number is to examine error rates that are considered high and very high outliers. Using the information available in Appendix H, within this restricted set, we do not detect a significant difference in how error rates are distributed by person number.

4.1.7 In addition to the factors in the above two questions, what can we say about the impact of data capture mode on hard or soft match nonblank error rates?

The three data capture modes are OCR, OMR, and KFI. The results of including data capture mode in the analysis can be found in section 4.5. For fields that are filled out for only one person on a form, the only significant factor affecting the nonblank error rate is form. There is no significant contribution of field category, data capture mode, or the interaction of field category and mode. The structure of the data set did not allow us to test field for significance.

For fields that are filled out for multiple persons on a form, the largest significant factor affecting the nonblank error rate is the interaction of field and mode. Interaction means that the effect of

field will change depending on the mode. The field and mode do not operate independently in their effect on the nonblank error rate. There is a significant secondary contribution of field category. The structure of the data set did not allow us to test field and person number for significance.

Outlier error rates by data capture mode do not appear when the data are analyzed at the field category level. They appear at the field level, and we see different issues highlighted for different forms. For the d1s, the Spanish mailout/mailback short form, name related fields is a dominant issue. For the d2, the English mailout/mailback long form, and the d2u, the English update/leave long form, the write-in fields for other race or ethnicity appear many times as outliers. The d2e, the English enumerator long form, shows several outliers for occupation related fields.

4.1.8 If we replace data capture mode with data capture center in the factors in the above question, what can we say about the impact of data capture center on hard or soft match nonblank error rates?

The four data capture centers are Baltimore, Jeffersonville, Phoenix, and Pomona. The results of including data capture center in the analysis for data capture center are covered in section 4.6. For fields that are filled out for only one person on a form, the largest significant factor affecting the nonblank error rate is form. There is a significant secondary contribution from field category. The structure of the data set did not allow us to test field for significance.

For fields that are filled out for multiple persons on a form, the largest significant factor affecting the nonblank error rate is field category. There is a significant secondary contribution from form. The structure of the data set did not allow us to test field and person number for significance.

Although not outliers in all four data capture centers, the categories Form Management and POP–Name have the highest nonblank error rates in all. Form Management covers the person added and person canceled fields on enumerator forms. It is encouraging to note only one of 52 outliers for Form Management was for adding or canceling persons.

4.1.9 If we replace data capture center with Census 2000 regional census center in the factors in the above question, what can we say about the impact of Census 2000 regional census center on hard or soft match nonblank error rates?

There were twelve Census 2000 regional census centers:

- 21 covered Connecticut, Maine, Massachusetts, New Hampshire, upstate New York, Puerto Rico, Rhode Island, and Vermont;
- 22 covered northern New Jersey and metropolitan New York City;
- 23 covered Delaware, the District of Columbia, Maryland, southern New Jersey, and Pennsylvania;

- 24 covered Michigan, Ohio, and West Virginia;
- 25 covered Illinois, Indiana, and Wisconsin;
- 26 covered Arkansas, Iowa, Kansas, Minnesota, Missouri, and Oklahoma;
- 27 covered Alaska, northern California, Idaho, Oregon, and Washington state;
- 28 covered Kentucky, North Carolina, South Carolina, Tennessee and Virginia;
- 29 covered Alabama, Florida, and Georgia;
- 30 covered Louisiana, Mississippi, and Texas;
- 31 covered Arizona, Colorado, Idaho, Montana, Nebraska, Nevada, New Mexico, North Dakota,South Dakota, Utah, and Wyoming; and
- 32 covered southern California and Hawaii.

We carried out the significance testing for Census 2000 regional census center in two ways. The main analysis was restricted to the 18,183 combinations of form, field, and regional census center used in the initial draft of this evaluation. This appears in sections 4.7.3 and 4.7.4.

The analysis on the full set of 27,254 combinations is in Appendix K. As discussed in section 4.7.1, we believe including all 27,254 combinations in the main analysis leads to major distortions. Our comments here are based on the discussion in section 4.7.

For fields that are filled out for only one person on a form, the largest significant factor affecting the nonblank error rate is form. There is a significant secondary contribution of field category. The structure of the data set did not allow us to test field for significance.

For fields that are filled out for multiple persons on a form, the largest significant factor in the nonblank error rate is field category. There is a significant secondary contribution of Census 2000 regional census center. The structure of the data set did not allow us to test field and person number for significance.

Field categories that are high outliers occur in regional census centers 22, 23, 26, 27, 29, 30, and 32. The outlying categories are consistently Form Management and POP–Name. Form Management includes the contact information and person added/canceled fields on the enumerator forms. We find the outliers in this category are concentrated in the contact information fields. Fields for information on the addition or cancellation of persons do not appear. We find this encouraging.

Regional census centers 22, 23, 27, 29, and 32 span Florida, Los Angeles, and New York City. These are areas with above average concentrations of immigrants. Immigrants of non-European extraction tend to have names with unusual spellings. Limited English skills of first generation immigrants may lead to poor handwriting. Either condition could present a challenge to the automated technology and might account at least partly for high error rates in POP–Name fields from these regional census centers.

4.1.10 If we replace Census 2000 regional census centers with KFI impact in the factors in the above question, what can we say about the impact of KFI on soft match nonblank error rates for fields that went to KFI?

The possible ways KFI can affect fields going through it is

- it can improve our ability to capture respondent intent,
- it can worsen our ability to capture respondent intent,
- it can be redundant in two ways, and
- we may not be able in a specific case to determine an effect.

We want to be sure KFI does not improve our ability to capture intent at the cost of a higher soft match error rate. The results of including KFI impact in the analysis are covered in section 4.8.

For fields that are filled out for only one person on a form, the largest significant factor affecting the nonblank error rate is form. There is a significant secondary contribution of field category. The structure of the data set did not allow us to test field for significance. For fields that are filled out for multiple persons on a form, the largest significant factor affecting the nonblank error rate is the interaction of field and KFI impact. Interaction means that the effect of field will change depending on the impact of KFI. Field and KFI impact do not operate independently in their effect on the nonblank error rate. There are significant secondary contributions of form and field category. The structure of the data set did not allow us to test field and person number for significance.

We find no evidence KFI improves the capture of intent at the cost of higher soft match errors. There are clues to partly explain the interaction of field and KFI impact on the error rate. First, the most frequent category of KFI impact is "Cannot be determined". The automated technology rejected the content, and the entry keyed by the operator was not judged to be the respondent intent, character for character. Such content tends to be especially hard to interpret.

Second, many of the outliers on the d1s, the Spanish mailout/mailback short form, are for name fields. It is possible these outliers reflect limits on the capability of the automated technology to understand special Spanish language characters.

Third, many of the outliers on the d2, the English mailout/mailback long form, and the d2u, the English update/leave form, are for fields in which respondents write in a race or ethnicity other than the ones provided. This might reflect the increased challenge of interpreting characters written by hand instead of checked off in a box, especially when the handwriting is poor.

4.1.11 If we consider the same factors as in the above question but restrict ourselves to fields that were sent to KFI unnecessarily, which factors significantly affect in the nonblank KFI redundancy rate?

The KFI redundancy rate is the rate at which fields are sent to KFI unnecessarily. Since KFI redundancy can occur in two varieties, we want to include it as a fixed factor in our testing. This would answer whether the effect of the other factors on the KFI redundancy rate depends on which variety of redundancy is being considered. However all of the occurrences of KFI redundancy in our raw data are for only one variety. We cannot test for statistical significance of a fixed factor when it appears at only one level in the data set. Therefore, we do not include KFI redundancy as a factor.

We test form, field category, field, and person number for their effects on the nonblank KFI redundancy rate. The results are discussed in section 4.9. For fields that are filled out for only one person on a form, the only significant factor affecting the nonblank redundancy rate is field category. The structure of the data set did not allow us to test field for significance.

For fields that are filled out for multiple persons on a form, the largest significant factor affecting the nonblank redundancy rate is field category. There is a secondary significant association with form. The structure of the data set did not allow us to test field and person number for significance.

The category POP–Name is the only one flagged a high or very high outlier. The specific fields in the POP–Name category that are high or very high outliers are for forms d1s and d2u, specifically the middle initial for higher numbered persons.

While we do not propose it as the only explanation, respondent fatigue is a possible one for the POP–Name outliers. By the time respondents supply name information for the fifth or sixth person in a household, it is reasonable to suppose accuracy or neatness in the middle initial is not a high priority. Ideally, no field should be sent to KFI redundantly. For a field consisting of single character, it is not clear to us the benefits of achieving the ideal is worth the cost.

4.1.12 If we consider the same factors as in the above question but replace KFI impact with the Person 1 Race check-box field, what can we say about the impact of this race field on the nonblank hard match error rate?

The results of including the Person 1 race response in the analysis are discussed in section 4.10. Restricting ourselves to the Person 1 Race check-box field eliminates the factors of field category and person number. We are left with form and race response. Both significantly affect the nonblank hard match error rate. Of the two, the race response has the larger effect. Within our limited data set, we cannot find any error rates for specific race response fields that are outliers. The effect of race may be tied up with other factors that still need to be identified and tested.

4.1.13 What were the major reasons for failure to capture respondent intent?

The intent of the respondent was based on the judgement of analysts who examined the content of the forms after they were captured first by the automated technology and then by KFI. Sometimes the analysts concluded the captured responses misinterpreted what was meant. The ways and reasons for misinterpreting intent are analyzed in section 4.11. At the level of field, the high or very high outliers in terms of misinterpreting respondent intent are for the reason Extra check-box. Extra check-box occurs when the output from the automated technology output marks more check-boxes than are marked on the scanned image.

At the more general level of field category, the errors

- Extra characters (the output from the automated technology output shows more check-boxes marked than are on the scanned image),

- Missing characters (the output from the automated technology has fewer characters than the scanned image), and

- Wrong character (the output from the automated technology and the scanned image have the same number of characters, but the output from the automated technology disagrees with the scanned image in one or more characters)

appear in seven or nine of the 13 categories. These problems are not confined to a particular field or field category but rather exist across a wide swath. The major reasons for the errors are

- poor handwriting (the respondent's handwriting makes one letter look like another, but one can tell what the respondent meant),

- no reason found (the response is written clearly and there is nothing to suggest why it was
  not captured correctly), and

- rules not followed (the rules for keying the response after Census 2000 processing were not followed).

These reasons cut across the most forms and fields.

4.1.14 What is the best single number to sum up the performance of the automated data capture and imaging technology in Census 2000?

We have placed this question next to last rather than first because we believe any single number answer provides the least useful information for our readers. Given that some may desire one, we propose the probability that write-in fields are captured with no soft match errors and as the respondent intends. We feel this task is the most challenging one for the technology.

For the automated technology to capture write-ins as intended, it must first read any intelligible write-in content in the field. Second, once read, the write-in content must be accepted, that is not sent to KFI. Third, once accepted, the write-in content must capture the intent of the respondent. Fourth, once write-in intent is correctly captured, there must be no soft match errors.

We can write this as a chain of conditional probabilities:

Probability that write-in fields are captured with no soft match errors and as the respondent intends =

P(write-in content is read by the automated technology|write-in content exists in field) x P(write-in content is accepted by the automated technology|write-in content exists and is read) x P(automated technology captures intent correctly|write-in content exists, read, and is accepted) x P(no soft match error|have intended response; and write-in content exists, read, and is accepted). For convenience, we adopt the following symbols:

- A = write-in content is read in field and write-in content exists
- B = write-in content is read in field
- C = write-in content is accepted
- D = write-in content is read in field and write-in content exists
- E = technology correctly captures write-in content
- F = write-in content exists, is read, and is accepted
- G = no soft match error
- H = have intended response; and write-in content exists, is read, and is accepted

So we can rewrite the probability as P(A|B) x P(C|D) x P(E|F) x P(G|H).

We estimate P(A|B) in part by using of the file consisting of the cases in which the clerical evaluators determined the automated technology and KFI disagreed on content and by

1. taking the number of unduplicated write-in records in all of our data files,
2. taking the number of unduplicated write-in records in the file where the automated technology and KFI disagree and for which the error code is Blanked Response (see Table 43) and,
3. computing (1)/[(1)+(2)].

We estimate P(C|D) by

1. taking the number of unduplicated write-in records in our data files with a data capture mode of OCR (see section 4.5.2 for explanation),
2. taking the number of unduplicated write-in records in our data files and,
3. computing (1)/(2).

The value for (2) is the same as the value for the numerator in our estimate of P(A|B).

P(E|F) is the most uncertain quantity to estimate. The only records for which the analysts judged the intent of the respondent are the ones for which the content read by the automated technology disagreed with the content read by KFI. Unfortunately for our purpose, these are exactly the kind of records in which we should expect to find more than the usual proportion of cases that are hard to interpret under any technology, mechanical or human. We should estimate P(E|F) with cases reflecting a mix of low, moderate, and high difficulty of interpretation.

Besides judging the intent of the respondent, the analysts also judged whether the automated technology, KFI, or both failed to capture the intent of the respondent. This opens up a next best strategy for estimating P(E|F). We can focus on the subset of records for which the analysts concluded the automated technology was not responsible for failure to capture intent. We can

1. take the number of unduplicated records in the file where the automated technology and KFI disagree and for which the automated technology was not responsible for a failure to capture intent,

2. within the write-in records contained in (1) take the number which have a capture mode OCR, and

3. compute P(E|F) as (2)/(1).

The next best strategy has two drawbacks we should note:

1. The records used to estimate P(E|F) may still not reflect a balanced mix between cases of low, moderate, and high difficulty.

2. The records may be such a small sample that the estimate has poor precision.

We estimate P(G|H) by

1. taking the number of unduplicated write-in records in the file where the automated technology and KFI disagree and for which the automated technology was not responsible for a failure to capture intent,

2. taking the number of write-in records contained in (1) which have a capture mode OCR,
3. taking the write-in number of records contained in (2) without a soft match error according to the soft match algorithm (see Appendix G for an explanation), and

4. computing P(G|H) as (3)/(2).

The value for (2) is the same as the value for the numerator in our estimate of P(E|F). This strategy for estimating P(G|H) has the same two drawbacks noted above for P(E|F).

Substituting the appropriate values from our raw data, our best single number works out as follows. For P(A|B), P(write-in content is read in field and write-in content exists|write-in content is read in field),

- (1) = 31,523,300
- (2) = 1,614.

So, the estimate of P(A|B) = 31,523,300 / (31,523,300+ 1,614) = 0.999949.

For P(C|D), P(write-in content is accepted|write-in content is read in field and write-in content exists),

- (1) = 24,857,562 and
- (2) = 31,523,300.

So the estimate of P(C|D) = 24,857,562 / 31,523,300 = 0.788546.

For P(E|F), P(technology correctly captures write-in content|write-in content exists, read, and is accepted),

- (1) = 565,371 and
- (2) = 149,685.

So the estimate of P(E|F) = 149,685 / 565,371 = 0.264755.

For P(G|H), P(no soft match error|have intended response; and write-in content exists, read, and is accepted),

- (1) = 565,371,
- (2) = 149,685, and
- (3) = 59,808.

So the estimate of P(G|H) = 59,808 / 149,685 = 0.399559. Our estimate for the probability the automated technology will accept and capture write-in fields without soft match errors and as the respondent intends is 0.999949 x 0.788546 x 0.264755 x 0.399559 = 0.083412.

4.1.15 What are the implications of the probability the automated technology will accept and capture write-in fields as the respondent intends?

First, since we did not design this evaluation with the goal of generating this probability, we concede the strong likelihood of serious limitations with respect to our assumptions and precision in the preceding calculations.

Second, if there is intelligible content in a field, the automated technology will detect it with nearly perfect certainty.

Third, although the probability is lower than we would like, applying that probability over the many millions of responses in the decennial census still means a sizeable portion of those responses will be captured and interpreted correctly at speeds that are orders of magnitude above KFI. This opens up the possibility of more opportunity to focus human talent on responses that are particularly difficult to process.

Fourth, the largest impediment to automation is not the quality of the hardware or software, but the quality of the responses supplied by human beings. Misspelling, misplacement, and illegibility occur in too many variations and combinations for complete automation to be practical.

The preceding results suggest the future role of the automated technology reduces to two possibilities.

- The automated technology has a supporting role in decennial census processing. It is used to rapidly complete the clear and easy responses. Traditional methods claim the majority of resources for especially difficult responses.

- The automated technology has a dominant role in decennial census processing. Census forms are dramatically streamlined and redesigned to eliminate the long form's vast sea of handwritten responses requiring interpretation.

Which role automation will have depends on whether we retain the long form. So long as we gather huge quantities of write-in responses during the decennial population count, a supporting role is far more likely.

4.2 Overall Median Data Capture Error Rates

4.2.1 Contents of This Section

In this section, we show the median nonblank error rates with associated 90 percent confidence intervals. The details of the method for approximating the 90 percent confidence intervals are in Appendix E. The computational procedure for determining the median is described in Appendix F. The distinction between nonblank and total error rates is explained below. For definitions of common or special terms in this section, see the glossary in Appendix M.

To arrive at the median nonblank error rate for this section, we divide the data into two groups: one from enumerator-returned forms and the other from respondent-returned forms. The group respondent- returned consists of forms

- d1 (English mailout/mailback short form),
- d1s (Spanish mailout/mailback short form),
- d1u (English update/leave short form),
- d1ur (English update/leave short form for Puerto Rico),
- d2 (English mailout/mailback long form),
- d2s (Spanish mailout/mailback long form),
- d2u (English update/leave long form), and
- d2ur (English update/leave long form for Puerto Rico).

The group enumerator-returned consists of forms

- d1e (English enumerator short form),
- d1er (English enumerator short form for Puerto Rico),
- d2e (English enumerator long form), and
- d2er (English enumerator long form for Puerto Rico).

We collected the data for all the forms belonging to a particular group. We subgrouped the fields belonging to each form into thirteen categories. A list appears in Appendix B. We calculated nonblank error rates for all the fields comprising a field category. The median rates in Table Two below are the medians of all the field error rates for the various categories. For all the combinations in the table, the error rate consolidates both hard and soft match cases.

4.2.2 Calculation of the Hard and Soft Match Error Rates

To understand Table Two, it helps to understand how the error rates are calculated. We begin by reviewing the definition of hard and soft match errors from section 2.2. If the content of a check-box field is captured incorrectly by the automated technology or KFI, we have a hard match error. If the content of a write-in field is captured incorrectly, we have a soft match error.

We compare the Census 2000 context value against the evaluation truth value. The context value is the characters returned by the automated technology after special editing. The editing removes extra characters inserted by the automated technology that are needed to execute its program. The evaluation truth value is the content that was judged to be the most likely intent of the respondent. This judgement was performed by the clerical evaluators in Jeffersonville, IN, mentioned in section 2.1.

For check-box fields, we compare the context value to the truth value check-box by check-box. If the sequence of marked and unmarked check-boxes fails to match exactly, the context value is a hard match error. We do not compare check-box fields that are trailing blanks.

For write-in fields, we take all the characters in the context value and the truth value and count how many times each appears. Then we pass this information to the soft match algorithm to score the degree to which context and truth diverge. If the returned score exceeds a threshold, the context value is soft match error case. The algorithm does not count trailing blanks in the scoring.

Pseudocode for the soft match algorithm appears in Appendix G.

A field can be check-box or write-in but never both. So if any particular context value is in error, it is either a hard or soft match error but never both. We add up the number of fields for which the context value is in error. This is the numerator of the error rate.

We compute two error rates: nonblank and total. The denominator of the nonblank error rate is the number of times the automated technology read nonblank content for a field. The denominator for the total error rate is the number of times the automated technology read the field regardless of whether there was any content in it. In other words, it includes blank cases.

As long as blanks are occasional occurrences for a field, the nonblank and total error rates will be close. This is the case for the great majority of fields in this evaluation. Fields that are prone to large numbers of blanks will lead to large differences in the error rates. In this latter case, we believe the nonblank error rate is a better measure of data quality. The great bulk of the discussion in the results section of this evaluation focuses exclusively on the nonblank error rate.

While the automated technology should be given credit for reading blank fields correctly, this is not the same level of challenge as reading nonblank fields correctly. We compute the error rate as 100 x (numerator/denominator). The rates for Table Two are the nonblank error rates only.

Table 2. Median Data Capture Error Rates With Approximate 90 Percent Confidence Intervals, Nonblank Error Rates by Field Category Within Groupings of Respondent-Returned and Enumerator-Returned, Averaged Across All Capture Modes

Form Group	Field Category	Median Nonblank Data Capture Error Rate	Approximate 90% Lower Confidence Bound for Median	Approximate 90% Upper Confidence Bound for Median
Respondent-returned	POP--Military	8.940%	3.593%	13.889%
	POP--Ethnic	3.931%	3.309%	4.370%
	POP--Income	3.497%	3.188%	3.966%
	POP--Race	3.296%	2.593%	3.721%
	POP--Name	3.226%	2.889%	3.537%
	POP--Occupation	2.766%	2.459%	2.963%
	Housing Profile	1.835%	1.276%	2.128%
	POP--Education	1.389%	1.135%	1.633%
	POP--Demographic	1.161%	1.085%	1.244%
	POP--Disability	0.916%	0.737%	1.058%
Enumerator-returned	POP--Military	20.516%	6.607%	61.429%
	Form Management	2.931%	2.389%	3.777%
	POP--Income	2.620%	2.073%	3.232%
	POP--Occupation	2.445%	2.170%	2.728%
	Special Housing	2.301%	1.996%	3.545%
	POP--Name	1.967%	1.610%	2.158%
	POP--Education	1.759%	0.786%	3.372%
	Housing Profile	1.506%	1.373%	1.921%
	POP--Ethnic	1.354%	0.643%	1.692%
	POP--Demographic	0.986%	0.858%	1.213%
	POP--Race	0.872%	0.688%	0.998%
	POP--Disability	0.812%	0.684%	1.960%
	Coverage*			

*There were too few data points for the coverage category to compute valid overall rates and confidence intervals.

The grouping enumerator-returned contains three categories not found for forms in respondent-returned. These are Coverage, Form Management, and Special Housing. That is why there are no rows for these categories in the respondent-returned part of Table Two.

The confidence limits overlap between the two groupings for Housing Profile, POP–Demographic, POP–Disability, POP–Education, POP–Income, POP–Military, and POP–Occupation. There is a statistically significant lower median error rate for POP-Ethnic, POP–Name, and POP–Race in the Enumerator-returned grouping. Although not the source for the majority of data in Census 2000, it is helpful the enumerator-returned forms show lower error rates for the critical variables of ethnicity and race.

4.3 Median Data Capture Error Rates by Form / Field Category Combination

In this section, we break down the median nonblank error rates further. In the previous section, our break out was by field category. The data for each field category included multiple forms. The break out here is still by field category, but there are separate field category results for each individual form. Additionally, Table Three in this section shows

- the median rate recomputed by including blank cases,
- the total number of data records for a form / field category combination,
- the total number of data records in error,
- the number of blank data records, and
- whether the nonblank error rate can be considered a high or very high outlier.

An error rate is considered to be a high outlier if for all the field category by form combinations it exceeds the median rate by at least 1.5 times and by not more than 3.0 times the interquartile range. Very high outliers are any error rates that exceed the median by more than 3.0 times the interquartile range. More details concerning the calculation of outliers are described in Appendix F. For all the combinations reflected in Table Three, the error rate includes both hard and soft match cases. The details concerning the calculation of errors follows section 4.2.2. For definitions of common or special terms in this section, see the glossary in Appendix M.

High outliers appear in the field category POP-Name for forms

- d1e, the English enumerator short form,
- d1s, the Spanish mailout/mailback short form,
- d2e, the English enumerator long form, and
- d2ur, the English update/leave long form.

The form with the most high or very high outliers is d2ur. The automated technology finds it a challenge to read some of the names from enumerator-returned or Spanish language forms. Better enumerator training or Spanish form design may be needed. The update/leave process in Puerto Rico is another possible challenge, at least for name and ethnicity fields on long forms.

Table 3. Median Nonblank Data Capture Error Rates by Field Category Within Form, With Additional Statistics Including Outlier Status

Form	Name	Field Category	Median Nonblank Error Rate	Error Rate Recomputed With Blanks	Total Data Records	Total Blank Records	Total Data Capture Errors	Outlier
d1		POP--Name	2.191%	2.191%	1,699,662	0	37,247	
		POP--Race	0.829%	0.829%	622,807	0	5,160	
		POP--Ethnic	0.637%	0.637%	627,390	0	3,994	
		POP--Demographic	0.627%	0.627%	4,244,375	0	26,595	
		Housing Profile	0.236%	0.236%	233,461	0	551	

Form Name	Field Category	Median Nonblank Error Rate	Error Rate Recomputed With Blanks	Total Data Records	Total Blank Records	Total Data Capture Errors	Outlier
d1e	POP--Name	4.159%	4.159%	819,652	0	34,087	High
	Form Management	2.625%	2.625%	2,317,899	0	60,834	
	Special Housing	1.968%	1.968%	19,820	0	390	
	POP--Race	0.737%	0.737%	238,402	0	1,757	
	POP--Demographic	0.725%	0.724%	1,770,662	348	12,826	
	Housing Profile	0.563%	0.563%	297,767	0	1,677	
	POP--Ethnic	0.365%	0.365%	221,387	187	808	
	Coverage	0.196%	0.196%	169,838	0	333	
d1er	Form Management	0.062%	0.062%	33,664	0	21	
	POP--Name	0.006%	0.006%	16,399	0	1	
d1s	POP--Name	7.052%	7.052%	30,588	0	2,157	Very High
	POP--Ethnic	2.976%	2.976%	15,288	0	455	
	POP--Race	2.781%	2.781%	11,580	0	322	
	POP--Demographic	1.046%	1.046%	73,412	0	768	
	Housing Profile	0.341%	0.341%	2,637	0	9	
d1u	Housing Profile	2.156%	2.156%	75,125	0	1,620	
	POP--Name	1.921%	1.921%	293,754	0	5,643	
	POP--Demographic	0.778%	0.778%	777,536	184	6,047	
	POP--Race	0.465%	0.465%	105,021	0	488	
	POP--Ethnic	0.386%	0.386%	102,680	0	396	
d1ur	Housing Profile	0.168%	0.168%	6,564	0	11	
	POP--Race	0.129%	0.129%	8,535	0	11	
	POP--Demographic	0.080%	0.080%	63,622	23	51	
	POP--Name	0.009%	0.009%	21,907	0	2	
d2	POP--Name	2.890%	2.890%	2,221,784	83	64,205	
	POP--Ethnic	2.442%	2.439%	752,955	985	18,363	
	POP--Occupation	2.442%	2.440%	4,780,477	4,207	116,634	
	POP--Race	1.728%	1.726%	414,640	512	7,155	
	POP--Income	1.589%	1.589%	2,693,587	10	42,804	
	POP--Education	1.550%	1.550%	916,067	8	14,203	
	POP--Military	1.290%	1.290%	401,507	4	5,178	
	Housing Profile	1.239%	1.239%	3,462,423	17	42,906	
	POP--Demographic	1.073%	1.073%	6,981,177	34	74,874	
	POP--Disability	0.672%	0.672%	2,177,729	6	14,626	
d2e	POP–Name	4.626%	4.626%	1,727,650	0	79,919	High
	Form Management	3.848%	3.848%	3,500,832	0	134,710	High
	POP--Military	3.382%	3.382%	206,180	0	6,973	
	POP--Occupation	2.240%	2.237%	2,636,454	4,669	58,965	
	Special Housing	2.151%	2.151%	48,494	0	1,043	
	POP--Education	1.893%	1.893%	526,909	0	9,977	
	Housing Profile	1.456%	1.456%	2,544,749	0	37,047	
	POP--Demographic	1.234%	1.234%	4,483,270	500	55,306	
	POP--Income	1.011%	1.011%	1,385,314	0	14,011	
	POP--Disability	0.849%	0.849%	1,270,897	0	10,796	
	POP--Ethnic	0.800%	0.798%	497,327	680	3,971	
	Coverage	0.673%	0.673%	196,825	0	1,324	
	POP--Race	0.452%	0.452%	306,910	0	1,386	

Form Name	Field Category	Median Nonblan	Error Rate Recomput	Total Data	Total Blank	Total Data Capture	Outlier
d2er	Form Management	0.075%	0.075%	38,584	0	29	
	POP--Name	0.035%	0.035%	26,039	0	9	
	POP--Education	0.020%	0.020%	10,227	0	2	
	POP--Income	0.003%	0.003%	30,276	0	1	
	Housing Profile	0.002%	0.002%	44,948	0	1	
	POP--Demographic	0.001%	0.001%	83,433	35	1	
d2s	POP--Name	0.331%	0.331%	39,828	0	132	
	POP--Income	0.021%	0.021%	28,764	0	6	
	POP--Demographic	0.009%	0.009%	141,520	0	13	
	POP--Occupation	0.007%	0.007%	58,841	107	4	
	Housing Profile	0.006%	0.006%	34,194	0	2	
d2u	POP--Name	2.254%	2.254%	805,598	8	18,158	
	POP--Occupation	2.046%	2.044%	1,658,387	1,959	33,894	
	POP--Income	1.612%	1.612%	936,654	3	15,099	
	POP--Ethnic	1.489%	1.487%	251,583	337	3,741	
	Housing Profile	1.436%	1.436%	1,295,760	3	18,601	
	POP--Education	1.368%	1.368%	321,740	1	4,401	
	POP--Military	1.282%	1.282%	143,854	0	1,844	
	POP--Demographic	1.232%	1.232%	2,504,652	369	30,859	
	POP--Race	1.214%	1.213%	146,853	181	1,781	
	POP--Disability	0.864%	0.864%	777,995	0	6,722	
d2ur	Housing Profile	7.849%	7.849%	34,718	0	2,725	Very High
	POP--Ethnic	5.080%	5.064%	8,413	27	426	High
	POP--Name	4.548%	4.548%	14,599	0	664	High
	POP--Demographic	3.034%	3.034%	85,995	9	2,609	
	POP--Occupation	2.653%	2.647%	37,546	75	994	
	POP--Income	0.915%	0.915%	24,590	0	225	
	POP--Education	0.900%	0.900%	11,663	0	105	
	POP--Military	0.763%	0.763%	4,064	0	31	
	POP--Disability	0.386%	0.386%	25,671	0	99	
	POP--Race	0.143%	0.143%	4,898	8	7	

4.4 Analysis of Hard and Soft Match Error Rates for All Fields

4.4.1 Contents of This Section

In this section, we continue to a lower level of detail in analyzing our data. To understand our perspective here, consider the following:

- we have various decennial census forms: d1, d1e, d2, etc.
- each form has several categories of fields: name fields, race fields, etc.
- each field category contains several fields: names for person 1, person 2, etc.

When we count all the fields that exist in all the categories on all the forms, there are 810 in all. See Appendix C for a list. For definitions of common or special terms in this section, see the glossary in Appendix M.

Another factor entering our consideration in this section is the distinction between a person and a nonperson field. Examples of person fields are name fields, race fields, gender fields, and ethnicity fields. With person fields, there is space on the form to collect data for multiple persons in a household. So we have name, race, gender, and ethnicity information for person 1, person 2, person 3, and so on for a given household.

Examples of nonperson fields are the housing questions asked on the long forms. The members of the household are considered to live in a single dwelling. So we ask on each long form one question about the age of the house, how much of a mortgage there is on it, what the property taxes are, and so on. The important distinction then is whether the same information is gathered once or more than once on a given form.

Our basic question in this section is this: does the nonblank error rate vary in a significant way depending on what form, field category, or type of field we are talking about? To answer this question, we construct an analysis of variance (ANOVA) where the nonblank error rate is the response variable and the independent variables are form, field category, and field.

4.4.2 Factors and Models for Testing Statistical Significance

Our factors for testing statistical significance are form, field category, field, and the number of the person for which data being collected if we are dealing with a person field. We regard these factors as fixed. For more details about the significance testing, see Appendix J. We analyze nonperson fields for statistical significance separately from person fields. For nonperson fields, our model includes the variables

- field nested within field category and
- field category nested within form.

For person fields, our model includes the variables

- person number nested within field,
- field nested within field category, and
- field category nested within form.

We present four analyses:

- nonperson fields excluding all outliers
- nonperson fields including all outliers
- person fields excluding all outliers
- person fields including all outliers.

4.4.3 Significance Testing for Nonperson Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model." Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 4a. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	80	31096.34495	388.70431	21.80	**<0.0001**
Error	51	909.37477	17.83088		
Corrected Total	131	32005.71972			

Table 4b. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	271.77016	24.70638	1.39	0.2084
Field Category	10	48.35769	4.83577	0.27	0.9848
Field	54	22637.98677	419.22198	23.51	**<0.0001**

Table 5a. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers,

Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	86	100857.7439	1172.7645	49.12	**<0.0001**
Error	68	1623.5993	23.8765		
Corrected Total	154	102481.3433			

Table 5b. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	1326.82219	120.62020	5.05	**<0.0001**
Field Category	12	674.78183	56.23182	2.36	**0.0135**
Field	58	53353.47341	919.88747	38.53	**<0.0001**

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables do not agree as to which individual factors are significant. Since outliers are known to distort results, it is preferable to conclude based on excluding outliers. For nonperson fields, therefore, the only significant factor is associated with field. Form or field category are not significant.

4.4.4 Significance Testing for Person Fields

The notation and interpretation of the output in this section is also that of an ANOVA table. PROC GLM in SAS version 8.2 was also used to test for significance. The significance level for testing is also 10 percent.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 6a. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	728	116299.5814	159.7522	25.08	**<0.0001**
Error	1688	10753.1878	6.3704		
Corrected Total	2416	127052.7692			

Table 6b. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers,

Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	285.720042	28.572004	4.49	**<0.0001**
Field Category	48	2295.559258	47.824151	7.51	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			

Table 7a. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	805	163801.2463	203.4798	19.07	**<0.0001**
Error	2035	21708.9489	10.6678		
Corrected Total	2840	185510.1951			

Table 7b. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	546.465873	54.646587	5.12	**<0.0001**
Field Category	50	3232.208834	64.644177	6.06	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables agree as to which individual factors are significant. For person fields, the largest significant factor rate is field category. There is a significant secondary contribution of form. The structure of the data set did not allow SAS to test field and person number for significance.

4.4.5 Outlier Data for This Section

We have reached the first point in our analysis where the volume of data becomes an issue in table construction. As mentioned in section 4.4.1, we have 810 fields to consider. These fields exist on the twelve forms listed in Appendix A. When we calculate the nonblank error rate for all the fields available in our data, we have 2,996 rates by the time we are done. This is because the same field can appear on more than one form. Some of these rates–almost 450--are high or very high outliers according to the procedure discussed in section 4.3. How do we communicate what these outliers have to say without forcing the reader to wade through a 450 line table?

We think a fair compromise is to restrict the table to the outliers that are based on a reasonably large number of records. It is hard to conclude much when the data behind an outlier consists of two, three, or some other small number of records. After experimenting with different possibilities, we believe 500 records is a reasonable minimum to require. This results in Table Eight. It consists of 168 outliers. It covers eight of the twelve forms in our raw data. It provides insight into the highest six percent of the nonblank error rates. We believe this emphasizes problem fields that occur often enough to be a priority for investigation and improvement.

Table 8. Field Nonblank Error Rates that are High and Very High Outliers and Based on at Least 500 Blank and Nonblank Data Records

Form Name	Field Name	Description	Nonblank Error Rate	Total Nonblank	Outlier
d1	p3_relo	2 - Person 3: Other Relative	7.816%	3,288	High
d1e	p4ocancl	Person 4: Cancel	30.000%	750	Very High
	p5ocancl	Person 5: Cancel	26.423%	685	Very High
	rilast	Respondent's Last Name	11.212%	131,961	High
	rifirst	Respondent's First Name	8.003%	133,156	High
d1s	p5mi	Person 5: Middle Initial	10.667%	600	High
	p4mi	Person 4: Middle Initial	10.226%	929	High
	p2hisp19	Person 2: Other Hispanic Origin	9.931%	1,017	High
	p1hisp19	Person 1: Other Hispanic Origin	9.930%	1,138	High
	p3mi	Person 3: Middle Initial	9.744%	1,211	High
	p1mi	Person 1: Middle Initial	9.196%	1,555	High
	p2mi	Person 2: Middle Initial	9.155%	1,409	High
	p1last	Person 1: Last Name	7.892%	2,699	High
	p1trib19	Person 1: Am. Indian, AK Native Tribe	7.843%	612	High
	p2last	Person 2: Last Name	7.677%	2,449	High
d1u	p1apt16a	Apartment Number	8.801%	3,136	High
d2	p4trib_1	Person 4: Am Indian, Alaska Native Tribe	30.460%	1,218	Very High
	p2trib_1	Person 2: Am Indian, Alaska Native Tribe	29.838%	2,785	Very High
	p3trib_1	Person 3: Am Indian, Alaska Native Tribe	28.197%	1,947	Very High
	p2asia_1	Person 2: Other Asian	27.814%	2,301	Very High
	p6oetype	Person 6: Class of Worker	27.167%	946	Very High
	p1asia_1	Person 1: Other Asian	26.512%	2,199	Very High
	p5hisp_1	Person 5: Other Hispanic Origin	25.896%	977	Very High
	p1trib_1	Person 1: Am Indian, Alaska Native Tribe	24.805%	2,689	Very High
	p5trib_1	Person 5: Am Indian, Alaska Native Tribe	24.662%	665	Very High
	p3asia_1	Person 3: Other Asian	24.506%	1,469	Very High

Form Name	Field Name	Description	Nonblank Error Rate	Total Nonblank	Outlier
d2	p5asia_1	Person 5: Other Asian	23.689%	591	Very High
	p3hisp_1	Person 3: Other Hispanic Origin	22.724%	2,614	Very High
	p4asia_1	Person 4: Other Asian	22.070%	947	Very High
	p1hisp_1	Person 1: Other Hispanic Origin	20.980%	4,428	Very High
	p6hisp_1	Person 6: Other Hispanic Origin	20.598%	602	Very High
	p2race_1	Person 2: Other Race	20.458%	4,414	Very High
	p3race_1	Person 3: Other Race	20.427%	2,952	Very High
	p2hisp_1	Person 2: Other Hispanic Origin	20.423%	3,829	Very High
	p4race_1	Person 4: Other Race	19.355%	2,046	Very High
	p5race_1	Person 5: Other Race	19.292%	1,187	Very High
	p6race_1	Person 6: Other Race	18.155%	672	Very High
	p6otrans	Person 6: Work Vehicle	17.318%	716	Very High
	p1race_1	Person 1: Other Race	16.792%	4,913	Very High
	p6otype	Person 6: Business Type	16.351%	740	Very High
	p1ointls	Person 1: Interest Loss	15.696%	1,357	Very High
	p6owork	Person 6: Work Last Year	15.392%	1,085	Very High
	p4_relo	Person 4: Other Relative	14.503%	1,248	Very High
	p5_relo	Person 5: Other Relative	14.041%	933	Very High
	p3addr_1	Person 3: Work Address	13.892%	12,907	Very High
	p2oresp	Person 2: How Long	13.639%	1,745	Very High
	p1addr_1	Person 1: Work Address	13.637%	91,310	Very High
	p6oam_pm	Person 6: Time to Work am/pm	13.468%	594	Very High
	p1ototls	Person 1: Total Income Loss	13.432%	1,489	Very High
	p2ototls	Person 2: Total Income Loss	13.427%	782	Very High
	p4addr_1	Person 4: Work Address	13.249%	4,091	Very High
	p5addr_1	Person 5: Work Address	12.950%	1,390	High
	p2addr_1	Person 2: Work Address	12.520%	56,468	High
	p3_relo	Person 3: Other Relative	12.316%	2,111	High
	p6addr_1	Person 6: Work Address	12.018%	649	High
	p1oresp	Person 1: How Long	11.781%	2,886	High
	p3oserve	Person 3: When on Active Duty	11.749%	1,115	High
	p6oint	Person 6: Interest	11.352%	1,427	High
	p6_relo	Person 6: Other Relative	11.079%	686	High
	p6oride	Person 6: Carpool	10.400%	500	High
	p2oslfls	Person 2: Self- Person 2:employment Loss	10.009%	1,119	High
	p1ssi	Person 1: SSI Amount	9.941%	7,605	High
	p6olayof	Person 6: Last Week Layoff	9.885%	1,133	High
	p6omilit	Person 6: Active Duty	9.699%	1,629	High
	p2_relo	Person 2: Other Relative	9.208%	4,746	High
	p3selfe	Person 3: Self Employment Income Amount	9.138%	1,160	High
	p4empl_1	Person 4: Employer	9.013%	5,625	High
	p2welfr	Person 2: Welfare Amount	8.875%	2,107	High
	p6octlmt	Person 6: Work Inside City Limits	8.859%	587	High
	p1welfr	Person 1: Welfare Amount	8.813%	4,346	High
	p1oslfls	Person 1: Self- Person 1:employment Loss	8.756%	2,501	High
	p6empl_1	Person 6: Employer	8.701%	816	High
	p2ssi	Person 2: SSI Amount	8.653%	3,733	High

Form Name	Field Name	Description	Nonblank Error Rate	Total Nonblank	Outlier
d2	p5otype	Person 5: Business Type	8.405%	1,749	High
	p2_other	Person 2: Other Income Amount	8.052%	5,340	High
	p1ograde	Person 1: Grade Level	8.002%	29,005	High
	p2oserve	Person 2: When on Active Duty	7.838%	4,950	High
	p5oride	Person 5: Carpool	7.832%	1,264	High
	p6oabsnt	Person 6: Last Week Absent	7.778%	990	High
	p5empl_1	Person 5: Employer	7.713%	1,828	High
	p3_other	Person 3: Other Income Amount	7.698%	1,299	High
	p1oarmed	Person 1: Armed Forces	7.677%	1,485	High
	p3welfr	Person 3: Welfare Amount	7.549%	861	High
d2e	p5oresp	Person 5: How Long	91.362%	903	Very High
	p3oresp	Person 3: How Long	86.052%	889	Very High
	p4oserve	Person 4: When on Active Duty	82.660%	1,782	Very High
	p2ototls	Person 2: Total Income Loss	74.372%	597	Very High
	p5ostart	Person 5: Could Start Last Week	57.649%	1,072	Very High
	p2oresp	Person 2: How Long	47.550%	1,918	Very High
	p5oneeds	Person 5: Responsible for Needs	44.915%	944	Very High
	p5oetype	Person 5: Class of Worker	44.375%	2,889	Very High
	p3ocancl	Person 3: Cancel	41.379%	522	Very High
	p1ocancl	Person 1: Cancel	39.893%	559	Very High
	p4otrans	Person 4: Work Vehicle	38.287%	5,242	Very High
	p1oarmed	Person 1: Armed Forces	37.452%	526	Very High
	p3oneeds	Person 3: Responsible for Needs	33.091%	3,025	Very High
	p1oadd	Person 1: Add	31.919%	542	Very High
	p3oserve	Person 3: When on Active Duty	29.475%	648	Very High
	p5otype	Person 5: Business Type	25.698%	2,043	Very High
	p1oslfls	Person 1: Self- Person 1:employment Loss	24.769%	650	Very High
	p5oborn	Person 5: Under 19	21.534%	1,291	Very High
	p4oride	Person 4: Carpool	19.620%	3,155	Very High
	rilast	Respondent's Last Name	17.553%	166,557	Very High
	p5owork	Person 5: Work Last Year	16.733%	2,008	Very High
	p5olook	Person 5: Looking for Work	16.530%	2,196	Very High
	p3ostart	Person 3: Could Start Last Week	15.497%	3,091	Very High
	p5otrans	Person 5: Work Vehicle	14.167%	1,447	Very High
	p5olstwk	Person 5: Last Worked	13.590%	2,156	Very High
	p5olvcty	Person 5: Live Inside City Limits	13.231%	3,847	Very High
	rifirst	Respondent's First Name	12.221%	168,452	High
	p1oserve	Person 1: When on Active Duty	11.557%	13,654	High
	p1omort	Household: No Payment	11.427%	1,724	High
	p3oyears	Person 3: Years on Active Duty	11.004%	518	High
	p3oborn	Person 3: Under 17	10.708%	5,267	High
	p4ostart	Person 4: Could Start Last Week	10.705%	1,205	High
	p1oresp	Person 1: How Long	10.240%	2,002	High
	p1stx16a	Street Name	9.958%	33,361	High
	p4orecal	Person 4: Will Be Recalled	9.873%	1,104	High
	p1ograde	Person 1: Grade Level	9.724%	8,176	High
	p2oserve	Person 2: When on Active Duty	9.304%	2,859	High

Form Name	Field Name	Description	Nonblank Error Rate	Total Nonblank	Outlier
d2e	p5ograndd	Person 5: Grandchildren	9.279%	3,233	High
	p3oetype	Person 3: Class of Worker	9.269%	11,781	High
	p3orecal	Person 3: Will Be Recalled	8.655%	2,126	High
	p4oam_pm	Person 4: Time to Work am/pm	8.432%	2,965	High
	p2ostart	Person 2: Could Start Last Week	8.375%	6,209	High
	p2oneeds	Person 2: Responsible for Needs	8.318%	7,838	High
d2u	p1asia_1	Person 1: Other Asian	22.016%	486	Very High
	p2asia_1	Person 2: Other Asian	19.083%	545	Very High
	p2trib_1	Person 2: Am Indian, Alaska Native Tribe	17.576%	990	Very High
	p2race_1	Person 2: Other Race	16.018%	899	Very High
	p3trib_1	Person 3: Am Indian, Alaska Native Tribe	15.949%	627	Very High
	p2hisp_1	Person 2: Other Hispanic Origin	15.326%	783	Very High
	p3hisp_1	Person 3: Other Hispanic Origin	14.865%	518	Very High
	p1trib_1	Person 1: Am Indian, Alaska Native Tribe	14.690%	953	Very High
	p1hisp_1	Person 1: Other Hispanic Origin	14.491%	904	Very High
	p1race_1	Person 1: Other Race	12.879%	924	High
	p3race_1	Person 3: Other Race	12.868%	544	High
	p1stx16a	Street Name	10.123%	29,874	High
	p1oelec	Household: Electricity	9.316%	1,535	High
	p1addr_1	Person 1: Work Address	9.281%	31,150	High
	p3_relo	Person 3: Other Relative	9.241%	606	High
	p1oslfls	Person 1: Self- Person 1:employment Loss	9.163%	1,899	High
	p2oresp	Person 2: How Long	8.696%	690	High
	p4addr_1	Person 4: Work Address	8.658%	1,155	High
	p2addr_1	Person 2: Work Address	8.563%	21,475	High
	p1welfr	Person 1: Welfare Amount	8.516%	1,503	High
	p1apt16a	Apartment Number	8.482%	4,374	High
	p3addr_1	Person 3: Work Address	8.261%	4,370	High
	p2oslfls	Person 2: Self- Person 2:employment Loss	8.052%	621	High
	p1ssi	Person 1: SSI Amount	8.046%	3,157	High
d2ur	p1oauto	Household: Number of Automobiles	72.310%	1,589	Very High
	p1obdrm	Household: Number of Bedrooms	71.420%	1,578	Very High
	p1lang	Person 1: Language	48.247%	1,198	Very High
	p3lang	Person 3: Language	46.006%	626	Very High
	p2lang	Person 2: Language	45.511%	958	Very High
	p1stx16a	Street Name	19.272%	1,126	Very High
	p1addr_1	Person 1: Work Address	18.474%	498	Very High
	p1hsn10a	House Number	12.796%	719	High
	p2last	Person 2: Last Name	9.111%	1,383	High
	p4ohisp	Person 4: Hispanic Origin	9.007%	544	High
	p3last	Person 3: Last Name	9.000%	900	High
	p2ohisp	Person 2: Hispanic Origin	8.676%	1,360	High
	p1actv_1	Person 1: Industry	8.380%	716	High
	p3ohisp	Person 3: Hispanic Origin	8.241%	898	High
	p4last	Person 4: Last Name	7.871%	559	High
	p1empl_1	Person 1: Employer	7.796%	744	High
	p2lvcity	Person 2: Migration City	7.769%	502	High
	p1lvcity	Person 1: Migration City	7.750%	671	High

As a first attempt to understand Table Eight further, we analyze the distribution by form type, form name, and person number. Details are in Appendix H. The analysis shows form d2, the English mailout/mailback long form, has a statistically greater presence in Table Eight than would be expected from its distribution in the entire group of 2,996 error rates. Further investigation should begin with this form.

4.5 Analysis of Individual Hard and Soft Match Error Rates By Data Capture Mode

4.5.1 Contents of This Section

In this section, we use a new grouping of the data called data capture mode to analyze the hard match and soft match error rates. In the previous section, we were concerned about how the nonblank error rate behaved depending on

- form (whether we are dealing with a d1, d2, etc.),
- field category (whether we are dealing with name fields, race fields, etc.), and
- field (whether we are dealing with name data for person 1, person 2, etc).

Our basic question in this section is this: does the nonblank error rate vary in a significant way depending on what form, field category, type of field, and data capture mode we are talking about? To answer this question, we construct an analysis of variance (ANOVA) where the nonblank error rate is the response variable and the independent variables are form, field category, field, and data capture mode.

In this section, we also distinguish between person and nonperson fields as discussed in section 4.4.1. For definitions of common or special terms in this section, see the glossary in Appendix M.

An explanation of data capture mode follows in section 4.5.2. After the ANOVA, we show Tables 13 and 14. The data for the tables are the same as for the ANOVA. After going through the different combinations of forms, fields, and data capture modes, we have a raw data set consisting of 4,308 hard and soft match error rates for the ANOVA and the tables.

In Table 13, we show nonblank error rates that are outliers for specific fields on specific forms. We aim for a sufficiently fine level of detail that makes it easy to identify the largest improvement opportunities.

Table 14 complements Table 13. We aim for a higher level of detail that supports a meaningful overall view of the data. We show the nonblank error rates for each field category. We show a separate field category result for each of the three modes of data capture. Any outliers in Table 14 identify field categories that stand out in terms of a high error rate.

The method for testing statistical significance follows sections 4.4.3 and 4.4.4. The details concerning the calculation of errors follows section 4.2.2. The rules concerning the determination of outliers is as described in section 4.3.

4.5.2 Explaining the Modes of Data Capture, OCR, OMR, and KFI

The three modes of data capture are OCR, OMR, and KFI. To understand these modes, we share more information about Census 2000 processing. After capturing the content for a field, the automated technology calculated a measure called a confidence level. The confidence level was the technology's estimate of the probability that it had captured intelligible content. While spaces does not allow us to explain in detail, in broad terms an algorithm compared the electronic patterns of the content with a stored library of patterns and looked for matches between the two.

The technology was programmed to reject content whose associated confidence level failed to meet a minimum threshold. In these cases, the fall back procedure was for a human operator to look at the scanned image of the form and key in an entry manually. In other words, KFI was used.

As a general rule, the content whose confidence level met or exceeded the threshold was accepted by the automated technology. Some fields went directly to KFI regardless of the confidence level. These were check-box fields where more than one box could be selected and still count as a valid response.

After being accepted, content advanced to the next field. So the first thing to understand about data capture modes is that the raw data for this evaluation are split between cases that met the threshold and cases that did not.

The cases that met or exceeded the threshold form two categories of data capture mode. If a successful case is for a check-box field, the mode is OMR. OMR stands for "optical mark recognition." If a successful case is for a write-in field, the mode is OCR. OCR stands for "optical character recognition."

The cases failing the threshold form the third category of data capture mode. A standard term for this category did not emerge during Census 2000 processing. Since the fall back procedure used KFI, the tendency was to adopt this term for convenience of description.

We follow this practice in this evaluation. To distinguish KFI from the independent keying of our predetermined sample of forms *after* Census 2000 processing, we use the term MIK for the latter. MIK stands for "manual inspection and keying." We believe this designation captures the essence of what happened to the content rejected by the automated technology.

Before proceeding with the analysis, we emphasize and reiterate some useful points. First, the same operation applied during Census 2000 processing to handle rejected content as applied afterwards in part of the creation of our raw data. A human being looked at a scanned image of a form and keyed in what he or she saw.

Second, this means some of the fields in our raw data were keyed twice: once during Census 2000 processing and once afterwards. Any remedial keying during processing is independent of the keying that took place after processing. The two keyings were performed by different groups of people who did not have a chance to interact and affect each other's work.

Third, the three modes of data capture permit us to analyze the fields that were keyed twice separately from those that were keyed once. We are in a position to check for consistency of conclusions between the two situations.

Finally, to understand the general performance of the automated technology for hard match error rates, refer to Table 14 in this section under OMR mode. For soft match error performance, refer to the OCR mode section of Table 14. The general performance for content rejected by the automated technology and keyed by a human operator can be found in the KFI section.

4.5.3 Factors and Models for Testing Statistical Significance

Our factors for testing statistical significance are mode, form, field, field category, and person number. We regard these factors as fixed. For more details about the significance testing, see Appendix J.

We analyze nonperson fields for statistical significance separately from person fields. For nonperson fields, our model is

- field nested within field category,
- field category nested within form, and
- mode crossed with field.

For person fields, our model is

- person number nested within field,
- field nested within field category,
- field category nested within form, and
- mode crossed with field.

We present four analyses:

- nonperson fields excluding all outliers
- nonperson fields including all outliers
- person fields excluding all outliers
- person fields including all outliers.

4.5.4 Significance Testing for Nonperson Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 9a. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	95	46152.58824	485.81672	31.70	**<0.0001**
Error	74	1134.08073	15.32542		
Corrected Total	169	47286.66897			

Table 9b. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	372.9771795	33.9070163	2.21	**0.0223**
Field Category	9	58.7980470	6.5331163	0.43	0.9169
Field	NA	NA			
Mode	1	6.0143276	6.0143276	0.39	0.5329
Field*Mode	12	69.1829862	5.7652489	0.38	0.9680

Table 10a. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	103	102692.7191	997.0167	52.82	**<0.0001**
Error	88	1661.1823	18.8771		
Corrected Total	191	104353.9014			

Table 10b. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	838.7812075	76.2528370	4.04	**<0.0001**
Field Category	12	507.5054506	42.2921209	2.24	**0.0161**
Field	NA	NA			
Mode	1	0.2792463	0.2792463	0.01	0.9035
Field*Mode	16	74.7559615	4.6722476	0.25	0.9986

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables do not agree as to which individual factors are significant. Since outliers are known to distort results, it is preferable to conclude based on excluding outliers. For nonperson fields, therefore, the only significant factor is form. There is no significant contribution of field category, mode, or the interaction of field and mode. The structure of the data set did not allow SAS to test field for significance.

4.5.5 Significance Testing for Person Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance..

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 11a. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1087	175927.1447	161.8465	35.47	**<0.0001**
Error	2514	11470.1455	4.5625		
Corrected Total	3601	187397.2902			

Table 11b. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	239.921265	23.992127	5.26	**<0.0001**
Field Category	48	1802.527318	37.552652	8.23	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			
Mode	2	2335.898722	1167.949361	255.99	**<0.0001**
Field*Mode	345	4247.311096	12.311047	2.70	**<0.0001**

Table 12a. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1161	233264.9021	200.9172	26.32	**<0.0001**
Error	2954	22551.4161	7.6342		
Corrected Total	4115	255816.3182			

Table 12b. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	513.783000	51.378300	6.73	**<0.0001**
Field Category	50	2667.128153	53.342563	6.99	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			
Mode	2	385.085264	192.542632	25.22	**<0.0001**
Field*Mode	354	5627.312804	15.896364	2.08	**<0.0001**

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables agree as to which individual factors are significant. For person fields, the largest significant factor is the interaction of field and mode. Interaction means that the effect of field will change depending on the mode. The field and mode do not operate independently in their effect on the nonblank error rate. There is a significant secondary contribution of field category. The structure of the data set did not allow SAS to test field and person number for significance.

4.5.6 Outlier Data for This Section

We have reached another point in our analysis where the volume of data becomes an issue in table construction. As mentioned in section 4.5.1, when we calculate the nonblank error rate for all the combinations of variables relevant to this analysis, we have 4,308 rates by the time we are done. Some of these rates–almost 550--are high or very high outliers according to the procedure discussed in section 4.3. How do we communicate what these outliers have to say without forcing the reader to wade through a 550 line table?

We think a fair compromise is to restrict the table to the outliers that are based on a reasonably large number of records. It is hard to conclude much when the data behind an outlier consist of two, three, or some other small number of records. After experimenting with different possibilities, we believe 500 records is a reasonable minimum to require. This results in Table 13. It consists of 149 outliers. It provides insight into the highest three percent of the nonblank error rates. We believe this emphasizes problem fields that occur often enough to be a priority for investigation and improvement.

Table 13. Field Nonblank Error Rates that are High and Very High Outliers and Based on at Least 500 Blank and Nonblank Data Records

Form Name	Field Name	Description	Mode	Nonblank Error %	Total Nonblank Records	Outlier
d1	p3_relo	Person 3: Other Relative	KFI	11.284%	2,118	High
	p2_relo	Person 2: Other Relative	KFI	10.160%	1,880	High
	p4_relo	Person 4: Other Relative	KFI	9.517%	2,028	High
d1e	p4ocancl	Person 4: Cancel	OMR	30.000%	750	Very High
	p5ocancl	Person 5: Cancel	OMR	26.423%	685	Very High
	rilast	Respondent's Last Name	OCR	11.212%	131,961	High
	p2orace	Person 2: Race	KFI	10.673%	1,649	High
	p3orace	Person 3: Race	KFI	10.173%	1,563	High
d1s	p1mi	Person 1: Middle Initial	KFI	21.333%	525	Very High
	p1hisp19	Person 1: Other Hispanic Origin	KFI	13.993%	536	High
	p1last	Person 1: Last Name	KFI	13.875%	1,009	High
	p4last	Person 4: Last Name	KFI	13.854%	628	High
	p2last	Person 2: Last Name	KFI	12.603%	968	High
	p3last	Person 3: Last Name	KFI	11.442%	874	High
	p3hisp19	Person 3: Other Hispanic Origin	KFI	10.558%	502	High
d1u	p1hsn10a	House Number	KFI	16.177%	3,950	High
d2	p4trib_1	Person 4: Am Indian, Alaska Native Tribe	OCR	30.460%	1,218	Very High
	p2trib_1	Person 2: Am Indian, Alaska Native Tribe	OCR	29.838%	2,785	Very High
	p3trib_1	Person 3: Am Indian, Alaska Native Tribe	OCR	28.197%	1,947	Very High
	p2asia_1	Person 2: Other Asian	OCR	27.814%	2,301	Very High
	p6oetype	Person 6: Class of Worker	OMR	27.167%	946	Very High
	p1asia_1	Person 1: Other Asian	OCR	26.512%	2,199	Very High
	p5hisp_1	Person 5: Other Hispanic Origin	OCR	25.896%	977	Very High
	p1trib_1	Person 1: Am Indian, Alaska Native Tribe	OCR	24.805%	2,689	Very High

Form Name	Field Name	Description	Mode	Nonblank Error %	Total Nonblank Records	Outlier
d2	p5trib_1	Person 5: Am Indian, Alaska Native Tribe	OCR	24.662%	665	Very High
	p3asia_1	Person 3: Other Asian	OCR	24.506%	1,469	Very High
	p5asia_1	Person 5: Other Asian	OCR	23.689%	591	Very High
	p4hisp_1	Person 4: Other Hispanic Origin	OCR	23.543%	1,699	Very High
	p3hisp_1	Person 3: Other Hispanic Origin	OCR	22.724%	2,614	Very High
	p4_relo	Person 4: Other Relative	KFI	22.343%	734	Very High
	p4asia_1	Person 4: Other Asian	OCR	22.070%	947	Very High
	p1hisp_1	Person 1: Other Hispanic Origin	OCR	20.980%	4,428	Very High
	p6hisp_1	Person 6: Other Hispanic Origin	OCR	20.598%	602	Very High
	p2race_1	Person 2: Other Race	OCR	20.458%	4,414	Very High
	p3race_1	Person 3: Other Race	OCR	20.427%	2,952	Very High
	p2hisp_1	Person 2: Other Hispanic Origin	OCR	20.423%	3,829	Very High
	p5_relo	Person 5: Other Relative	KFI	20.000%	605	Very High
	p4race_1	Person 4: Other Race	OCR	19.355%	2,046	Very High
	p5race_1	Person 5: Other Race	OCR	19.292%	1,187	Very High
	p6race_1	Person 6: Other Race	OCR	18.155%	672	Very High
	p3_relo	Person 3: Other Relative	KFI	17.922%	1,328	Very High
	p6otrans	Person 6: Work Vehicle	OMR	17.318%	716	Very High
	p1race_1	Person 1: Other Race	OCR	16.792%	4,913	Very High
	p6otype	Person 6: Business Type	OMR	16.351%	740	High
	p1ointls	Person 1: Interest Loss	OMR	15.696%	1,357	High
	p6owork	Person 6: Work Last Year	OMR	15.392%	1,085	High
	p3addr_1	Person 3: Work Address	KFI	13.892%	12,907	High
	p3selfe	Person 3: Self Employment Income Amount	KFI	13.826%	745	High
	p2_other	Person 2: Other Income Amount	KFI	13.663%	2,869	High
	p2oresp	Person 2: How Long	OMR	13.639%	1,745	High
	p1addr_1	Person 1: Work Address	KFI	13.637%	91,310	High
	p6oam_pm	Person 6: Time to Work am/pm	OMR	13.468%	594	High
	p1ototls	Person 1: Total Income Loss	OMR	13.432%	1,489	High
	p2ototls	Person 2: Total Income Loss	OMR	13.427%	782	High
	p4addr_1	Person 4: Work Address	KFI	13.249%	4,091	High
	p1ssi	Person 1: SSI Amount	KFI	13.068%	5,081	High
	p1_other	Person 1: Other Income Amount	KFI	13.052%	6,681	High
	p5addr_1	Person 5: Work Address	KFI	12.950%	1,390	High
	p2ssi	Person 2: SSI Amount	KFI	12.672%	2,320	High
	p1yrmvus	Person 1: Migration Year	KFI	12.547%	4,264	High
	p2addr_1	Person 2: Work Address	KFI	12.520%	56,468	High
	p6addr_1	Person 6: Work Address	KFI	12.018%	649	High
	p1welfr	Person 1: Welfare Amount	KFI	11.976%	2,789	High
	p1oresp	Person 1: How Long	OMR	11.781%	2,886	High
	p3oserve	Person 3: When on Active Duty	OMR	11.749%	1,115	High
	r1last	Roster: Person 1 Last Name	KFI	11.515%	58,706	High
	p2welfr	Person 2: Welfare Amount	KFI	11.503%	1,504	High
	p6oint	Person 6: Interest	OMR	11.352%	1,427	High
	p2selfe	Person 2: Self Employment Income Amount	KFI	11.231%	3,437	High
	p1selfe	Person 1: Self Employment Income Amount	KFI	11.127%	6,920	High
	p2_relo	Person 2: Other Relative	KFI	11.114%	3,302	High
	p4empl_1	Person 4: Employer	KFI	11.097%	3,956	High

Form Name	Field Name	Description	Mode	Nonblank Error %	Total Nonblank Records	Outlier
d2	p3_other	Person 3: Other Income Amount	KFI	10.497%	886	High
	r2last	Roster: Person 2 Last Name	KFI	10.477%	41,376	High
	p6oride	Person 6: Carpool	OMR	10.400%	500	High
	p1last	Person 1: Last Name	KFI	10.032%	60,464	High
	p2oslfls	Person 2: Self- Person 2:employment Loss	OMR	10.009%	1,119	High
	p6empl_1	Person 6: Employer	KFI	9.907%	646	High
	p6olayof	Person 6: Last Week Layoff	OMR	9.885%	1,133	High
	p2yrmvus	Person 2: Migration Year	KFI	9.770%	3,787	High
	r3last	Roster: Person 3 Last Name	KFI	9.751%	23,484	High
	p5empl_1	Person 5: Employer	KFI	9.714%	1,328	High
	p6omilit	Person 6: Active Duty	OMR	9.699%	1,629	High
	p1retir	Person 1: Retirement Income Amount	KFI	9.690%	10,206	High
	p3yrmvus	Person 3: Migration Year	KFI	9.681%	2,665	High
d2e	p5oresp	Person 5: How Long	OMR	91.362%	903	Very High
	p3oresp	Person 3: How Long	OMR	86.052%	889	Very High
	p4oserve	Person 4: When on Active Duty	OMR	82.660%	1,782	Very High
	p2ototls	Person 2: Total Income Loss	OMR	74.372%	597	Very High
	p5ostart	Person 5: Could Start Last Week	OMR	57.649%	1,072	Very High
	p2oresp	Person 2: How Long	OMR	47.550%	1,918	Very High
	p5oneeds	Person 5: Responsible for Needs	OMR	44.915%	944	Very High
	p3ocancl	Person 3: Cancel	OMR	41.379%	522	Very High
	p1ocancl	Person 1: Cancel	OMR	39.893%	559	Very High
	p4otrans	Person 4: Work Vehicle	OMR	38.287%	5,242	Very High
	p1oarmed	Person 1: Armed Forces	OMR	37.452%	526	Very High
	p3oneeds	Person 3: Responsible for Needs	OMR	33.091%	3,025	Very High
	p1oadd	Person 1: Add	OMR	31.919%	542	Very High
	p3oserve	Person 3: When on Active Duty	OMR	29.475%	648	Very High
	p5otype	Person 5: Business Type	OMR	25.698%	2,043	Very High
	p1oslfls	Person 1: Self- Person 1:employment Loss	OMR	24.769%	650	Very High
	p5oborn	Person 5: Under 19	OMR	21.534%	1,291	Very High
	p4oride	Person 4: Carpool	OMR	19.620%	3,155	Very High
	rilast	Respondent's Last Name	OCR	17.555%	166,529	Very High
	p5owork	Person 5: Work Last Year	OMR	16.733%	2,008	Very High
	p5olook	Person 5: Looking for Work	OMR	16.530%	2,196	High
	p3ostart	Person 3: Could Start Last Week	OMR	15.497%	3,091	High
	p5otrans	Person 5: Work Vehicle	OMR	14.167%	1,447	High
	p5olstwk	Person 5: Last Worked	OMR	13.590%	2,156	High
	p5olvcty	Person 5: Live Inside City Limits	OMR	13.231%	3,847	High
	rifirst	Respondent's First Name	OCR	12.222%	168,443	High
	p1oserve	Person 1: When on Active Duty	OMR	11.557%	13,654	High
	p1omort	Household: No Payment	OMR	11.427%	1,724	High
	p3oyears	Person 3: Years on Active Duty	OMR	11.004%	518	High
	p1zip5a	Zip Code	KFI	10.780%	5,575	High
	p3oborn	Person 3: Under 17	OMR	10.708%	5,267	High
	p4ostart	Person 4: Could Start Last Week	OMR	10.705%	1,205	High
	p3_relo	Person 3: Other Relative	KFI	10.649%	601	High
	p5lvzip	Person 5: Migration Zip Code	KFI	10.626%	1,007	High
	p1oresp	Person 1: How Long	OMR	10.240%	2,002	High
	p1stx16a	Street Name	KFI	9.958%	33,361	High
	p4orecal	Person 4: Will Be Recalled	OMR	9.873%	1,104	High
	p1ograde	Person 1: Grade Level	OMR	9.724%	8,176	High

Form Name	Field Name	Description	Mode	Nonblank Error %	Total Nonblank Records	Outlier
d2u	p2asia_1	Person 2: Other Asian	OCR	19.083%	545	Very High
	p2trib_1	Person 2: Am Indian, Alaska Native Tribe	OCR	17.576%	990	Very High
	p2race_1	Person 2: Other Race	OCR	16.018%	899	High
	p3trib_1	Person 3: Am Indian, Alaska Native Tribe	OCR	15.949%	627	High
	p2hisp_1	Person 2: Other Hispanic Origin	OCR	15.326%	783	High
	p3hisp_1	Person 3: Other Hispanic Origin	OCR	14.865%	518	High
	p1trib_1	Person 1: Am Indian, Alaska Native Tribe	OCR	14.690%	953	High
	p1hisp_1	Person 1: Other Hispanic Origin	OCR	14.491%	904	High
	p1race_1	Person 1: Other Race	OCR	12.879%	924	High
	p3race_1	Person 3: Other Race	OCR	12.868%	544	High
	p1welfr	Person 1: Welfare Amount	KFI	10.794%	982	High
	p2yrmvus	Person 2: Migration Year	KFI	10.720%	681	High
	p1_other	Person 1: Other Income Amount	KFI	10.327%	2,537	High
	p1ssi	Person 1: SSI Amount	KFI	10.189%	2,061	High
	p1stx16a	Street Name	KFI	10.123%	29,874	High
	p2_other	Person 2: Other Income Amount	KFI	9.758%	1,117	High
	p2welfr	Person 2: Welfare Amount	KFI	9.552%	513	High
d2ur	p1oauto	Household: Number of Automobiles	OMR	72.310%	1,589	Very High
	p1obdrm	Household: Number of Bedrooms	OMR	71.420%	1,578	Very High
	p2lang	Person 2: Language	OCR	68.484%	587	Very High
	p1lang	Person 1: Language	OCR	67.950%	805	Very High
	p1stx16a	Street Name	KFI	19.272%	1,126	Very High
	p1addr_1	Person 1: Work Address	KFI	18.474%	498	Very High
	p1hsn10a	House Number	KFI	12.796%	719	High
	p2last	Person 2: Last Name	KFI	11.950%	636	High
	p1last	Person 1: Last Name	KFI	9.873%	709	High

Table 14. Field Category Nonblank Error Rates by Mode of Data Capture

Mode of Data Capture	Field Category	Nonblank Error %	Outlier
KFI	POP--Income	7.051%	
	POP--Occupation	6.141%	
	POP--Name	5.842%	
	POP--Ethnic	5.116%	
	Housing Profile	4.841%	
	POP--Race	4.687%	
	POP--Demographic	4.474%	
	Special Housing	2.606%	
	Form Management	1.723%	

Mode of Data Capture	Field Category	Nonblank Error %	Outlier
OCR	POP--Race	7.214%	
	Form Management	5.817%	
	POP--Name	2.212%	
	POP--Ethnic	2.182%	
	Special Housing	1.633%	
	POP--Income	1.167%	
	POP--Occupation	0.786%	
	Housing Profile	0.776%	
	POP--Demographic	0.571%	
OMR	POP--Military	1.857%	
	POP--Occupation	1.729%	
	POP--Education	1.614%	
	Housing Profile	1.150%	
	POP--Income	0.909%	
	POP--Disability	0.759%	
	POP--Demographic	0.739%	
	Form Management	0.672%	
	Coverage	0.452%	
	POP--Race	0.353%	
	POP--Ethnic	0.306%	

From Table 14, we see none of the field category error rates are outliers. Understanding of outliers has to take place at the level of individual fields. This information is found in Table 13. We see different issues highlighted for different forms. For the d1s, the Spanish mailout/mailback short form, name related fields is a dominant issue. For the d2, the English mailout/mailback long form, and the d2u, the English update/leave long form, the write-in fields for other race or ethnicity appear many times on the outlier list. The d2e, the English enumerator long form, shows several outliers for occupation related fields.

4.6 Analysis of Hard and Soft Match Error Rates By Data Capture Center

4.6.1 Contents of This Section

In this section, we use a new grouping of the data called data capture center to analyze the hard match and soft match error rates. In the previous section, we were concerned about how the nonblank error rate behaved depending on

- form (whether we are dealing with a d1, d2, etc.),
- field category (whether we are dealing with name fields, race fields, etc.),
- field (whether we are dealing with name data for person 1, person 2, etc), and
- data capture mode (OCR, OMR, or KFI).

The data capture center are the four locations in Census 2000 at which forms were received, scanned, and converted into useable electronic files. We refer to the data capture centers by their cities of location: Baltimore, Jeffersonville, Phoenix, and Pomona.

Our basic question in this section is this: does the nonblank error rate vary in a significant way depending on what form, field category, type of field, and data capture center we are talking about? To answer this question, we construct an analysis of variance (ANOVA) where the nonblank error rate is the response variable and the independent variables are form, field category, field, and data capture center.

In this section, we also distinguish between person and nonperson fields as discussed in section 4.4.1. For definitions of common or special terms in this section, see the glossary in Appendix M.

After the ANOVA, we show Tables 19 and 20. The data for the tables are the same as for the ANOVA. After going through the different combinations of forms, fields, and data capture centers, we have a raw data set consisting of 9,883 hard and soft match error rates for the ANOVA and the tables. In Table 19, we show nonblank error rates that are outliers for specific fields on specific forms. We aim for a sufficiently fine level of detail that makes it easy to identify the largest improvement opportunities.

Table 20 complements Table 19. We aim for a higher level of detail that supports a meaningful overall view of the data. We show the nonblank error rates for each field category. We show a separate field category result for each of the four data capture centers. Any outliers in Table 20 identify field categories that stand out in terms of a high error rate.

The method for testing statistical significance follows sections 4.4.3 and 4.4.4 The details concerning the calculation of errors follows section 4.2.2. The rules concerning the determination of outliers is as described in section 4.3.

4.6.2 Factors and Models for Testing Statistical Significance

Our factors for testing statistical significance are data capture center (identified by the abbreviation DCC), form, field, field category, and person number. We regard these factors as fixed. For more details about the significance testing, see Appendix J.

We analyze nonperson fields for statistical significance separately from person fields. For nonperson fields, our model is

- field nested within field category,
- field category nested within form, and
- DCC crossed with field.

For person fields, our model is

- person number nested within field,
- field nested within field category,
- field category nested within form, and
- DCC crossed with field.

We present four analyses:

- nonperson fields excluding all outliers
- nonperson fields including all outliers
- person fields excluding all outliers
- person fields including all outliers.

4.6.3 Significance Testing for Nonperson Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 15a. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers,

Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	241	35518.60153	147.38009	30.39	**<0.0001**
Error	213	1033.03152	4.84991		
Corrected Total	454	36551.63306			

Table 15b. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	298.2262713	27.1114792	5.59	**<0.0001**
Field Category	11	148.7294909	13.5208628	2.79	**0.0021**
Field	NA	NA			
DCC	3	2.0949027	0.6983009	0.14	0.9334
Field*DCC	156	224.9933534	1.4422651	0.30	1.0000

Table 16a. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	276	101307.0898	367.0547	47.66	**<0.0001**
Error	266	2048.5499	7.7013		
Corrected Total	542	103355.6397			

Table 16b. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	1322.895597	120.263236	15.62	**<0.0001**
Field Category	12	683.682893	56.973574	7.40	**<0.0001**
Field	NA	NA			
DCC	3	3.670158	1.223386	0.16	0.9239
Field*DCC	187	297.584533	1.591361	0.21	1.0000

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables agree as to which individual factors are significant. For nonperson fields, therefore, the largest

significant factor is form. There is a significant secondary contribution from field category. The structure of the data set did not allow SAS to test field for significance.

4.6.4 Significance Testing for Person Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 17a. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2727	118461.9974	43.4404	15.91	**<0.0001**
Error	5198	14194.7383	2.7308		
Corrected Total	7925	132656.7357			

Table 17b. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	249.913247	24.991325	9.15	**<0.0001**
Field Category	48	2289.274122	47.693211	17.46	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			
DCC	3	12.657393	4.219131	1.55	0.2007
Field*DCC	1965	1845.212756	0.939040	0.34	1.0000

Table 18a. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3033	166775.1743	54.9869	13.24	**<0.0001**
Error	6306	26193.0635	4.1537		
Corrected Total	9339	192968.2378			

Table 18b. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	416.473326	41.647333	10.03	**<0.0001**
Field Category	50	3091.937365	61.838747	14.89	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			
DCC	3	40.155894	13.385298	3.22	**0.0217**
Field*DCC	2225	3147.278035	1.414507	0.34	1.0000

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables do not agree as to which individual factors are significant. Since outliers are known to distort results, it is preferable to conclude based on excluding outliers. For person fields, therefore, the largest significant factor is field category. There is a significant secondary contribution from form. The structure of the data set did not allow SAS to test field and person number for significance.

4.6.5 Outlier Data for This Section

We have reached another point in our analysis where the volume of data becomes an issue in table construction. As mentioned in section 4.6.1, when we calculate the nonblank error rate for all the combinations of variables relevant to this analysis, we have 9,883 rates by the time we are done. Some of these rates–almost 1,500--are high or very high outliers according to the procedure discussed in section 4.3. How do we communicate what these outliers have to say without forcing the reader to wade through a 1,500 line table?

We think a fair compromise is to restrict the table to the outliers that are based on a reasonably large number of records. It is hard to conclude much when the data behind an outlier consist of two, three, or some other small number of records. After experimenting with different possibilities, we believe 500 records is a reasonable minimum to require. This results in Table 19. It consists of 234 outliers. It provides insight into the highest two percent of the nonblank error rates. We believe this emphasizes problem fields that occur often enough to be a priority for investigation and improvement. In Tables 19 and 20, the data capture centers are abbreviated as follows:

- BAL means Baltimore,
- JEF means Jeffersonville,
- PHX means Phoenix, and
- POM means Pomona.

Table 19. Field Nonblank Error Rates that are High and Very High Outliers and Based on at Least 500 Blank and Nonblank Data Records

Form Name	Field Name	Description	Data Capture Center	Nonblank Error %	Total Nonblank Records	Outlier
d1	p3_relo	Person 3: Other Relative	PHX	8.370%	920	High
d1e	p4ocancl	Person 4: Cancel	POM	16.110%	509	Very High
	rilast	Respondent's Last Name	PHX	11.709%	47,058	High
	rilast	Respondent's Last Name	JEF	11.039%	13,262	High
	rilast	Respondent's Last Name	BAL	10.987%	30,772	High
	rilast	Respondent's Last Name	POM	10.866%	40,869	High
	rifirst	Respondent's First Name	PHX	8.597%	47,412	High
	rifirst	Respondent's First Name	JEF	8.521%	13,414	High
d1s	p5mi	Person 5: Middle Initial	PHX	10.847%	590	High
	p4mi	Person 4: Middle Initial	PHX	10.262%	916	High
	p1hisp19	Person 1: Other Hispanic Origin	PHX	10.000%	1,120	High
	p2hisp19	Person 2: Other Hispanic Origin	PHX	10.000%	1,000	High
	p3mi	Person 3: Middle Initial	PHX	9.783%	1,196	High
	p2mi	Person 2: Middle Initial	PHX	9.261%	1,393	High
	p1mi	Person 1: Middle Initial	PHX	9.215%	1,541	High
d1u	p1apt16a	Apartment Number	POM	9.988%	851	High
	p1stx16a	Street Name	JEF	9.001%	911	High
	p1apt16a	Apartment Number	BAL	8.923%	650	High
	p1apt16a	Apartment Number	PHX	8.068%	1,475	High
d2	p2trib_1	Person 2: Am Indian, Alaska Native Tribe	POM	31.002%	1,258	Very High
	p1asia_1	Person 1: Other Asian	BAL	30.856%	619	Very High

p2trib_1	Person 2: Am Indian, Alaska Native Tribe	PHX	29.868%	606 Very High
p4trib_1	Person 4: Am Indian, Alaska Native Tribe	POM	29.577%	568 Very High
p2trib_1	Person 2: Am Indian, Alaska Native Tribe	BAL	28.319%	678 Very High
p3trib_1	Person 3: Am Indian, Alaska Native Tribe	POM	28.074%	862 Very High
p2asia_1	Person 2: Other Asian	POM	27.453%	958 Very High
p2asia_1	Person 2: Other Asian	BAL	27.076%	602 Very High
p4hisp_1	Person 4: Other Hispanic Origin	POM	26.817%	619 Very High
p1trib_1	Person 1: Am Indian, Alaska Native Tribe	BAL	26.480%	642 Very High
p1trib_1	Person 1: Am Indian, Alaska Native Tribe	POM	26.117%	1,164 Very High
p1asia_1	Person 1: Other Asian	PHX	25.519%	482 Very High
p1asia_1	Person 1: Other Asian	POM	25.457%	876 Very High
p3hisp_1	Person 3: Other Hispanic Origin	POM	24.080%	951 Very High
p3hisp_1	Person 3: Other Hispanic Origin	PHX	23.384%	727 Very High
p3asia_1	Person 3: Other Asian	POM	23.370%	629 Very High
p4hisp_1	Person 4: Other Hispanic Origin	BAL	22.330%	515 Very High
p2hisp_1	Person 2: Other Hispanic Origin	POM	22.230%	1,408 Very High
p1hisp_1	Person 1: Other Hispanic Origin	POM	21.786%	1,680 Very High
p2race_1	Person 2: Other Race	BAL	21.682%	1,070 Very High
p3hisp_1	Person 3: Other Hispanic Origin	BAL	21.305%	751 Very High
p1hisp_1	Person 1: Other Hispanic Origin	PHX	21.237%	1,229 Very High
p5race_1	Person 5: Other Race	POM	21.053%	608 Very High
p2hisp_1	Person 2: Other Hispanic Origin	PHX	21.013%	1,066 Very High
p2race_1	Person 2: Other Race	POM	20.998%	2,024 Very High
p3race_1	Person 3: Other Race	POM	20.899%	1,402 Very High
p4race_1	Person 4: Other Race	POM	20.659%	1,002 Very High
p3race_1	Person 3: Other Race	PHX	20.408%	637 Very High
p3race_1	Person 3: Other Race	BAL	20.061%	658 Very High
p1hisp_1	Person 1: Other Hispanic Origin	BAL	19.637%	1,212 Very High
p2hisp_1	Person 2: Other Hispanic Origin	BAL	18.416%	1,086 Very High
p2race_1	Person 2: Other Race	PHX	18.162%	925 Very High
p1race_1	Person 1: Other Race	BAL	17.563%	1,264 Very High
p1race_1	Person 1: Other Race	PHX	17.238%	1,050 Very High
p2_other	Person 2: Other Income Amount	PHX	17.211%	1,255 Very High
p3addr_1	Person 3: Work Address	POM	16.776%	3,815 Very High
p1ssi	Person 1: SSI Amount	PHX	16.676%	1,799 Very High
p4addr_1	Person 4: Work Address	POM	16.413%	1,249 Very High
p1addr_1	Person 1: Work Address	POM	16.031%	28,040 Very High
p1race_1	Person 1: Other Race	POM	15.904%	2,207 Very High
p2addr_1	Person 2: Work Address	POM	14.961%	17,111 Very High
p1welfr	Person 1: Welfare Amount	PHX	14.917%	905 Very High
p2ssi	Person 2: SSI Amount	PHX	14.330%	963 Very High
p2oresp	Person 2: How Long	POM	14.206%	535 Very High
p5addr_1	Person 5: Work Address	BAL	13.992%	486 Very High

Form Name	Field Name	Description	Data Capture Center	Nonblank Error %	Total Nonblank Records	Outlier
d2	p1_other	Person 1: Other Income Amount	PHX	13.549%	3,218	High
	p3addr_1	Person 3: Work Address	BAL	13.525%	4,510	High
	p3_relo	Person 3: Other Relative	PHX	13.106%	557	High
	p1addr_1	Person 1: Work Address	BAL	13.020%	29,515	High
	p4addr_1	Person 4: Work Address	PHX	12.990%	816	High
	p3_relo	Person 3: Other Relative	BAL	12.836%	670	High
	p3addr_1	Person 3: Work Address	PHX	12.788%	2,776	High
	p1oresp	Person 1: How Long	POM	12.768%	838	High
	p1addr_1	Person 1: Work Address	PHX	12.580%	21,685	High
	p1oresp	Person 1: How Long	BAL	12.470%	826	High
	p2addr_1	Person 2: Work Address	BAL	11.690%	18,478	High
	p2addr_1	Person 2: Work Address	PHX	11.628%	13,081	High
	p6omilit	Person 6: Active Duty	BAL	11.554%	502	High
	p4addr_1	Person 4: Work Address	BAL	11.536%	1,465	High
	p1addr_1	Person 1: Work Address	JEF	11.483%	12,070	High
	p3_relo	Person 3: Other Relative	POM	11.111%	666	High
	p4addr_1	Person 4: Work Address	JEF	11.052%	561	High
	p2addr_1	Person 2: Work Address	JEF	10.631%	7,798	High
	p2oresp	Person 2: How Long	PHX	10.546%	531	High
	p3addr_1	Person 3: Work Address	JEF	10.410%	1,806	High
	p1oresp	Person 1: How Long	PHX	10.200%	902	High
	p1oslfls	Person 1: Self- Person 1:employment Loss	BAL	10.116%	692	High
	p2_relo	Person 2: Other Relative	POM	9.651%	1,492	High
	p4empl_1	Person 4: Employer	BAL	9.629%	1,942	High
	p1oslfls	Person 1: Self- Person 1:employment Loss	PHX	9.593%	615	High
	p5otype	Person 5: Business Type	BAL	9.582%	574	High
	p5empl_1	Person 5: Employer	BAL	9.412%	595	High
	p1osecpy	Household: No Payment	JEF	9.372%	1,227	High
	p4empl_1	Person 4: Employer	POM	9.158%	1,758	High
	p1oslfls	Person 1: Self- Person 1:employment Loss	POM	8.948%	827	High
	p2oserve	Person 2: When on Active Duty	PHX	8.830%	1,461	High
	p4empl_1	Person 4: Employer	PHX	8.821%	1,145	High
	p1ograde	Person 1: Grade Level	JEF	8.754%	3,073	High
	p1selfe	Person 1: Self Employment Income Amount	PHX	8.736%	3,514	High
	p2selfe	Person 2: Self Employment Income Amount	PHX	8.715%	1,595	High
	p2oserve	Person 2: When on Active Duty	POM	8.684%	1,520	High
	p1oelec	Household: Electricity	JEF	8.472%	779	High
	p2_relo	Person 2: Other Relative	BAL	8.461%	1,501	High
	p2welfr	Person 2: Welfare Amount	BAL	8.258%	666	High
	p1ssi	Person 1: SSI Amount	BAL	8.180%	2,604	High
	p6omilit	Person 6: Active Duty	POM	8.130%	615	High
	p1ograde	Person 1: Grade Level	PHX	8.117%	7,798	High
	p4otrans	Person 4: Work Vehicle	JEF	8.112%	678	High
	p3yrmvus	Person 3: Migration Year	PHX	8.093%	1,631	High

Form Name	Field Name	Description	Data Capture Center	Nonblank Error %	Total Nonblank Records	Outlier
d2e	p4oserve	Person 4: When on Active Duty	POM	87.444%	669	Very High
	p4oserve	Person 4: When on Active Duty	PHX	82.765%	528	Very High
	p3oneeds	Person 3: Responsible for Needs	JEF	66.960%	569	Very High
	p2oresp	Person 2: How Long	POM	53.826%	745	Very High
	p5oetype	Person 5: Class of Worker	BAL	45.568%	722	Very High
	p5oetype	Person 5: Class of Worker	PHX	45.398%	804	Very High
	p4otrans	Person 4: Work Vehicle	JEF	44.863%	584	Very High
	p4otrans	Person 4: Work Vehicle	PHX	40.776%	1,469	Very High
	p5oetype	Person 5: Class of Worker	POM	39.670%	1,031	Very High
	p4otrans	Person 4: Work Vehicle	POM	37.534%	1,833	Very High
	p2oresp	Person 2: How Long	PHX	37.234%	564	Very High
	p3oneeds	Person 3: Responsible for Needs	BAL	36.402%	945	Very High
	p4otrans	Person 4: Work Vehicle	BAL	33.776%	1,356	Very High
	p5otype	Person 5: Business Type	BAL	33.739%	575	Very High
	p3oborn	Person 3: Under 17	JEF	24.525%	579	Very High
	p5olook	Person 5: Looking for Work	POM	20.592%	845	Very High
	p3oneeds	Person 3: Responsible for Needs	POM	20.476%	757	Very High
	rilast	Respondent's Last Name	JEF	20.396%	18,759	Very High
	p4oride	Person 4: Carpool	POM	20.362%	1,105	Very High
	rilast	Respondent's Last Name	POM	19.178%	51,930	Very High
	p4oride	Person 4: Carpool	PHX	18.374%	898	Very High
	p3ostart	Person 3: Could Start Last Week	POM	17.941%	1,059	Very High
	p5owork	Person 5: Work Last Year	PHX	17.235%	586	Very High
	p1stx16a	Street Name	POM	16.985%	10,680	Very High
	rilast	Respondent's Last Name	PHX	16.644%	53,312	Very High
	p5otype	Person 5: Business Type	POM	16.374%	684	Very High
	p1oserve	Person 1: When on Active Duty	JEF	16.203%	1,401	Very High
	p3oneeds	Person 3: Responsible for Needs	PHX	16.048%	754	Very High
	rilast	Respondent's Last Name	BAL	15.455%	42,556	Very High
	p5otype	Person 5: Business Type	PHX	14.881%	504	Very High
	p5owork	Person 5: Work Last Year	POM	14.774%	731	Very High
	p3ostart	Person 3: Could Start Last Week	PHX	14.472%	919	Very High
	p5olstwk	Person 5: Last Worked	PHX	14.016%	635	Very High
	p5olvcty	Person 5: Live Inside City Limits	POM	13.961%	1,540	Very High
	p3oborn	Person 3: Under 17	BAL	13.932%	1,414	Very High
	rifirst	Respondent's First Name	JEF	13.573%	18,950	High
	p1oserve	Person 1: When on Active Duty	POM	13.347%	4,765	High
	p5olook	Person 5: Looking for Work	BAL	13.297%	549	High
	p3owork	Person 3: Work Last Year	JEF	13.084%	1,284	High
	rifirst	Respondent's First Name	POM	12.947%	52,576	High
	p3oetype	Person 3: Class of Worker	JEF	12.901%	1,248	High
	p1ograde	Person 1: Grade Level	JEF	12.516%	775	High
	rifirst	Respondent's First Name	PHX	12.458%	53,774	High
	p5olvcty	Person 5: Live Inside City Limits	BAL	12.247%	841	High
	p4ospkwl	Person 4: Speak English Well	JEF	11.975%	643	High
	p5olstwk	Person 5: Last Worked	POM	11.958%	761	High
	p4owages	Person 4: Wages	JEF	11.532%	581	High
	p3ograND	Person 3: Grandchildren	JEF	11.340%	1,896	High

Form Name	Field Name	Description	Data Capture Center	Nonblank Error %	Total Nonblank Records	Outlier
d2e	p2ostart	Person 2: Could Start Last Week	JEF	11.073%	578	High
	p3ostart	Person 3: Could Start Last Week	BAL	11.056%	805	High
	p2ostart	Person 2: Could Start Last Week	POM	10.549%	2,057	High
	p1oserve	Person 1: When on Active Duty	PHX	10.463%	4,100	High
	rifirst	Respondent's First Name	BAL	10.449%	43,152	High
	p4oam_pm	Person 4: Time to Work am/pm	PHX	10.294%	816	High
	p2oserve	Person 2: When on Active Duty	PHX	10.235%	938	High
	p3addr_1	Person 3: Work Address	POM	10.163%	2,027	High
	p5olook	Person 5: Looking for Work	PHX	10.02%	589	High
	p5olvcty	Person 5: Live Inside City Limits	PHX	9.804%	1,071	High
	p5ojob	Person 5: Difficulty Working	JEF	9.774%	532	High
	p3orecal	Person 3: Will Be Recalled	POM	9.587%	678	High
	p4ototal	Person 4: Total Income None	PHX	9.478%	823	High
	p5olstwk	Person 5: Last Worked	BAL	9.416%	531	High
	p4omilit	Person 4: Active Duty	POM	9.320%	2,736	High
	p3oetype	Person 3: Class of Worker	PHX	9.247%	3,201	High
	p2oserve	Person 2: When on Active Duty	POM	9.211%	912	High
	p2oneeds	Person 2: Responsible for Needs	JEF	9.172%	785	High
	p4omilit	Person 4: Active Duty	JEF	8.938%	772	High
	p1osecpy	Household: No Payment	POM	8.929%	672	High
	p3oetype	Person 3: Class of Worker	BAL	8.866%	3,316	High
	p2oneeds	Person 2: Responsible for Needs	POM	8.785%	2,470	High
	p4addr_1	Person 4: Work Address	POM	8.675%	830	High
	p5otrans	Person 5: Work Vehicle	POM	8.671%	519	High
	p1ograde	Person 1: Grade Level	PHX	8.633%	2,502	High
	p2oneeds	Person 2: Responsible for Needs	PHX	8.562%	2,628	High
	p2oserve	Person 2: When on Active Duty	BAL	8.545%	749	High
	p3oetype	Person 3: Class of Worker	POM	8.491%	4,016	High
	p4addr_1	Person 4: Work Address	BAL	8.464%	638	High
	p1oserve	Person 1: When on Active Duty	BAL	8.442%	3,388	High
	p1omort	Household: No Payment	PHX	8.392%	572	High
	p3orecal	Person 3: Will Be Recalled	PHX	8.199%	683	High
	p4ospkwl	Person 4: Speak English Well	PHX	8.196%	2,184	High
	p4oproft	Person 4: Work Last Week	JEF	8.149%	724	High
	p2oborn	Person 2: Under 16	JEF	8.130%	861	High
	p5ojob	Person 5: Difficulty Working	POM	8.086%	2,090	High

Form Name	Field Name	Description	Data Capture Center	Nonblank Error %	Total Nonblank Records	Outlier
d2u	p2trib_1	Person 2: Am Indian, Alaska Native Tribe	POM	16.192%	562	Very High
	p1trib_1	Person 1: Am Indian, Alaska Native Tribe	POM	14.748%	556	Very High
	p1stx16a	Street Name	JEF	12.785%	2,190	High
	p2_other	Person 2: Other Income Amount	PHX	11.252%	631	High
	p1_other	Person 1: Other Income Amount	PHX	10.619%	1,535	High
	p2ograde	Person 2: Grade Level	JEF	10.363%	579	High
	p1ssi	Person 1: SSI Amount	PHX	10.261%	1,150	High
	p1stx16a	Street Name	POM	10.224%	10,524	High
	p1addr_1	Person 1: Work Address	BAL	10.202%	7,636	High
	p1addr_1	Person 1: Work Address	JEF	10.122%	2,460	High
	p2ssi	Person 2: SSI Amount	PHX	9.980%	511	High
	p1stx16a	Street Name	BAL	9.875%	6,694	High
	p1oagric	Household: Agricultural Products	POM	9.821%	10,987	High
	p2addr_1	Person 2: Work Address	JEF	9.707%	1,772	High
	p1stx16a	Street Name	PHX	9.622%	10,466	High
	p1ograde	Person 1: Grade Level	JEF	9.408%	574	High
	p1apt16a	Apartment Number	POM	9.227%	1,398	High
	p3addr_1	Person 3: Work Address	BAL	9.195%	1,131	High
	p1addr_1	Person 1: Work Address	POM	9.190%	11,795	High
	p2addr_1	Person 2: Work Address	BAL	9.008%	5,273	High
	p1oelec	Household: Electricity	POM	8.886%	664	High
	p1addr_1	Person 1: Work Address	PHX	8.413%	9,259	High
	p1apt16a	Apartment Number	BAL	8.410%	1,082	High
	p1ograde	Person 1: Grade Level	POM	8.407%	3,069	High
	p2addr_1	Person 2: Work Address	POM	8.313%	8,228	High
	p2addr_1	Person 2: Work Address	PHX	8.191%	6,202	High
	p3addr_1	Person 3: Work Address	PHX	8.138%	1,278	High
d2ur	p1oauto	Household: Number of Automobiles	POM	72.292%	1,588	Very High
	p1obdrm	Household: Number of Bedrooms	POM	71.401%	1,577	Very High
	p1lang	Person 1: Language	POM	48.204%	1,197	Very High
	p3lang	Person 3: Language	POM	45.920%	625	Very High
	p2lang	Person 2: Language	POM	45.455%	957	Very High
	p1stx16a	Street Name	POM	19.200%	1,125	Very High
	p1addr_1	Person 1: Work Address	POM	18.310%	497	Very High
	p1hsn10a	House Number	POM	12.813%	718	High
	p2last	Person 2: Last Name	POM	9.117%	1,382	High
	p3last	Person 3: Last Name	POM	9.010%	899	High
	p4ohisp	Person 4: Hispanic Origin	POM	9.007%	544	High
	p2ohisp	Person 2: Hispanic Origin	POM	8.683%	1,359	High
	p1actv_1	Person 1: Industry	POM	8.392%	715	High
	p3ohisp	Person 3: Hispanic Origin	POM	8.250%	897	High

Table 20. Field Category Nonblank Error Rates by Data Capture Center

Data Capture Center	Field Category	Nonblank Error %	Outlier
BAL	Form Management	3.128%	
	POP--Name	2.987%	
	Special Housing	2.340%	
	POP--Occupation	2.281%	
	POP--Military	1.503%	
	POP--Education	1.440%	
	POP--Income	1.329%	
	POP--Ethnic	1.305%	
	Housing Profile	1.165%	
	POP--Demographic	0.922%	
	POP--Race	0.825%	
	POP--Disability	0.703%	
	Coverage	0.440%	
JEF	Form Management	3.662%	High
	POP--Name	3.491%	High
	POP--Occupation	2.455%	
	POP--Military	2.348%	
	Special Housing	2.130%	
	POP--Education	1.949%	
	POP--Income	1.612%	
	Housing Profile	1.484%	
	POP--Ethnic	1.436%	
	POP--Demographic	1.106%	
	POP--Disability	1.086%	
	POP--Race	0.942%	
	Coverage	0.578%	
PHX	Form Management	3.421%	High
	POP--Name	3.237%	High
	POP--Occupation	2.196%	
	Special Housing	2.121%	
	POP--Military	1.905%	
	POP--Income	1.560%	
	POP--Education	1.551%	
	Housing Profile	1.289%	
	POP--Ethnic	1.128%	
	POP--Demographic	1.000%	
	POP--Race	0.827%	
	POP--Disability	0.724%	
	Coverage	0.391%	

Data Capture Center	Field Category	Nonblank Error %	Outlier
POM	Form Management	3.361%	High
	POP--Name	3.178%	
	POP--Occupation	2.396%	
	POP--Military	1.981%	
	Special Housing	1.962%	
	POP--Education	1.719%	
	Housing Profile	1.443%	
	POP--Ethnic	1.426%	
	POP--Income	1.364%	
	POP--Race	1.249%	
	POP--Demographic	1.047%	
	POP--Disability	0.734%	
	Coverage	0.485%	

From Table 20, we see that although they are not outliers in all four centers, the categories Form Management and POP–Name have the highest nonblank error rates in all four. Form Management covers the person added and person canceled fields on the enumerator forms. It is encouraging to note that only one of the 52 outlier rates in Table 19 for Form Management was for adding or canceling persons. While the entries in Table 19 should be gleaned to identify opportunities for improvement, the higher level view of Table 20 suggests an interesting follow up question. What specifically is there about the nature of the Form Management and POP–Name categories that leads them to occupy the top two positions in all four data capture centers?

4.7 Analysis of Hard and Soft Match Error Rates By Census 2000 Regional Census Center

4.7.1 Contents of This Section and a Special Issue Affecting the Analysis

In this section, we use a new grouping of the data called Census 2000 regional census centers to analyze the hard match and soft match error rates. In the previous section, we were concerned about how the nonblank error rate behaved depending on

- form (whether we are dealing with a d1, d2, etc.),
- field category (whether we are dealing with name fields, race fields, etc.),
- field (whether we are dealing with name data for person 1, person 2, etc), and
- data capture center (Baltimore, Jeffersonville Phoenix, or Pomona).

In Census 2000, the twelve regional census centers across the United States were the next layer of management below Suitland, MD, headquarters. The twelve regional census centers were numbered from 21 to 32.

Our basic question in this section is this: does the nonblank error rate vary in a significant way depending on what form, field category, type of field, and Census 2000 regional census center we are talking about? To answer this question, we construct an analysis of variance (ANOVA) where the nonblank error rate is the response variable and the independent variables are form, field category, field, and Census 2000 regional census center.

As explained in section 3.3, the analysis in this final draft of this evaluation includes 666,711 records that were left out of the analysis in the initial draft. By including these records, the analysis of this section is affected in a way not pertinent to the other sections. We originally excluded the records because we were unable to match them to the twelve regional census center files.

Although we could not match them, we concluded with the help of our contractor that they could be treated as if they did match. In calculating the hard and soft match error rates by regional census center, the analysis for the final draft produces 27,254 combinations of field, form, and regional census center. This is 9,071 more than the 18,183 combinations produced by the analysis in the initial draft.

There are many combinations of field, form, and Census 2000 regional census center where all the records have a hard or soft match error, leading to an error rate of 100 percent for that combination. This can happen especially when the total number of cases for a combination is small.

There are enough combinations where the error rate is 100 percent that when the 666,711 unmatched records are included, 100 percent is the boundary of the third quartile when the error rates are sorted in ascending order. Since outliers are a function of the interquartile range, and the interquartile range depends on the value for the boundary of the third quartile, none of the error rates in the set of 27,254 can be classified as an outlier.

The interquartile range is nearly 100 percent. Outliers occur at a distance from the median at least equal to 1.5 times the interquartile range, or nearly 150 percent in this case. When the raw data are in the form of percents as it is here, outliers are impossible under these conditions.

We face two choices: include all 27,254 error rates in the analysis or exclude the 9,071 rates that lead to the condition of no outliers. We do not believe it is prudent to put forth an analysis in which the structure of the data rules out the possibility of outliers. A case could be made that the 27,254 error rates should be regarded not as one universe but at least two.

In this section, we choose the second option. The analysis is restricted to the 18,183 combinations of field, form, and Census 2000 regional census center used in the initial draft of this evaluation. Some of these exist within the 666,711 unmatched records. We include these cases in the analysis so the results will not duplicate the initial draft of this evaluation.

In the interest of a full comparison, we add an extra appendix to the final draft. In Appendix K, we include all 27,254 error rates in testing factors for statistical significance. We conclude the appendix by noting any similarities or differences to the findings of this section. Where the findings conflict, we believe the results of this section should be preferred.

After the ANOVA, we show Tables 25 and 26. The data for the tables are the same as for the ANOVA. In this section, we also distinguish between person and nonperson fields as discussed in section 4.4.1.

In Table 25, we show nonblank error rates that are outliers for specific fields on specific forms. We aim for a sufficiently fine level of detail that makes it easy to identify the largest improvement opportunities. Table 26 complements Table 25. We aim for a higher level of detail that supports a meaningful overall view of the data. We show the nonblank error rates for each field category. We show a separate field category result for each of the twelve Census 2000 regional census centers. Any outliers in Table 26 identify field categories that stand out in terms of a high error rate.

Additional tables appear in Appendix L. They show the nonblank error rates by each field category within Census 2000 regional census center but broken out further between respondent-returned and enumerator-returned forms. The method for testing statistical significance follows section 4.4.3 and 4.4.4. The details concerning the calculation of errors follows section 4.2.2. The rules concerning the determination of outliers is as described in section 4.3. For definitions of common or special terms in this section, see the glossary in Appendix M.

4.7.2 Factors and Models for Testing Statistical Significance

Our factors for testing statistical significance are Census 2000 regional census center (abbreviated as RCC), form, field, field category, and person number. We regard these factors as fixed. For more details about the significance testing, see Appendix J.

We analyze nonperson fields for statistical significance separately from person fields. For nonperson fields, our model is

- field nested within field category,
- field category nested within form, and
- regional census center crossed with field.

For person fields, our model is

- person number nested within field,
- field nested within field category,
- field category nested within form, and
- regional census center.

We present four analyses:

- nonperson fields excluding all outliers
- nonperson fields including all outliers
- person fields excluding all outliers
- person fields including all outliers.

4.7.3 Significance Testing for Nonperson Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 21a. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	620	32885.15615	53.04057	28.67	**<0.0001**
Error	520	962.00422	1.85001		
Corrected Total	1140	33847.16037			

Table 21b. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	199.6940704	18.1540064	9.81	**<0.0001**
Field Category	10	40.4267420	4.0426742	2.19	**0.0175**
Field	NA	NA			
RCC	11	64.9103424	5.9009402	3.19	**0.0003**
Field*RCC	526	542.3153681	1.0310178	0.56	1.0000

Table 22a. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	713	97825.39284	137.20251	41.51	**<0.0001**
Error	650	2148.35164	3.30516		
Corrected Total	1363	99973.74447			

Table 22b. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	1015.756488	92.341499	27.94	**<0.0001**
Field Category	12	621.284623	51.773719	15.66	**<0.0001**
Field	NA	NA			
RCC	11	56.871296	5.170118	1.56	0.1049
Field*RCC	616	731.420683	1.187371	0.36	1.0000

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables almost

agree as to which individual factors are significant. Form and field category are significant regardless of including outliers. When outliers are excluded, regional census center is significant. When outliers are included, regional census center is just below the threshold of significance. For nonperson fields, the largest significant factor is form. There is a significant secondary contribution of field category. The structure of the data set did not allow SAS to test field for significance.

4.7.4 Significance Testing for Person Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 23a. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	769	85846.2147	111.6336	106.14	**<0.0001**
Error	13586	14289.4062	1.0518		
Corrected Total	14355	100135.6209			

Table 23b. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	177.716261	17.771626	16.90	**<0.0001**
Field Category	48	1813.919223	37.789984	35.93	**<0.0001**
Field	NA	NA.			
Person Number	NA	NA			
RCC	11	739.626950	67.238814	63.93	**<0.0001**

Table 24a. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	816	122095.6298	149.6270	92.93	**<0.0001**
Error	16002	25764.1040	1.6101		
Corrected Total	16818	147859.7339			

Table 24b. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	351.972403	35.197240	21.86	**<0.0001**
Field Category	50	2494.339702	49.886794	30.98	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			
RCC	11	791.290444	71.935495	44.68	**<0.0001**

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables agree as to which individual factors are significant. For person fields, the largest significant factor is field category. There is a significant secondary contribution of regional census center. The structure of the data set did not allow SAS to test field and person number for significance. We did not include a test for the interaction of regional census center and field in the person field analysis. Unlike the nonperson analysis, the memory resources available to SAS did not allow enough capacity to test the model with this interaction included.

4.7.5 Outlier Data for This Section

We have reached another point in our analysis where the volume of data becomes an issue in table construction. As mentioned in section 4.7.1, when we calculate the nonblank error rate for all the combinations of variables relevant to this analysis, we have 18,183 rates by the time we are done. Some of these rates–almost 2,700--are high or very high outliers according to the procedure discussed in section 4.3. How do we communicate what these outliers have to say without forcing the reader to wade through a 2,700 line table?

We think a fair compromise is to restrict the table to the outliers that are based on a reasonably large number of records. It is hard to conclude much when the data behind an outlier consist of two, three, or some other small number of records. After experimenting with different possibilities, we believe 1000 records is a reasonable minimum to require. This results in Table 25. It consists of 153 outliers. It provides insight into the highest 0.8 percent of the nonblank error rates. We believe this emphasizes problem fields that occur often enough to be a priority for investigation and improvement.

Unfortunately, the limits of space do not leave enough room in Tables 25 and 26 to write out in words the areas represented by the regional census center numbers 21 to 32. To make Tables 25 and 26 easier to read, we provide here a list to use in combination with them. It indicates the states covered by the twelve regional census centers.

The twelve Census 2000 regional census centers were organized as follows:

- 21 covered Connecticut, Maine, Massachusetts, New Hampshire, upstate New York, Puerto Rico, Rhode Island, and Vermont;
- 22 covered northern New Jersey and metropolitan New York City;
- 23 covered Delaware, the District of Columbia, Maryland, southern New Jersey, and Pennsylvania;
- 24 covered Michigan, Ohio, and West Virginia;
- 25 covered Illinois, Indiana, and Wisconsin;
- 26 covered Arkansas, Iowa, Kansas, Minnesota, Missouri, and Oklahoma;
- 27 covered Alaska, northern California, Idaho, Oregon, and Washington state;
- 28 covered Kentucky, North Carolina, South Carolina, Tennessee and Virginia;
- 29 covered Alabama, Florida, and Georgia;
- 30 covered Louisiana, Mississippi, and Texas;
- 31 covered Arizona, Colorado, Idaho, Montana, Nebraska, Nevada, New Mexico, North Dakota South Dakota, Utah, and Wyoming; and
- 32 covered southern California and Hawaii.

Table 25. Field Nonblank Error Rates that are High and Very High Outliers and Based on at Least 1000 Blank and Nonblank Data Records

Form Name	Field Name	Description	Nonblank Error %	Total Nonblank Records	RCC	Outlier
d1e	rilast	R1 - Respondent's Last Name	15.820%	9,096	22	Very High
	rc_oc6	6 - Outcome	14.439%	1,212	22	Very High
	rilast	R1 - Respondent's Last Name	13.396%	8,779	23	Very High
	rifirst	R1 - Respondent's First Name	11.936%	9,157	22	High
	rilast	R1 - Respondent's Last Name	11.873%	8,852	21	High
	rilast	R1 - Respondent's Last Name	11.691%	14,644	30	High
	rilast	R1 - Respondent's Last Name	11.621%	9,896	32	High
	rilast	R1 - Respondent's Last Name	11.440%	15,997	29	High
	rilast	R1 - Respondent's Last Name	10.969%	8,433	24	High

Form Name	Field Name	Description	Nonblank Error %	Total Nonblank	RCC	Outlier
d1e	rilast	R1 - Respondent's Last Name	10.820%	9,168	26	High
	rilast	R1 - Respondent's Last Name	10.437%	15,455	28	High
	rilast	R1 - Respondent's Last Name	10.107%	10,013	27	High
	rilast	R1 - Respondent's Last Name	9.238%	11,106	31	High
	rifirst	R1 - Respondent's First Name	9.155%	8,957	23	High
	rifirst	R1 - Respondent's First Name	8.786%	14,682	30	High
	rifirst	R1 - Respondent's First Name	8.398%	8,847	21	High
	rifirst	R1 - Respondent's First Name	8.343%	10,104	32	High
	rifirst	R1 - Respondent's First Name	8.290%	16,284	29	High
	rilast	R1 - Respondent's Last Name	8.268%	10,522	25	High
	rifirst	R1 - Respondent's First Name	7.769%	15,472	28	High
	rifirst	R1 - Respondent's First Name	7.741%	9,198	26	High
d2	p1addr_1	22a - Person 1: Work Address	19.744%	6,331	32	Very High
	p2addr_1	22a - Person 2: Work Address	17.482%	3,781	32	Very High
	p1addr_1	22a - Person 1: Work Address	16.275%	7,447	27	Very High
	p3addr_1	22a - Person 3: Work Address	15.996%	1,044	22	Very High
	p1addr_1	22a - Person 1: Work Address	15.588%	6,614	29	Very High
	p2addr_1	22a - Person 2: Work Address	15.542%	4,581	27	Very High
	p1addr_1	22a - Person 1: Work Address	15.141%	7,635	23	Very High
	p3addr_1	22a - Person 3: Work Address	14.892%	1,014	26	Very High
	p1addr_1	22a - Person 1: Work Address	14.232%	6,380	31	Very High
	p1addr_1	22a - Person 1: Work Address	14.107%	8,173	26	Very High
	p3addr_1	22a - Person 3: Work Address	14.105%	1,184	23	Very High
	p1addr_1	22a - Person 1: Work Address	13.847%	8,529	24	Very High
	p2addr_1	22a - Person 2: Work Address	13.796%	3,849	31	Very High
	p2addr_1	22a - Person 2: Work Address	13.656%	4,855	23	Very High
	p1_other	31h - Person 1: Other Income Amount	13.436%	1,042	28	Very High
	p1addr_1	22a - Person 1: Work Address	13.163%	5,994	22	Very High
	p2addr_1	22a - Person 2: Work Address	13.143%	5,090	26	Very High
	p3addr_1	22a - Person 3: Work Address	12.872%	1,243	24	High
	p2addr_1	22a - Person 2: Work Address	12.224%	3,493	22	High
	p2addr_1	22a - Person 2: Work Address	12.192%	5,405	24	High
	p3addr_1	22a - Person 3: Work Address	11.695%	1,009	28	High
	p1addr_1	22a - Person 1: Work Address	11.391%	11,474	25	High
	p1addr_1	22a - Person 1: Work Address	11.378%	7,286	30	High
	p1addr_1	22a - Person 1: Work Address	11.187%	7,929	28	High
	p3addr_1	22a - Person 3: Work Address	10.968%	1,085	21	High
	p2addr_1	22a - Person 2: Work Address	10.649%	4,789	28	High
	p2addr_1	22a - Person 2: Work Address	10.623%	7,418	25	High
	p3addr_1	22a - Person 3: Work Address	10.304%	1,679	25	High
	p2addr_1	22a - Person 2: Work Address	10.179%	4,254	30	High
	p1addr_1	22a - Person 1: Work Address	9.963%	7,518	21	High
	p1osecpy	48b - Household: No Payment	9.413%	1,158	25	High
	p1selfe	31b - Person 1: Self Employment Income Amount	9.012%	1,154	28	High
	p1ograde	8b - Person 1: Grade Level	8.955%	2,870	25	High
	p1selfe	31b - Person 1: Self Employment Income Amount	8.905%	1,123	30	High
	p2addr_1	22a - Person 2: Work Address	8.880%	4,831	21	High
	p3empl_1	27a - Person 3: Employer	8.289%	1,315	22	High

Form Name	Field Name	Description	Nonblank Error %	Total Nonblank Records	RCC	Outlier
d2	p2ograde	8b - Person 2: Grade Level	8.213%	2,642	25	High
	p1ograde	8b - Person 1: Grade Level	8.212%	2,058	26	High
	p1ograde	8b - Person 1: Grade Level	8.102%	2,888	32	High
	p1ograde	8b - Person 1: Grade Level	8.099%	2,025	22	High
	p1ograde	8b - Person 1: Grade Level	8.075%	2,390	28	High
	p1ograde	8b - Person 1: Grade Level	7.942%	2,531	30	High
	p1ograde	8b - Person 1: Grade Level	7.902%	2,050	21	High
	p3empl_1	27a - Person 3: Employer	7.893%	1,495	32	High
	p1lvcity	15b - Person 1: Migration City	7.844%	4,628	22	High
	p2ograde	8b - Person 2: Grade Level	7.706%	2,232	27	High
	p1_other	31h - Person 1: Other Income Amount	7.705%	1,259	23	High
	p1retir	31g - Person 1: Retirement Income Amount	7.663%	1,579	30	High
d2e	rilast	R1 - Respondent's Last Name	21.410%	9,827	32	Very High
	rilast	R1 - Respondent's Last Name	21.240%	9,642	22	Very High
	rilast	R1 - Respondent's Last Name	19.361%	16,146	26	Very High
	rilast	R1 - Respondent's Last Name	19.044%	15,202	25	Very High
	rilast	R1 - Respondent's Last Name	18.196%	11,596	27	Very High
	rilast	R1 - Respondent's Last Name	18.035%	16,224	29	Very High
	rilast	R1 - Respondent's Last Name	17.595%	12,765	23	Very High
	p1stx16a	H2 - Street Name	17.217%	3,270	31	Very High
	p1stx16a	H2 - Street Name	17.182%	5,785	26	Very High
	rilast	R1 - Respondent's Last Name	16.991%	13,354	31	Very High
	p1stx16a	H2 - Street Name	16.823%	1,064	27	Very High
	rifirst	R1 - Respondent's First Name	16.287%	9,670	22	Very High
	rilast	R1 - Respondent's Last Name	15.928%	12,594	21	Very High
	p1oserve	20b - Person 1: When on Active Duty	15.811%	1,246	25	Very High
	rilast	R1 - Respondent's Last Name	15.795%	17,822	30	Very High
	rilast	R1 - Respondent's Last Name	15.174%	18,143	28	Very High
	rifirst	R1 - Respondent's First Name	14.669%	10,089	32	Very High
	rilast	R1 - Respondent's Last Name	14.628%	13,242	24	Very High
	p1oserve	20b - Person 1: When on Active Duty	14.031%	1,461	26	Very High
	p1oserve	20b - Person 1: When on Active Duty	13.909%	1,215	31	Very High
	p3owork	30a - Person 3: Work Last Year	13.391%	1,165	25	Very High
	rifirst	R1 - Respondent's First Name	13.088%	16,168	26	High
	rifirst	R1 - Respondent's First Name	12.905%	16,621	29	High
	rifirst	R1 - Respondent's First Name	12.551%	15,250	25	High
	rifirst	R1 - Respondent's First Name	12.138%	17,870	30	High
	rifirst	R1 - Respondent's First Name	12.064%	11,903	27	High
	p3oetype	29 - Person 3: Class of Worker	11.875%	1,120	25	High
	p3ograND	19a - Person 3: Grandchildren	11.792%	1,696	25	High
	rifirst	R1 - Respondent's First Name	11.679%	13,135	23	High
	rifirst	R1 - Respondent's First Name	11.524%	18,171	28	High
	rifirst	R1 - Respondent's First Name	11.362%	13,395	31	High
	p1oserve	20b - Person 1: When on Active Duty	10.983%	1,211	27	High
	p1oserve	20b - Person 1: When on Active Duty	10.831%	1,228	30	High
	rifirst	R1 - Respondent's First Name	10.748%	12,607	21	High

Form Name	Field Name	Description	Nonblank Error %	Total Nonblank Records	RCC	Outlier
d2e	p1oserve	20b - Person 1: When on Active Duty	9.515%	1,608	28	High
	p1stx16a	H2 - Street Name	9.431%	4,379	30	High
	rifirst	R1 - Respondent's First Name	9.187%	13,573	24	High
	p1stx16a	H2 - Street Name	9.011%	2,952	29	High
	p3oetype	29 - Person 3: Class of Worker	8.948%	1,017	26	High
	p3oetype	29 - Person 3: Class of Worker	8.761%	1,130	28	High
	p4odegre	9 - Person 4: Highest Degree Completed	8.742%	2,345	26	High
	p3oetype	29 - Person 3: Class of Worker	8.481%	1,014	31	High
	p4odegre	9 - Person 4: Highest Degree Completed	8.368%	2,175	31	High
	p3oetype	29 - Person 3: Class of Worker	8.276%	1,160	30	High
	p1addr_1	22a - Person 1: Work Address	8.082%	2,747	22	High
	p2addr_1	22a - Person 2: Work Address	7.645%	1,452	22	High
	p4odegre	9 - Person 4: Highest Degree Completed	7.633%	2,083	27	High
	p4ograde	8b - Person 4: Grade Level	7.615%	1,602	25	High
d2u	p1stx16a	H2 - Street Name	13.219%	1,929	25	Very High
	p1addr_1	22a - Person 1: Work Address	12.033%	2,327	24	High
	p1stx16a	H2 - Street Name	11.397%	1,009	27	High
	p2addr_1	22a - Person 2: Work Address	10.671%	1,565	24	High
	p1oagric	44c - Household: Agricultural Products	10.518%	3,109	31	High
	p1oagric	44c - Household: Agricultural Products	10.301%	6,873	26	High
	p1addr_1	22a - Person 1: Work Address	10.266%	2,104	23	High
	p1stx16a	H2 - Street Name	10.189%	6,046	26	High
	p1stx16a	H2 - Street Name	10.185%	1,787	23	High
	p1addr_1	22a - Person 1: Work Address	10.154%	2,206	25	High
	p1stx16a	H2 - Street Name	9.921%	3,810	30	High
	p1stx16a	H2 - Street Name	9.820%	3,279	31	High
	p2addr_1	22a - Person 2: Work Address	9.530%	1,574	25	High
	p2addr_1	22a - Person 2: Work Address	9.121%	1,491	23	High
	p1addr_1	22a - Person 1: Work Address	9.058%	6,966	26	High
	p1stx16a	H2 - Street Name	8.978%	4,600	28	High
	p1addr_1	22a - Person 1: Work Address	8.910%	3,816	31	High
	p1addr_1	22a - Person 1: Work Address	8.784%	3,199	21	High
	p1addr_1	22a - Person 1: Work Address	8.484%	3,041	30	High
	p2addr_1	22a - Person 2: Work Address	8.453%	2,579	31	High
	p1stx16a	H2 - Street Name	8.346%	2,624	21	High
	p1ograde	8b - Person 1: Grade Level	8.324%	1,802	26	High
	p2addr_1	22a - Person 2: Work Address	8.317%	2,020	30	High
	p1oagric	44c - Household: Agricultural Products	8.033%	1,805	25	High
	p2addr_1	22a - Person 2: Work Address	7.975%	4,978	26	High
	p2addr_1	22a - Person 2: Work Address	7.865%	2,225	21	High
	p1addr_1	22a - Person 1: Work Address	7.841%	4,515	28	High
	p1hsn10a	H2 - House Number	7.687%	1,353	25	High
	p2addr_1	22a - Person 2: Work Address	7.660%	3,068	28	High
d2ur	p1oauto	43 - Household: Number of Automobiles	72.310%	1,589	21	Very High
	p1obdrm	38 - Household: Number of Bedrooms	71.420%	1,578	21	Very High
	p1lang	11b - Person 1: Language	48.247%	1,198	21	Very High
	p1stx16a	H2 - Street Name	19.272%	1,126	21	Very High
	p2last	1 - Person 2: Last Name	9.111%	1,383	21	High
	p2ohisp	5 - Person 2: Hispanic Origin	8.676%	1,360	21	High

Table 26. Field Category Nonblank Error Rates by Census 2000 Regional Census Center

Census 2000 RCC	Field Category	Nonblank Error %	Outlier
21	Form Management	3.070%	
	POP--Name	3.029%	
	POP--Occupation	2.221%	
	Special Housing	2.107%	
	POP--Military	1.556%	
	Housing Profile	1.525%	
	POP--Ethnic	1.397%	
	POP--Education	1.347%	
	POP--Income	1.293%	
	POP--Demographic	1.034%	
	POP--Race	0.696%	
	POP--Disability	0.674%	
	Coverage	0.453%	
22	POP--Name	4.441%	High
	Form Management	4.071%	High
	Special Housing	3.422%	
	POP--Occupation	2.618%	
	POP--Ethnic	1.878%	
	POP--Military	1.719%	
	POP--Education	1.669%	
	POP--Race	1.510%	
	POP--Income	1.403%	
	Housing Profile	1.339%	
	POP--Demographic	1.071%	
	POP--Disability	0.720%	
	Coverage	0.583%	
23	POP--Name	3.879%	High
	Form Management	3.425%	
	POP--Occupation	3.102%	
	POP--Ethnic	2.759%	
	Special Housing	2.302%	
	POP--Income	2.110%	
	POP--Military	1.922%	
	POP--Education	1.571%	
	Housing Profile	1.321%	
	POP--Demographic	1.062%	
	Coverage	0.465%	
	POP--Race	0.404%	
	POP--Disability	0.368%	

Census 2000 RCC	Field Category	Nonblank Error %	Outlier
24	POP--Name	3.233%	
	POP--Ethnic	3.177%	
	Form Management	3.098%	
	POP--Occupation	2.579%	
	Special Housing	2.326%	
	POP--Income	2.127%	
	POP--Education	1.581%	
	Housing Profile	1.368%	
	POP--Demographic	1.102%	
	POP--Race	1.094%	
	POP--Military	0.543%	
	Coverage	0.464%	
	POP--Disability	0.387%	
25	Form Management	3.429%	
	POP--Name	3.230%	
	POP--Occupation	2.424%	
	POP--Military	2.276%	
	Special Housing	1.994%	
	POP--Education	1.894%	
	POP--Income	1.593%	
	Housing Profile	1.452%	
	POP--Ethnic	1.441%	
	POP--Demographic	1.070%	
	POP--Disability	1.067%	
	POP--Race	0.903%	
	Coverage	0.531%	
26	Form Management	3.445%	High
	POP--Name	2.952%	
	POP--Occupation	2.199%	
	POP--Military	1.885%	
	Special Housing	1.665%	
	POP--Education	1.633%	
	POP--Income	1.389%	
	Housing Profile	1.350%	
	POP--Ethnic	1.152%	
	POP--Demographic	1.049%	
	POP--Race	0.718%	
	POP--Disability	0.705%	
	Coverage	0.526%	

Census 2000 RCC	Field Category	Nonblank Error %	Outlier
27	POP--Military	10.983%	Very High
	POP--Name	3.850%	High
	Form Management	3.421%	
	POP--Occupation	3.364%	
	Special Housing	3.245%	
	POP--Ethnic	2.223%	
	POP--Education	1.685%	
	POP--Income	1.518%	
	Housing Profile	1.328%	
	POP--Demographic	1.123%	
	POP--Race	0.852%	
	POP--Disability	0.606%	
	Coverage	0.419%	
28	Form Management	3.270%	
	POP--Name	2.886%	
	POP--Occupation	2.085%	
	Special Housing	1.988%	
	POP--Military	1.882%	
	POP--Income	1.499%	
	POP--Education	1.489%	
	Housing Profile	1.223%	
	POP--Demographic	0.954%	
	POP--Ethnic	0.909%	
	POP--Disability	0.707%	
	POP--Race	0.599%	
	Coverage	0.367%	
29	POP--Name	4.392%	High
	Form Management	3.354%	
	POP--Occupation	3.221%	
	Special Housing	2.163%	
	POP--Education	1.771%	
	POP--Income	1.297%	
	Housing Profile	1.270%	
	POP--Demographic	1.086%	
	POP--Disability	0.920%	
	POP--Ethnic	0.718%	
	Coverage	0.633%	
	POP--Race	0.403%	
	POP--Military	0.343%	

Census 2000 RCC	Field Category	Nonblank Error %	Outlier
30	Form Management	3.469%	High
	POP--Name	3.272%	
	POP--Occupation	2.163%	
	Special Housing	2.032%	
	POP--Military	1.835%	
	POP--Income	1.524%	
	POP--Education	1.503%	
	Housing Profile	1.364%	
	POP--Ethnic	1.180%	
	POP--Demographic	1.005%	
	POP--Race	0.992%	
	POP--Disability	0.737%	
	Coverage	0.366%	
31	Form Management	2.960%	
	POP--Name	2.944%	
	POP--Occupation	2.263%	
	POP--Military	2.070%	
	Special Housing	1.784%	
	POP--Education	1.742%	
	Housing Profile	1.312%	
	POP--Income	1.296%	
	POP--Ethnic	1.188%	
	POP--Demographic	0.990%	
	POP--Race	0.984%	
	POP--Disability	0.728%	
	Coverage	0.486%	
32	POP--Name	4.016%	High
	Form Management	3.948%	High
	POP--Occupation	3.874%	High
	POP--Ethnic	3.122%	
	POP--Education	2.071%	
	POP--Income	1.876%	
	Special Housing	1.818%	
	Housing Profile	1.491%	
	POP--Race	1.259%	
	POP--Demographic	1.236%	
	POP--Military	0.491%	
	POP--Disability	0.485%	
	Coverage	0.465%	

From Table 26, we see field categories that are high outliers in regional census centers 22, 23, 26, 27, 29, 30, and 32. The outlying categories are consistently Form Management and POP–Name. Form Management includes the contact information and person added/canceled fields on the enumerator forms. Studying Table 25, we find the outliers in this field category are concentrated in the contact information fields. Fields for information on the addition or cancellation of persons do not appear. We find this last observation encouraging. The RCC's with the outliers correspond to the following geographic areas:

- **22 covered northern New Jersey and metropolitan New York City;**
- **23 covered Delaware, the District of Columbia, Maryland, southern New Jersey, and**
  Pennsylvania;
- **26 covered Arkansas, Iowa, Kansas, Minnesota, Missouri, and Oklahoma;**
- **27 covered Alaska, northern California, Idaho, Oregon, and Washington state;**
- **29 covered Alabama, Florida, and Georgia;**
- **30 covered Louisiana, Mississippi, and Texas; and**
- **32 covered southern California and Hawaii.**

Regional census centers 22, 23, 27, 29, and 32 cover Florida, Los Angeles, and New York City. These are areas with above average concentrations of immigrants. Immigrants of non-European extraction tend to have names with unusual spellings. Limited English skills of first generation immigrants may lead to poor handwriting. Either condition could present a challenge to the automated technology and might account at least partly for high soft match error rates in POP–Name fields from these RCC's.

4.8 Analysis of KFI Impact on Soft Match Error Rates

4.8.1 Contents of This Section

In this section, we use a new grouping of the data called KFI Impact to analyze the soft match error rates. In the previous section, we were concerned about how the nonblank error rate behaved depending on

- form (whether we are dealing with a d1, d2, etc.),
- field category (whether we are dealing with name fields, race fields, etc.),
- field (whether we are dealing with name data for person 1, person 2, etc), and
- Census 2000 regional census center (21, 22, and so on up to 32).

Our basic question in this section is this: does the nonblank error rate vary in a significant way depending on what form, field category, type of field, and KFI impact we are talking about? To answer this question, we construct an analysis of variance (ANOVA) where the nonblank error rate is the response variable and the independent variables are form, field category, field, and KFI impact.

In this section, we also distinguish between person and nonperson fields as discussed in section 4.4.1. For definitions of common or special terms in this section, see the glossary in Appendix M.

KFI as a mode of data capture is explained in detail in section 4.5.2. We will summarize and repeat the explanation here for convenience.

Occasionally during Census 2000 processing, the automated technology rejected the content it read for a field if it did not meet a minimum threshold for confidence. Confidence is the technology's estimate of the probability it has captured intelligible content. The technology estimates by comparing the electronic profile of the content to a stored library of patterns.

In cases of content rejected by the technology, a human operator would examine the information on the form and key in a response manually. The keyed content passed through the rest of Census 2000 processing as the response for the corresponding field. We refer to this keying operation in this evaluation as KFI for "Key From Image."

The raw data for this evaluation are a combination of fields that the automated technology accepted and the fields processed by KFI. This section focuses on the question of whether our ability under KFI to capture the intent of the respondent affects the chance of a soft match error. Our attention is restricted to fields for write-in responses. Write-in responses are more challenging to capture automatically than check-boxes. They are more likely to require KFI. Since we are concerned only with write-in responses, we cannot consider hard match errors since they occur only for check-box fields.

KFI has four possible impacts on our ability to capture intent:

- it can improve it,
- it can worsen it, and
- it can be unnecessary in two ways.

It is also possible to perform KFI and not be able to determine what its impact is. To determine the impact of KFI, either the content rejected by the technology or the content supplied by KFI has to match the content intended by the respondent. In this evaluation, for purposes of determining the impact of KFI, the match has to be character by character. We ignore any trailing blanks.

We need to elaborate some on how KFI can be unnecessary. First, the automated technology may reject content in error. If the content matches what the respondent intended, but the automated technology reads it in error, KFI is triggered unnecessarily.

Second, the automated technology may reject content it should reject. KFI is triggered, and the operator enters what he or she believes the respondent meant. The operator's belief, however, may be mistaken. In this situation, we have content the technology refused to accept and an operator-provided response that is not what the respondent intended. KFI brings us no closer to understanding what the respondent meant and so can be considered unnecessary.

Table 27 summarizes the possible impacts of KFI.

Table 27 Determining the Impact of KFI

If the automated technology...	and if the KFI content	and if the content intended by the respondent...	then we conclude....
incorrectly rejects content	matches the rejected content character for character except for	matches the KFI content character for character except	KFI was unnecessary, case 1
	does not match the rejected content character for character	does not match the KFI content character for character	KFI worsened our ability to capture
correctly rejects content	does not match the rejected content character for character	matches the KFI content character for character except	KFI improved our ability to capture
		does not match the KFI content character for character	the impact of KFI cannot be determined
	matches the rejected content character for character except for	does not match the KFI content character for character	KFI was unnecessary, case 2

We are grateful if KFI improves our ability to capture the intent of the respondent. At least we hope for no negative impact. What is unacceptable is for KFI to improve our ability to capture intent at the risk of a higher soft match error rate. We analyze the soft match error rates over the various ways KFI affected our ability to capture intent. If the soft match errors in the "KFI improves" cases are not significantly higher compared to the other KFI impacts, we conclude KFI is safe with respect to soft match errors.

After the ANOVA, we show Tables 32 and 33. The data for the tables are the same as for the ANOVA. After going through the different combinations of forms, fields, and KFI impact, we have a raw data set consisting of 2,787 soft match error rates for the ANOVA and the tables.

In Table 32, we show nonblank error rates that are outliers for specific fields on specific forms. We aim for a sufficiently fine level of detail that makes it easy to identify the largest improvement opportunities.

Table 33 complements Table 32. We aim for a higher level of detail that supports a meaningful overall view of the data. We show the nonblank error rates for each field category. We show a separate field category result for each of the varieties of KFI impact in our data. Any outliers in Table 33 identify field categories that stand out in terms of a high error rate.

The method for testing statistical significance follows sections 4.4.3 and 4.4.4. The details concerning the calculation of errors follows section 4.2.2. The rules concerning the determination of outliers is as described in section 4.3.

4.8.2 Factors and Models for Testing Statistical Significance

Our factors for testing statistical significance are KFI impact, form, field, field category, and person number. We regard these factors as fixed. For more details about the significance testing, see Appendix J. We analyze nonperson fields for statistical significance separately from person fields. For nonperson fields, our model is

- field nested within field category,
- field category nested within form, and
- KFI impact crossed with field.

For person fields, our model is

- person number nested within field,
- field nested within field category,
- field category nested within form, and
- KFI impact crossed with field.

We present four analyses:

- nonperson fields excluding all outliers
- nonperson fields including all outliers
- person fields excluding all outliers
- person fields including all outliers.

4.8.3 Significance Testing for Nonperson Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 28a. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	73	30633.88219	419.64222	65.91	**<0.0001**
Error	45	286.50088	6.36669		
Corrected Total	118	30920.38307			

Table 28b. ANOVA For Nonblank Error Rates For Nonperson Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	644.7316550	58.6119686	9.21	**<0.0001**
Field Category	4	176.6871672	44.1717918	6.94	**0.0002**
Field	NA	NA			
KFI Impact	2	4.8571366	2.4285683	0.38	0.6851
Field*KFI Impact	13	44.2431523	3.4033194	0.53	0.8903

Table 29a. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	78	66425.41379	851.60787	93.12	**<0.0001**
Error	58	530.39885	9.14481		
Corrected Total	136	66955.81264			

Table 29b. ANOVA For Nonblank Error Rates For Nonperson Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	1045.379517	95.034502	10.39	<**0.0001**
Field Category	6	547.856047	91.309341	9.98	<**0.0001**
Field	NA	NA			
KFI Impact	2	4.645587	2.322793	0.25	0.7765
Field*KFI Impact	17	49.003084	2.882534	0.32	0.9946

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank soft match error rate and the factors included in our model. The tables agree as to which individual factors are significant. For nonperson fields, the largest significant factor is form. There is a significant secondary contribution of field category. The structure of the data set did not allow SAS to test field for significance.

4.8.4 Significance Testing for Person Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 30a. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	878	109591.3100	124.8193	8.55	**<0.0001**
Error	1520	22187.6992	14.5972		
Corrected Total	2398	131779.0092			

Table 30b. ANOVA For Nonblank Error Rates For Person Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	9	412.949576	45.883286	3.14	**0.0009**
Field Category	34	772.369355	22.716746	1.56	**0.0220**
Field	NA	NA			
Person Number	NA	NA			
KFI Impact	3	1646.504390	548.834797	37.60	**<0.0001**
Field*KFI Impact	472	8129.368080	17.223237	1.18	**0.0118**

Table 31a. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	919	134310.9182	146.1490	9.98	**<0.0001**
Error	1730	25330.5326	14.6419		
Corrected Total	2649	159641.4508			

Table 31b. ANOVA For Nonblank Error Rates For Person Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	9	735.03850	81.67094	5.58	**<0.0001**
Field Category	35	1270.67313	36.30495	2.48	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			
KFI Impact	3	214.54969	71.51656	4.88	**0.0022**
Field*KFIImpact	495	10860.84229	21.94110	1.50	**<0.0001**

Regardless of whether outliers are included, there is an overall significant relationship

between the nonblank error rate and the factors included in our model. The tables agree as to which individual factors are significant. For person fields, the largest significant factor is the interaction of field and KFI impact. Interaction means that the effect of KFI will change depending on the specific field being considered. Field and KFI impact do not operate independently in their effect on the nonblank soft match error rate. Here is an example to illustrate the interaction of field and KFI impact.

Field	Description	KFI Impact	Nonblank Error %
p1age	**Age of Person 1**	**Redundant, Case 2**	**6.599%**
		Cannot determine	**2.639%**
p1dob_y	**Date of Birth, Person 1**	**Redundant, Case 2**	**3.867%**
		Cannot determine	**4.035%**

The average error rate for "p1age" is higher for the KFI impact value of "Redundant, Case 2" than it is for "Cannot determine." For "p1dob_y", the average error rate for "Redundant, Case 2" is lower than for "Cannot determine." The reversal of the relationship in going from one field to another is a case of an interaction between KFI impact and field.

Besides the above interaction, there are significant secondary contributions of form and field category. The structure of the data set did not allow SAS to test field and person number for significance.

4.8.5 Outlier Data for This Section

As mentioned in section 4.8.1, when we calculate the nonblank error rate for all the combinations of variables relevant to this analysis, we have 2,787 rates by the time we are done. Some of these rates–almost 269--are high or very high outliers according to the procedure discussed in
section 4.3. While we could print the entire table, we prefer to avoid listing entries based on too small a number of cases. After experimenting with different possibilities, we believe 100 records is a reasonable minimum to require for a listing in the table below. This results in Table 32. It consists of 133 outliers. It provides insight into the highest five percent of the nonblank error rates. We believe this emphasizes problem fields that occur often enough to be a priority for investigation and improvement.

Table 32. Field Nonblank Error Rates that are High and Very High Outliers and Based on at Least 100 Blank and Nonblank Data Records

Form Name	Field Name	Description	KFI Impact	Nonblank Error %	Total Nonblank Records	Outlier
d1	p3_relo	Person 3: Other Relative	Cannot determine	11.284%	2,118	High
d1e	rilast	Respondent's Last Name	Unnecessary, Case 2	11.212%	131,961	High
d1s	p5mi	Person 5: Middle Initial	Unnecessary, Case 2	25.000%	208	Very High
	p6mi	Person 6: Middle Initial	Unnecessary, Case 2	24.615%	130	Very High
	p4mi	Person 4: Middle Initial	Cannot determine	22.115%	312	Very High
	p2mi	Person 2: Middle Initial	Cannot determine	21.443%	485	Very High
	p1mi	Person 1: Middle Initial	Cannot determine	21.333%	525	Very High
	p3mi	Person 3: Middle Initial	Cannot determine	21.114%	431	Very High
	p6_relo	Person 6: Other Relative	Cannot determine	19.271%	192	Very High
	p5_relo	Person 5: Other Relative	Cannot determine	18.327%	251	High
	p7last	Person 7: Last Name	Cannot determine	14.948%	194	High
	p2hisp19	Person 2: Other Hispanic Origin	Cannot determine	14.141%	495	High
	p1hisp19	Person 1: Other Hispanic Origin	Cannot determine	13.993%	536	High
	p1last	Person 1: Last Name	Cannot determine	13.875%	1,009	High
	p4last	Person 4: Last Name	Cannot determine	13.854%	628	High
	p1age	Person 1: Age	Cannot determine	13.740%	393	High
	p6last	Person 6: Last Name	Unnecessary, Case 2	13.475%	282	High
	p8first	Person 8: First Name	Cannot determine	13.235%	136	High
	p2last	Person 2: Last Name	Cannot determine	12.603%	968	High
	p1race19	Person 1: Other Race	Cannot determine	12.108%	223	High
d1s	p5last	Person 5: Last Name	Cannot determine	12.081%	447	High
	p3_relo	Person 3: Other Relative	Cannot determine	11.859%	312	High
	p8last	Person 8: Last Name	Cannot determine	11.852%	135	High
	p4_relo	Person 4: Other Relative	Cannot determine	11.498%	287	High
	p3last	Person 3: Last Name	Unnecessary, Case 2	11.442%	874	High
	p1asia19	Person 1: Other Asian	Unnecessary, Case 2	11.111%	153	High
	p3asia19	Person 3: Other Asian	Cannot determine	11.111%	117	High
	p1trib19	Person 1: Am. Indian, AK Native Tribe	Cannot determine	10.881%	386	High
d1u	p1hsn10a	House Number	Cannot determine	16.177%	3,950	High
	p3_relo	Person 3: Other Relative	Cannot determine	14.676%	293	High
	p7last	Person 7: Last Name	Cannot determine	11.968%	376	High
	p1asia19	Person 1: Other Asian	Cannot determine	11.364%	176	High
d2	p4trib_1	Person 4: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	30.460%	1,218	Very High
	p2trib_1	Person 2: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	29.838%	2,785	Very High
	p3trib_1	Person 3: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	28.197%	1,947	Very High
	p2asia_1	Person 2: Other Asian	Unnecessary, Case 2	27.814%	2,301	Very High
	p1asia_1	Person 1: Other Asian	Unnecessary, Case 2	26.512%	2,199	Very High
	p5hisp_1	Person 5: Other Hispanic Origin	Unnecessary, Case 2	25.896%	977	Very High
	p1trib_1	Person 1: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	24.805%	2,689	Very High
	p5trib_1	Person 5: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	24.662%	665	Very High
	p5asia_1	Person 5: Other Asian	Unnecessary, Case 2	23.689%	591	Very High

Form Name	Field Name	Description	KFI Impact	Nonblank Error %	Total Nonblank Records	Outlier
	p4hisp_1	Person 4: Other Hispanic Origin	Unnecessary, Case 2	23.543%	1,699	Very High
	p6trib_1	Person 6: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	22.798%	386	Very High
	p3hisp_1	Person 3: Other Hispanic Origin	Unnecessary, Case 2	22.724%	2,614	Very High
	p4_relo	Person 4: Other Relative	Cannot determine	22.343%	734	Very High
	p4asia_1	Person 4: Other Asian	Unnecessary, Case 2	22.070%	947	Very High
	p1hisp_1	Person 1: Other Hispanic Origin	Unnecessary, Case 2	20.980%	4,428	Very High
	p6hisp_1	Person 6: Other Hispanic Origin	Unnecessary, Case 2	20.598%	602	Very High
	p2race_1	Person 2: Other Race	Unnecessary, Case 2	20.458%	4,414	Very High
	p3race_1	Person 3: Other Race	Unnecessary, Case 2	20.427%	2,952	Very High
	p2hisp_1	Person 2: Other Hispanic Origin	Unnecessary, Case 2	20.423%	3,829	Very High
	p5_relo	Person 5: Other Relative	Cannot determine	20.000%	605	Very High
	p4race_1	Person 4: Other Race	Unnecessary, Case 2	19.355%	2,046	Very High
	p5race_1	Person 5: Other Race	Unnecessary, Case 2	19.292%	1,187	Very High
	p6race_1	Person 6: Other Race	Unnecessary, Case 2	18.155%	672	High
	p3_relo	Person 3: Other Relative	Worse	17.922%	1,328	High
	p6asia_1	Person 6: Other Asian	Unnecessary, Case 2	17.277%	382	High
	p1race_1	Person 1: Other Race	Unnecessary, Case 2	16.792%	4,913	High
	p6_relo	Person 6: Other Relative	Cannot determine	15.418%	467	High
	p3addr_1	Person 3: Work Address	Cannot determine	13.892%	12,907	High
	p3selfe	Person 3: Self Employment Income Amount	Cannot determine	13.826%	745	High
	p2_other	Person 2: Other Income Amount	Cannot determine	13.663%	2,869	High
	p1addr_1	Person 1: Work Address	Cannot determine	13.637%	91,310	High
	p4addr_1	Person 4: Work Address	Cannot determine	13.249%	4,091	High
	p5selfe	Person 5: Self Employment Income Amount	Cannot determine	13.174%	167	High
	p1ssi	Person 1: SSI Amount	Cannot determine	13.068%	5,081	High
	p1_other	Person 1: Other Income Amount	Cannot determine	13.052%	6,681	High
	p5addr_1	Person 5: Work Address	Cannot determine	12.950%	1,390	High
	p2ssi	Person 2: SSI Amount	Cannot determine	12.672%	2,320	High
	p1yrmvus	Person 1: Migration Year	Cannot determine	12.547%	4,264	High
	p2addr_1	Person 2: Work Address	Unnecessary, Case 2	12.520%	56,468	High
	p6addr_1	Person 6: Work Address	Cannot determine	12.018%	649	High
	p1welfr	Person 1: Welfare Amount	Cannot determine	11.976%	2,789	High
	r1last	Roster: Person 1 Last Name	Worse	11.515%	58,706	High
	p2welfr	Person 2: Welfare Amount	Cannot determine	11.503%	1,504	High
	p6int	Person 6: Interest Amount	Cannot determine	11.268%	142	High
	p2selfe	Person 2: Self Employment Income Amount	Cannot determine	11.231%	3,437	High
	p1selfe	Person 1: Self Employment Income Amount	Unnecessary, Case 2	11.127%	6,920	High
	p2_relo	Person 2: Other Relative	Cannot determine	11.114%	3,302	High
	p4empl_1	Person 4: Employer	Cannot determine	11.097%	3,956	High
d2e	rilast	Respondent's Last Name	Unnecessary, Case 2	17.555%	166,529	High
	p5ssi	Person 5: SSI Amount	Cannot determine	15.652%	115	High
	rifirst	Respondent's First Name	Unnecessary, Case 2	12.222%	168,443	High
	p4_relo	Person 4: Other Relative	Worse	12.179%	468	High
	p5_relo	Person 5: Other Relative	Cannot determine	11.485%	357	High
	p3selfe	Person 3: Self Employment Income Amount	Cannot determine	11.215%	107	High
	p5socl	Person 5: Social Security, Railroad Retirement	Cannot determine	11.000%	100	High

Form Name	Field Name	Description	KFI Impact	Nonblank Error %	Total Nonblank Records	Outlier
d2u	p1asia_1	Person 1: Other Asian	Unnecessary, Case 2	22.016%	486	Very High
	p6trib_1	Person 6: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	20.588%	102	Very High
	p4trib_1	Person 4: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	20.260%	385	Very High
	p2asia_1	Person 2: Other Asian	Unnecessary, Case 2	19.083%	545	Very High
	p4_relo	Person 4: Other Relative	Unnecessary, Case 2	18.100%	221	High
	p2trib_1	Person 2: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	17.576%	990	High
	p5_relo	Person 5: Other Relative	Cannot determine	17.333%	150	High
	p5trib_1	Person 5: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	17.010%	194	High
	p2race_1	Person 2: Other Race	Unnecessary, Case 2	16.018%	899	High
	p3trib_1	Person 3: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	15.949%	627	High
	p3asia_1	Person 3: Other Asian	Unnecessary, Case 2	15.932%	295	High
	p2hisp_1	Person 2: Other Hispanic Origin	Unnecessary, Case 2	15.326%	783	High
	p3hisp_1	Person 3: Other Hispanic Origin	Unnecessary, Case 2	14.865%	518	High
	p1trib_1	Person 1: Am Indian, Alaska Native Tribe	Unnecessary, Case 2	14.690%	953	High
	p5hisp_1	Person 5: Other Hispanic Origin	Unnecessary, Case 2	14.535%	172	High
	p1hisp_1	Person 1: Other Hispanic Origin	Unnecessary, Case 2	14.491%	904	High
	p1yrmvus	Person 1: Migration Year	Cannot determine	14.374%	807	High
	p4asia_1	Person 4: Other Asian	Unnecessary, Case 2	13.690%	168	High
	p4hisp_1	Person 4: Other Hispanic Origin	Unnecessary, Case 2	13.003%	323	High
	p3_relo	Person 3: Other Relative	Cannot determine	12.997%	377	High
	p1race_1	Person 1: Other Race	Unnecessary, Case 2	12.879%	924	High
	p3race_1	Person 3: Other Race	Unnecessary, Case 2	12.868%	544	High
	p4race_1	Person 4: Other Race	Unnecessary, Case 2	12.195%	369	High
	p3_other	Person 3: Other Income Amount	Cannot determine	11.679%	274	High
	p3welfr	Person 3: Welfare Amount	Cannot determine	11.340%	194	High
	p3selfe	Person 3: Self Employment Income Amount	Cannot determine	11.111%	270	High
	p1condo	Household: Condo Fee	Worse	10.903%	321	High
d2ur	p2lang	Person 2: Language	Unnecessary, Case 2	68.484%	587	Very High
	p1lang	Person 1: Language	Unnecessary, Case 2	67.950%	805	Very High
	p4lang	Person 4: Language	Unnecessary, Case 2	67.257%	226	Very High
	p3lang	Person 3: Language	Unnecessary, Case 2	66.667%	405	Very High
	p1stx16a	Street Name	Cannot determine	19.272%	1,126	Very High
	p1addr_1	Person 1: Work Address	Cannot determine	18.474%	498	High
	p3addr_1	Person 3: Work Address	Cannot determine	17.054%	129	High
	p2lvcity	Person 2: Migration City	Cannot determine	12.969%	293	High
	p1hsn10a	House Number	Cannot determine	12.796%	719	High
	p1apt16a	Apartment Number	Cannot determine	12.707%	362	High
	p1lvcity	Person 1: Migration City	Cannot determine	12.208%	385	High
	p2addr_1	Person 2: Work Address	Cannot determine	12.027%	291	High
	p2last	Person 2: Last Name	Cannot determine	11.950%	636	High
	p1age	Person 1: Age	Cannot determine	11.818%	110	High
	p1city	Person 1: Work City	Unnecessary, Case 2	11.297%	239	High
	p3empl_1	Person 3: Employer	Cannot determine	11.180%	161	High
	p3last	Person 3: Last Name	Cannot determine	11.086%	442	High
	p3kind_1	Person 3: Occupation Kind of Work	Cannot determine	10.857%	175	High

Table 33. Field Category Nonblank Error Rates by KFI Impact

KFI Impact	Field Category	Nonblank Error %	Outlier
Cannot determine	POP--Income	7.196%	
	POP--Occupation	6.366%	
	POP--Name	6.117%	
	POP--Race	5.969%	
	POP--Ethnic	5.506%	
	Housing Profile	5.322%	
	POP--Demographic	4.797%	
	Special Housing	2.562%	
	Form Management	1.859%	
Unnecessary, Case 1	POP--Name	2.759%	
	POP--Demographic	0.741%	
Unnecessary, Case 2	POP--Race	7.435%	
	Form Management	5.816%	
	POP--Name	2.457%	
	POP--Ethnic	2.230%	
	Special Housing	1.765%	
	POP--Income	1.417%	
	POP--Occupation	1.300%	
	Housing Profile	1.108%	
	POP--Demographic	0.747%	
Worse	POP--Occupation	4.377%	
	POP--Income	4.370%	
	POP--Ethnic	3.957%	
	POP--Name	3.826%	
	POP--Race	3.317%	
	Housing Profile	2.490%	
	Special Housing	2.241%	
	POP--Demographic	1.760%	

From Table 33, we see none of the field categories are outliers. Also, there are no instances in the table where the KFI impact was "Improved." Our primary concern, whether "Improved" is associated with higher soft match error rates, turns out not to be an issue. There were no write-in fields where we simultaneously had a soft match error and an KFI impact of "Improved."

From Table 32, there are some clues to partly explain the interaction of field and KFI impact on the soft match error rate. First, the most frequent category of KFI impact is "Cannot be determined." The automated technology rejected the content, and the entry keyed by the human operator was ultimately not judged to reflect the intent of the respondent, character for character. These are examples of content that tend to be especially difficult to interpret.

Second, many of the outliers on the d1s, the Spanish mailout/mailback short form, are for name fields. It is possible these outliers reflected limits on the capability of the automated technology to understand special Spanish language characters.

Third, many of the outliers on the d2, the English mailout/mailback long form and d2u, the English update/leave long form, are for fields in which respondents write in a race or ethnicity other than the ones provided. This might reflect the increased challenge of interpreting characters written by hand instead of checked off in a box, especially when the handwriting is poor.

The ability of the data capture software to read Spanish language characters might need more evaluation. Another possible improvement is increasing the number of choices respondents can check off for race or ethnicity. The benefit of more choices has to be weighed against the costs of a more complex form.

4.9 Analysis of the Impact of KFI Redundancy on KFI Workload

4.9.1 Contents of This Section

In this section, we are not concerned about error rates but about KFI redundancy rates. KFI redundancy rates measure how often field are sent to KFI unnecessarily. This concept is explained further below. In the previous section, we were concerned about how the nonblank error rate behaved depending on

- form (whether we are dealing with a d1, d2, etc.),
- field category (whether we are dealing with name fields, race fields, etc.),
- field (whether we are dealing with name data for person 1, person 2, etc), and
- KFI impact ("Better", "Worse", and so on as explained in section 4.8.1).

Our basic question in this section is this: does the KFI redundancy rate vary in a significant way depending on what form, field category, type of field, and type of KFI redundancy we are talking about? To answer this question, we construct an analysis of variance (ANOVA) where the KFI redundancy rate is the response variable and the independent variables are form, field category, and field. Unfortunately, type of KFI redundancy does not appear in enough varieties in our raw data to be included as a factor.

In this section, we also distinguish between person and nonperson fields as discussed in section 4.4.1. For definitions of common or special terms in this section, see the glossary in Appendix M. A full explanation of KFI appears in section 4.5.2. An abbreviated one appears in section 4.8.1. For convenience, we repeat the two ways in which KFI can be redundant.

The KFI redundancy data reflects an editing rule in effect at the time of Census 2000 processing. As explained in section 4.5.2, some content went directly to KFI regardless of how confidently the automatic technology judged it as acceptable for processing. If the set of content automatically sent to KFI changes in the future, the behavior of KFI redundancy will change even if the automated technology retains the same hardware and software design.

Table 34. Forms of KFI Redundancy

If the automated technology...	and if the KFI content	and if the content intended by the respondent...	then we conclude....
incorrectly rejects content	matches the rejected content character for character except for trailing blanks	matches the KFI content character for character except for trailing blanks	KFI was redundant, case 1
correctly rejects content	matches the rejected content character for character except for trailing blanks	does not match the KFI content character for character	KFI was redundant, case 2

KFI redundancy is a waste of resources, particularly during the compressed operations of a decennial census. It should be eliminated as much as possible. To progress toward that goal, we must first understand the possible drivers of KFI. We aim to do that here.

After the ANOVA, we show Tables 38 and 39. The data for the tables are the same as for the ANOVA. After going through the different combinations of forms, fields, and types of KFI redundancy, we have a raw data set consisting of 189 redundancy rates for the ANOVA and the tables.

In Table 38, we show nonblank redundancy rates that are outliers for specific fields on specific forms. We aim for a sufficiently fine level of detail that makes it easy to identify the largest improvement opportunities.

Table 39 complements Table 38. We aim for a higher level of detail that supports a meaningful overall view of the data. We show the nonblank redundacy rates for each field category. Any outliers in Table 39 identify field categories that stand out in terms of a high redundancy rate.

The method for testing statistical significance follows sections 4.4.3 and 4.4.4. The details concerning the calculation of redundancy rates follows below. The rules concerning the determination of outliers is as described in section 4.3.

4.9.2 Calculation of the KFI Redundancy Rates

Before proceeding to the analysis, we explain an important contributing concept, the KFI redundancy rate. For each field that went to KFI, we add up the number of times KFI was redundant. This is the numerator of the redundancy rate.

We can compute two redundancy rates: nonblank and total. The denominator of the nonblank

redundancy rate is the number of times the automated technology read content for a field. The denominator for the total redundancy rate is the number of times the automated technology read the field regardless of whether it saw any content. In other words, it includes blank cases.

As long as blanks are occasional occurrences for a field, the nonblank and total redundancy rates will be close. This is the case for the great majority of KFI redundant fields. Fields that are prone to large numbers of blanks will lead to large differences in the redundancy rates. In this latter case, we believe the nonblank error rate is a better measure of data quality. While the automated technology should be given credit for reading blank fields correctly, this is not the same level of challenge as reading nonblank fields correctly. A redundancy rate dominated by a large occurrence of blanks will make redundancy for the corresponding field look better than it probably deserves.

We compute the redundancy rate as 100 x (numerator/denominator). The rates for the Tables 38 and 39 in this section are for nonblank redundancy only.

4.9.3 Factors and Model for Testing Statistical Significance

Our factors for testing statistical significance are form, field, field category, and person number. We regard these factors as fixed. For more details about the significance testing, see Appendix J.

Since KFI redundancy can occur in two varieties, we want to include it as another fixed factor in our model. This would answer whether the effect of the other factors on the KFI redundancy rate depends on which variety of redundancy is being considered. However all of the occurrences of KFI redundancy in our raw data are for only one variety, case 2. We cannot test for statistical significance of a fixed factor when it appears at only one level in the data set. Therefore, we will not include KFI redundancy in our models.

We analyze nonperson fields for statistical significance separately from person fields. For nonperson fields, our model is

- field category nested within form.

For person fields, our model is

- person number nested within field,
- field nested within field category, and
- field category nested within form.

We present three analyses:

- nonperson fields
- person fields excluding all outliers
- person fields including all outliers.

There were no outliers in the nonperson fields so one test for significance will suffice for those.

4.9.4 Significance Testing for Nonperson Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 35a. ANOVA For Nonblank Redundancy Rates For Nonperson Fields, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	65.24864030	8.15608004	69.85	**0.0142**
Error	2	0.23354342	0.11677171		
Corrected Total	10	65.48218372			

Table 35b. ANOVA For Nonblank Redundancy Rates For Nonperson Fields, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	4	1.54355612	0.38588903	3.30	0.2456
Field Category	4	58.12468804	14.53117201	124.44	**0.0080**

There is an overall significant relationship between the nonblank redundancy rate and the factors included in our model. For nonperson fields, the only significant factor is field category. The structure of the data set did not allow SAS to test field for significance.

4.9.5 Significance Testing for Person Fields

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 36a. ANOVA For Nonblank Redundancy Rates For Person Fields Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	133	3018.094226	22.692438	14.85	**<0.0001**
Error	25	38.208794	1.528352		
Corrected Total	158	3056.303020			

Table 36b. ANOVA For Nonblank Redundancy Rates For Person Fields Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	8	37.86735065	4.73341883	3.10	**0.0143**
Field Category	10	84.02753595	8.40275359	5.50	**0.0003**
Field	NA	NA			
Person Number	NA	NA			

Table 37a. ANOVA For Nonblank Redundancy Rates For Person Fields Including Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	142	3141.177920	22.120971	8.96	**<0.0001**
Error	35	86.368502	2.467671		
Corrected Total	177	3227.546422			

Table 37b. ANOVA For Nonblank Redundancy Rates For Person Fields Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	8	56.5926926	7.0740866	2.87	**0.0146**
Field Category	10	116.6160173	11.6616017	4.73	**0.0003**
Field	NA	NA			
Person Number	NA	NA			

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank redundancy rate and the factors included in our model. The tables agree as to which individual factors are significant. For person fields, the largest significant factor is field category. There is a secondary significant association with form. The structure of the data set did not allow SAS to test field and person number for significance.

4.9.6 Outlier Data for This Section

As mentioned in section 4.9.1, when we calculate the nonblank redundancy rate for all the combinations of variables relevant to this analysis, we have 189 rates by the time we are done. Some of these rates–19–are high or very high outliers according to the procedure discussed in section 4.3. While we could print the entire table, we prefer to avoid listing entries based on too small a number of cases. After experimenting with different possibilities, we believe 100 records is a reasonable minimum to require for a listing in the table below. This results in Table 38. It consists of 10 outliers. It provides insight into the highest half of the nonblank redundancy rates. We believe this emphasizes problem fields that occur often enough to be a priority for investigation and improvement.

Table 38. Field Nonblank Redundancy Rates that are High and Very High Outliers and Based on at Least 100 Blank and Nonblank Data Records

Form Name	Field Name	Description	KFI Redundancy	Nonblank Redundancy %	Total Nonblank Records	Outlier
d1	p1dob_y	Person 1: Year of Birth	Redundant, Case 2	4.638%	33,657	Very High
d1s	p5mi	Person 5: Middle Initial	Redundant, Case 2	12.500%	208	Very High
	p6mi	Person 6: Middle Initial	Redundant, Case 2	10.769%	130	Very High
	p6dob_y	Person 6: Year of Birth	Redundant, Case 2	3.593%	167	High
d2e	p5int	Person 5: Interest Amount	Redundant, Case 2	2.913%	103	High
d2u	r7mi	Roster: Person 7 Middle Initial	Redundant, Case 2	4.918%	122	Very High
	p4_relo	Person 4: Other Relative	Redundant, Case 2	4.072%	221	High
	p6mi	Person 6: Middle Initial	Redundant, Case 2	3.020%	298	High
	p1last	Person 1: Last Name	Redundant, Case 2	2.896%	19,923	High
d2ur	p1phpre	Person 1: Phone Number Exchange	Redundant, Case 2	4.848%	165	Very High

Table 39. Field Category Nonblank Redundancy Rates for KFI

KFI Redundancy	Field Category	Nonblank Redundancy %	Outlier
Redundant, Case 2	POP--Name	1.466%	High
	POP--Demographic	1.183%	
	POP--Income	0.936%	
	Housing Profile	0.835%	
	Special Housing	0.478%	
	Form Management	0.341%	
	POP--Occupation	0.316%	
	POP--Race	0.237%	
	POP--Ethnic	0.162%	

From Table 39, we see the field category POP–Name is the only one flagged a high or very high outlier. From Table 38, specific fields in the POP–Name category appear as high or very high outliers for d1s, the Spanish mailout/mailback short form, and d2u, the English update/leave long form, specifically the middle initial for higher numbered persons.

While we do not propose it as the only explanation, respondent fatigue is a possible one for the POP–Name outliers. By the time respondents supply name information for the fifth or sixth person in a household, it is reasonable to suppose accuracy or neatness in the middle initial is not a high priority. Ideally, no field should be sent to KFI redundantly. One practical option with potential to reduce redundant KFI is to experiment with allowing the automated technology greater freedom to adjust its field acceptance criteria according to the particular field being read.

4.10 Analysis of Hard Match Errors in the Person 1 Race Check-Box Field

4.10.1 Contents of This Section

In this section, we return to hard match errors. In the previous section, we were concerned about how the nonblank redundancy rate behaved depending on

- form (whether we are dealing with a d1, d2, etc.),
- field category (whether we are dealing with name fields, race fields, etc.), and
- field (whether we are dealing with name data for person 1, person 2, etc).

Our focus here is restricted to a single field: the race check-box field for person 1. Since many statutory, administrative, and social policy applications of decennial census data depend on an accurate racial profile for the United States, it is proper to dedicate a portion of our analysis to how well the automated technology captures race related fields.

Our basic question in this section is this: does the nonblank error rate for the person 1 race check-box field vary in a significant way depending on what form or race response we are talking about? To answer this question, we construct an analysis of variance (ANOVA) where the nonblank error rate is the response variable and the independent variables are form and race response.

To keep the analysis as simple as possible,

- we look at the race check-box field for only one person on the form, and
- we examine the capture of only five of the more commonly expected responses.

The responses we examine are

- white;
- black, African American, or Negro;
- American Indian or Alaska native;
- the response "Some other race"; and
- cases where a person selects more than one race response.

We believe these limitations are reasonable because we assume any problems the automated technology has with race fields do not depend on which member of the household the response is for or which check-box is selected to indicate race.

In this section, we also distinguish between person and nonperson fields as discussed in section 4.4.1. For definitions of common or special terms in this section, see the glossary in Appendix M.

After the ANOVA, we show Table 42. The data for the tables are the same as for the ANOVA. After going through the different combinations of forms and race responses, we have a raw data set consisting of 18 hard match error rates for the ANOVA and the tables.

In Table 42, we show nonblank error rates that are outliers for specific race responses on specific forms. We aim for a sufficiently fine level of detail that makes it easy to identify the largest improvement opportunities.

The method for testing statistical significance follows sections 4.4.3 and 4.4.4. The details concerning the calculation of errors follows section 4.2.2. The rules concerning the determination of outliers is as described in section 4.3.

4.10.2 Factors and Model for Testing Statistical Significance

Our factors for testing statistical significance are form and race response. We regard these factors as fixed. The race check-box field is a person field. Therefore, nonperson fields are not tested for significance in this section. For more details about the significance testing, see Appendix J. Our model for this section is

- form and
- race response.

We wanted to include the interaction of form with race, but the data set did not have enough observations in the right combinations of form and race to allow this. We present two analyses:

- excluding all outliers
- including all outliers.

4.10.3 Significance Testing for Person 1 Race Check-Box Field

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model". Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table 40a. ANOVA For Nonblank Error Rates For Person 1 Race Check-Box Field

Excluding Outliers, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	356.0236500	39.5581833	20.36	**0.0054**
Error	4	7.7704374	1.9426093		
Corrected Total	13	363.7940874			

Table 40b. ANOVA For Nonblank Error Rates For Person 1 Race Check-Box Field Excluding Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form Name	8	72.3272766	9.0409096	4.65	**0.0771**
Race	1	287.9841750	287.9841750	148.25	**0.0003**

Table 41a. ANOVA For Nonblank Error Rates For Person 1 Race Check-Box Field Including Outliers, Overall Model

Number of observations 18

Note: Due to missing values, only 16 observations can be used in this analysis. The missing values pertain to error rates for combinations of form and race response where the check-box field was read as missing. The computer program interprets this to mean there is no value for the race response variable. We believe this interpretation is sound. As the exclusion only applies to 2 of 2,142 person 1 race check-box fields with hard match errors, we do not feel the exclusion introduces any major distortion.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	487.2319960	54.1368884	11.21	**0.0041**
Error	6	28.9879742	4.8313290		
Corrected Total	15	516.2199702			

Table 41b. Analysis For Nonblank Error Rates For Person 1 Race Check-Box Field Including Outliers, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form Name	8	77.0760150	9.6345019	1.99	0.2080
Race	1	408.7732479	408.7732479	84.61	**<0.0001**

Regardless of whether outliers are included, there is an overall significant relationship between the nonblank error rate and the factors included in our model. The tables do not agree as to which individual factors are significant. Since outliers are known to distort results, it is preferable to conclude based on excluding outliers. The largest significant factor is the race response. There is a significant secondary contribution of form.

4.10.4 Outlier Data for This Section

We are able to show all the nonblank error rates for race, both outliers and nonoutliers. One of the rates is calculated over a denominator of only five nonblank records. Another is calculated over a denominator of only two blank records. We leave these rates out to keep from distorting the table. We show the error rates in descending order.

Table 42. Field Nonblank Error Rates for Person 1 Race Check-box Field

Form Name	Field Name	Description	Race Response Selection	Nonblank Error %	Total Nonblank Records	Outlier
d1	p1orace	Person 1: Race	Other Single	0.194%	227,155	
d1e	p1orace	Person 1: Race	Other Single	0.311%	82,620	
d1s	p1orace	Person 1: Race	Other Single	0.804%	1,865	
d1u	p1orace	Person 1: Race	Other Single	0.054%	38,898	
d1ur	p1orace	Person 1: Race	Other Single	0.038%	2,657	
d2	p1orace	Person 1: Race	Other Single	0.140%	158,393	
d2e	p1orace	Person 1: Race	Other Single	0.271%	104,321	
d2u	p1orace	Person 1: Race	Other Single	0.437%	56,769	
d2ur	p1orace	Person 1: Race	Other Single	0.063%	1,596	

None of the nonblank error rates in the table is an outlier. With the race response testing as significant, the absence of outliers suggests the effect of the race response might be part of an interaction with other factors not included in our ANOVA. The next step from here is to expand the model and test other reasonable factors. We have not pursued this step owing to time constraints. Since the race response will remain an important topic of study for the Census Bureau, it would be helpful for future evaluations of the automated technology to provide for a more extensive analysis of its effect.

4.11 Analysis of Failure to Find Intent & Reasons Why

4.11.1 Contents of This Section

In this section, we switch from hard and soft match errors rates to misinterpretation rates. By misinterpretation, we mean not capturing the intent of the respondent. There are many ways this can happen. For each way, there are many reasons why. The possible manners and reasons for misinterpretation are explained in section 4.11.4. For definitions of common or special terms in this section, see the glossary in Appendix M.

In some previous sections, we explored how the nonblank error rate behaved depending on

- form (whether we are dealing with a d1, d2, etc.),
- field category (whether we are dealing with name fields, race fields, etc.), and
- field (whether we are dealing with name data for person 1, person 2, etc).

Our basic questions in this section are this: (1) In what manner was the intent of the respondent most frequently misinterpreted?, and (2) What were the most frequent reasons for misinterpretation? To answer this question, we define and explain how to calculate misinterpretation rates. This is done in section 4.11.3. Then we present a series of tables that shows misinterpretation rates that are outliers. The tables are broken out by the manner of misinterpretation and the reason for it.

There are four tables. In Table 47, we show misinterpretation rates that are outliers for specific fields on specific forms. We aim for a sufficiently fine level of detail that makes it easy to identify the largest improvement opportunities. The break out in Table 47 is by form, field, mode of data capture, and the manner of misinterpretation.

In Table 48, we aim for a higher level of detail that supports a meaningful overall view of the data. We show misinterpretation rates for each field category. We show a separate field category result for each manner of misinterpretation. Any outliers in Table 48 identify field categories that stand out in terms of a high misinterpretation rate.

After going through the different combinations of forms, fields, modes, and manners of misinterpretation, we have a data set consisting of 13,046 misinterpretation rates. This data set is the source for Tables 47 and 48.

In Table 49 and Table 50, we show a finer break out of the data. For the various ways in which misinterpretation can occur, we present separate rates for the individual reasons why. Table 49 shows misinterpretation rates that are outliers for specific fields on specific forms. As with the Table 47, we aim for a sufficiently fine level of detail that makes it easy to identify the largest improvement opportunities. The break out in Table 49 is by form, field, mode of data capture, and manner of misinterpretation, and reason why.

In Table 50, as in Table 48, we aim for a higher level of detail that supports a meaningful overall view of the data. We show misinterpretation rates for each field category. To save on space, the entries in Table 50 are limited to outliers. These identify field categories that stand out in terms of a high misinterpretation rate. The full list of misinterpretation rates by field category can be found in Appendix I.

After going through the different combinations of forms, fields, modes, manners of misinterpretation, and reasons why, we have a data set consisting of 37,303 misinterpretation rates. This data set is the source for Tables 49 and 50.

The rules concerning the determination of outliers are as described in section 4.3.

4.11.2 Determining the Intent of the Respondent

The intent of the respondent was judged by analysts who worked independently of the Census 2000 processing. They were also independent of the evaluation KFI operation. The analysts based their judgement on the set of rules they were provided with in their training.

If the analysts thought either the automated technology or KFI failed to capture the intent of the respondent, they entered codes into a computer file that eventually became part of the raw data for this evaluation. There were two sets of codes. The analysts picked from one set to identify the type of failure. They picked from another set to identify the reason for the failure.

Occasionally, an analyst found it difficult to determine whether the respondent's intent was captured properly. They consulted their supervisor for help. In our analysis for this section, we sometimes find records showing a decision by both a supervisor and an analyst. In these cases, we use the supervisor's decision. We use the analyst's when that is the only one available.

Within the set of codes for type of failure, some were reserved for write-in fields and the rest were reserved for check-box fields. Within the set of codes used to explain the failures, the situation was a little more complicated. The training materials for the analysts shows the reasons are worded differently depending on whether check-box fields or write-in fields are being considered. However, the substance of the description clearly shows in some cases the same reason could apply to either a check-box or write-in field.

We document the separate lists for check-box fields and write-in fields. We consider Big "X" through person, Poor image, and No reason found to be reasons that apply to both types. After providing the descriptions for error types and error reasons, we use the procedure of Appendix F to identify specific fields and field categories that can be considered high or very high outliers for failure to capture intent.

At the level of individual fields, our error rates are broken out by mode of capture: KFI, OCR, OMR. For an explanation of data capture mode, see section 4.5.2.

4.11.3 Calculation of the Misinterpretation Rates

Before proceeding to the tables, we explain an important contributing concept, the misinterpretation rate. For each field, we add up the number of times the analyst or supervisor concluded the respondent's intent was not captured. This is the numerator of the redundancy rate.
We compute the misinterpretation rate as 100 x (numerator/denominator).

We can compute two misinterpretation rates: nonblank and total. The denominator of the nonblank misinterpretation rate is the number of times the automated technology read content for a field. The denominator for the total misinterpretation rate is the number of times the automated technology read the field regardless of whether it saw any content. In other words, it includes blank cases. For our purposes, we only use nonblank misinterpretation rates in this section.

4.11.4 Manners of Interpretation and the Reasons Why

The ways in which we could misinterpret check-box or write-in fields are described in Tables 43 and 45. Tables 44 and 46 describe the possible reasons why.

Table 43. Possible Ways of Misinterpreting Write-in Fields

Way of Misinterpretation	Description
Extra characters	The output from the automated technology shows more characters than are on the scanned image.
Missing characters	The output from the automated technology has fewer characters than are on the scanned image.
Position reversed	The output from the automated technology and the scanned image have the same number of characters, but two characters in the automated technology output are in reverse order.
Wrong character	The output from the automated technology and the scanned image have the same number of characters, but the output from the automated technology disagrees with the scanned image.
Added response	The output from the automated technology shows content but the scanned image is blank.
Blanked response	The output from the automated technology is blank and the scanned image shows content.

Table 44. Possible Reasons for Misinterpreting Write-In Fields

Reason for Misinterpretation	Description
Poor handwriting	The respondent′s or enumerator's handwriting makes one letter look like another, but one can tell what the respondent meant.
Characters too close	The respondent's or enumerator's characters touch each other, or the respondent tries to squeeze characters in at the end of the field.
Response crossed out	The respondent or enumerator draws a line through the response.
Big "X" through person	The respondent or enumerator draws an "X" through the fields for an entire person. This is an attempt by the respondent to cross out all of the fields.
[illegible]	[illegible]
Decimal point	The respondent wrote a decimal point and it was ignored, or the respondent used an implied decimal point, and it was ignored.
Mixed upper case & lower case letter	The response has both uppercase and lowercase characters.
Spanish accent	The response is in Spanish, and the only difference between the scanned image and the output from the automated technology is an accent on a character.
Character goes out of field	The response is written so part of a character is outside of the spaces for the field.
No reason found	The response is written clearly and there is nothing to suggest why it was not captured correctly.

Table 45. Possible Ways of Misinterpreting Check-box Fields

Way of Misinterpretation	Description
Extra check-box	The output from the automated technology output shows more check-boxes marked than are on the scanned image.
Missing check-box	The output from the automated technology has fewer check-boxes marked than are on the scanned image.
Wrong Character	The output from the automated technology shows the same number of check-boxes marked as on the scanned image, but the boxes are not in the same positions on both.

Table 46. Possible Reasons for Misinterpreting Check-Box Fields

Reason for Misinterpretation	Description
Mark touches another box	The mark from one box hits a second box. This second box is picked up as a response.
Mark Outside box	The respondent's mark is outside of the box. This mark is not picked up as a response.
Box is crossed out	The respondent crosses out a box because he or she made a mistake. The box is picked up as a response.
Stray mark or spot	There is a spot on the paper and it is picked up as a response.
Big "X" through person	The respondent draws an "X" through the fields for an entire person. This is an attempt by the respondent to cross out all of the questions for that person. The check-boxes hit by the"X" are picked up as responses.
Poor image	There is a dark horizontal line drawn across the entire image. The boxes hit by the line are picked up as responses.
No reason found	The response is marked clearly and there is nothing to suggest why it was not captured correctly.

4.11.5 Outlier Rates by Manner of Misinterpretation

As mentioned in section 4.11.1, when we calculate the misinterpretation rate for all the combinations of variables relevant to Table 47, we have 13,046 rates by the time we are done. Some of these rates–almost 2,250--are high or very high outliers according to the procedure discussed in section 4.3. How do we communicate what these outliers have to say without forcing the reader to wade through a 2,250 line table?

We think a fair compromise is to restrict the table to the outliers that are based on a reasonably large number of records. It is hard to conclude much when the data behind an outlier consists of two, three, or some other small number of records. After experimenting with different possibilities, we believe 20,000 records is a reasonable minimum to require. This results in Table 47. It consists of 153 outliers. It provides insight into the highest 1.1 percent of the nonblank error rates. We believe this emphasizes problem fields that occur often enough to be a priority for investigation and improvement. We display the outliers by form, field, mode, and manner of misinterpretation.

Table 47. Field Nonblank Misinterpretation Rates that are High and Very High Outliers, And Based on at Least 20,000 Blank and Nonblank Data Records

Form Name	Field Name	Description	Mode	Type of Error	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d1	p2last	1 - Person 2: Last Name	KFI	Wrong character	3.733%	64,740	Very High
	p1phext	2 - Person 1: Phone Number Digits	KFI	Wrong character	3.448%	24,132	Very High
	p3last	1 - Person 3: Last Name	KFI	Wrong character	3.354%	36,316	High
	p1phpre	2 - Person 1: Phone Number Exchange	KFI	Wrong character	3.341%	20,295	High
	p1last	3 - Person 1: Last Name	KFI	Wrong character	3.185%	85,962	High
	p4last	1 - Person 4: Last Name	KFI	Wrong character	3.113%	21,684	High
	p2first	1 - Person 2: First Name	KFI	Wrong character	2.956%	44,580	High
	p1first	3 - Person 1: First Name	KFI	Wrong character	2.945%	50,770	High
	p3first	1 - Person 3: First Name	KFI	Wrong character	2.716%	27,581	High
	p1dob_y	6 - Person 1: Year of Birth	KFI	Wrong character	1.899%	33,657	High
d1e	rilast	R1 - Respondent's Last Name	OCR	Wrong character	9.873%	131,961	Very High
	rifirst	R1 - Respondent's First Name	OCR	Wrong character	7.153%	133,156	Very High
	p1last	3 - Person 1: Last Name	KFI	Wrong character	3.329%	29,681	High
	p2last	1 - Person 2: Last Name	KFI	Wrong character	3.106%	20,025	High
	p1first	3 - Person 1: First Name	KFI	Wrong character	2.463%	22,293	High
	rilast	R1 - Respondent's Last Name	OCR	Missing characters	2.395%	131,961	High
d2	p1addr_1	22a - Person 1: Work Address	KFI	Missing characters	18.114%	91,310	Very High
	p2addr_1	22a - Person 2: Work Address	KFI	Missing characters	16.135%	56,468	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Missing characters	8.831%	78,439	Very High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Missing characters	8.749%	60,098	Very High
	p2empl_1	27a - Person 2: Employer	KFI	Missing characters	7.943%	51,441	Very High
	p2duty_1	28b - Person 2: Occupation Duties	KFI	Missing characters	7.764%	39,761	Very High
	p1ograde	8b - Person 1: Grade Level	OMR	Extra check-box	7.040%	29,004	Very High
	p1actv_1	27b - Person 1: Industry	KFI	Missing characters	6.659%	52,455	Very High
	p2ograde	8b - Person 2: Grade Level	OMR	Extra check-box	6.207%	26,133	Very High
	p3ethn_1	10 - Person 3: Ancestry	KFI	Missing characters	6.178%	25,996	Very High
	p1lvcity	15b - Person 1: Migration City	KFI	Missing characters	5.703%	40,154	Very High
	p2actv_1	27b - Person 2: Industry	KFI	Missing characters	5.634%	34,312	Very High
	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Missing characters	5.419%	52,833	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Wrong character	5.037%	78,439	Very High
	p2empl_1	27a - Person 2: Employer	KFI	Wrong character	4.739%	51,441	Very High
	p2kind_1	28a - Person 2: Occupation Kind of Work	KFI	Missing characters	4.701%	35,397	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Wrong character	4.665%	91,310	Very High
	r1last	Roster: Person 1 Last Name	KFI	Wrong character	4.613%	58,706	Very High

Form Name	Field Name	Description	Mode	Manner of Misinterpretation	Nonblank Misinterpretation %	Total Nonblank Records	Outliers
d2	p1city	22b - Person 1: Work City	KFI	Missing characters	4.369%	40,145	Very High
	r2last	Roster: Person 2 Last Name	KFI	Wrong character	4.273%	41,376	Very High
	p1orecal	25c - Person 1: Will Be Recalled	OMR	Extra check-box	4.249%	21,698	Very High
	r3last	Roster: Person 3 Last Name	KFI	Wrong character	4.079%	23,484	Very High
	p2addr_1	22a - Person 2: Work Address	KFI	Wrong character	3.993%	56,468	Very High
	p3last	1 - Person 3: Last Name	KFI	Wrong character	3.985%	25,820	Very High
	p2lvcity	15b - Person 2: Migration City	KFI	Missing characters	3.983%	27,617	Very High
	r2first	Roster: Person 2 First Name	KFI	Wrong character	3.927%	27,654	Very High
	p1last	3 - Person 1: Last Name	KFI	Wrong character	3.852%	60,464	Very High
	p1olayof	25a - Person 1: Last Week Layoff	OMR	Extra check-box	3.655%	64,926	Very High
	p1oabsnt	25b - Person 1: Last Week Absent	OMR	Extra check-box	3.607%	57,247	Very High
	p2empl_1	27a - Person 2: Employer	OCR	Wrong character	3.603%	21,512	Very High
	p2last	1 - Person 2: Last Name	KFI	Wrong character	3.595%	45,652	Very High
	p2first	1 - Person 2: First Name	KFI	Wrong character	3.589%	31,734	Very High
	r1first	Roster: Person 1 First Name	KFI	Wrong character	3.423%	33,539	Very High
	p2city	22b - Person 2: Work City	KFI	Missing characters	3.362%	24,928	High
	p1empl_1	27a - Person 1: Employer	OCR	Wrong character	3.310%	32,119	High
	p3oalone	17c - Person 3: Difficulty Shopping	OMR	Extra check-box	3.231%	41,222	High
	p2ethn_1	10 - Person 2: Ancestry	KFI	Missing characters	3.220%	40,810	High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Wrong character	3.211%	60,098	High
	p1first	3 - Person 1: First Name	KFI	Wrong character	3.188%	36,671	High
	p1ethn_1	10 - Person 1: Ancestry	KFI	Missing characters	3.052%	50,779	High
	p1olook	25d - Person 1: Looking for Work	OMR	Extra check-box	3.021%	54,159	High
	p1lvcity	15b - Person 1: Migration City	KFI	Wrong character	3.011%	40,154	High
	p1total	32 - Person 1: Total Income Amount	KFI	Wrong character	2.990%	46,552	High
	p2olayof	25a - Person 2: Last Week Layoff	OMR	Extra check-box	2.974%	54,031	High
	p1zip	22f - Person 1: Work Zip Code	KFI	Wrong character	2.872%	20,888	High
	p2duty_1	28b - Person 2: Occupation Duties	KFI	Wrong character	2.812%	39,761	High
	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Wrong character	2.796%	52,833	High
	p1city	22b - Person 1: Work City	KFI	Wrong character	2.792%	40,145	High
	p1elec	45a - Household: Electricity Cost	KFI	Wrong character	2.769%	41,926	High
	p2lvcity	15b - Person 2: Migration City	KFI	Wrong character	2.766%	27,617	High
	r1mi	Roster: Person 1 Middle Initial	KFI	Wrong character	2.756%	25,327	High
	p1actv_1	27b - Person 1: Industry	KFI	Wrong character	2.726%	52,455	High
	p4oalone	17c - Person 4: Difficulty Shopping	OMR	Extra check-box	2.726%	20,212	High
	p1county	22d - Person 1: Work County	KFI	Wrong character	2.722%	26,338	High
	p1lvcnty	15b - Person 1: Migration County	KFI	Wrong character	2.722%	23,185	High
	p1mi	3 - Person 1: Middle Initial	KFI	Wrong character	2.648%	28,285	High
	p2oabsnt	25b - Person 2: Last Week Absent	OMR	Extra check-box	2.645%	48,012	High
	p1bnus	12 - Person 1: Name of State	KFI	Missing characters	2.637%	35,453	High
	p2kind_1	28a - Person 2: Occupation Kind of Work	KFI	Wrong character	2.599%	35,397	High
	p2city	22b - Person 2: Work City	KFI	Wrong character	2.595%	24,928	High
	p1wages	31a - Person 1: Wages Amount	KFI	Wrong character	2.594%	37,775	High
	p2total	32 - Person 2: Total Income Amount	KFI	Wrong character	2.348%	24,272	High
	p2actv_1	27b - Person 2: Industry	KFI	Wrong character	2.320%	34,312	High
	p1int	31c - Person 1: Interest Amount	KFI	Wrong character	2.279%	22,734	High
	p1empl_1	27a - Person 1: Employer	OCR	Missing characters	2.142%	32,119	High
	p2bnus	12 - Person 2: Name of State	KFI	Missing characters	2.140%	29,211	High
	p1gas	45b - Household: Gas Cost	KFI	Wrong character	2.100%	23,862	High

Form Name	Field Name	Description	Mode	Manner of Misinterpretation	Nonblank Misinterpretation %	Total Nonblank Records	Outliers
d2	p1oserve	20b - Person 1: When on Active Duty	OMR	Missing check-box	2.023%	36,934	High
	p1esttax	49 - Household: Real Estate Tax Amount	KFI	Wrong character	1.996%	29,505	High
	p1water	45c - Household: Water and Sewer Cost	KFI	Wrong character	1.989%	22,824	High
	p1oneeds	19b - Person 1: Responsible for Needs	OMR	Extra check-box	1.949%	29,201	High
	p2wages	31a - Person 2: Wages Amount	KFI	Wrong character	1.913%	21,220	High
	p2olook	25d - Person 2: Looking for Work	OMR	Extra check-box	1.898%	45,089	High
	p1flood	50 - Household: Insurance Payment	KFI	Wrong character	1.880%	27,760	High
	p3ojob	17d - Person 3: Difficulty Working	OMR	Extra check-box	1.864%	39,116	High
	p3ospkwl	11c - Person 3: Speak English Well	OMR	Extra check-box	1.855%	23,235	High
	p1lvcnty	15b - Person 1: Migration County	KFI	Missing characters	1.829%	23,185	High
d2e	rilast	R1 - Respondent's Last Name	OCR	Wrong character	17.286%	166,529	Very High
	rifirst	R1 - Respondent's First Name	OCR	Wrong character	12.080%	168,443	Very High
	p1stab2a	H2 - State	OCR	Wrong character	6.107%	21,386	Very High
	p1last	3 - Person 1: Last Name	KFI	Wrong character	5.396%	36,841	Very High
	p1phext	2 - Person 1: Phone Number Digits	KFI	Wrong character	5.338%	23,341	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Missing characters	5.312%	45,994	Very High
	p4odegre	9 - Person 4: Highest Degree Completed	OMR	Extra check-box	5.275%	25,955	Very High
	p2last	1 - Person 2: Last Name	KFI	Wrong character	5.133%	25,796	Very High
	p2empl_1	27a - Person 2: Employer	KFI	Wrong character	5.111%	22,695	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Missing characters	5.002%	36,328	Very High
	p2addr_1	22a - Person 2: Work Address	KFI	Missing characters	4.974%	26,498	Very High
	p2empl_1	27a - Person 2: Employer	KFI	Missing characters	4.776%	22,695	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Wrong character	4.765%	36,328	Very High
	p1lasta	7 - Person 1: Last Name	KFI	Wrong character	4.620%	30,841	Very High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Missing characters	4.555%	27,267	Very High
	p2first	1 - Person 2: First Name	KFI	Wrong character	4.423%	20,575	Very High
	p1first	3 - Person 1: First Name	KFI	Wrong character	3.956%	25,406	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Wrong character	3.840%	45,994	Very High
	p2lasta	7 - Person 1: Last Name	KFI	Wrong character	3.616%	21,679	Very High
	p1actv_1	27b - Person 1: Industry	KFI	Missing characters	3.534%	24,306	Very High
	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Missing characters	3.482%	24,527	Very High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Wrong character	3.411%	27,267	Very High
	p1cty16a	H2 - City	KFI	Wrong character	3.410%	26,660	Very High
	p2addr_1	22a - Person 2: Work Address	KFI	Wrong character	3.396%	26,498	Very High
	p1zip5a	H1- Zip Code	OCR	Wrong character	3.160%	27,819	High
	p1stx16a	H2 - Street Name	KFI	Missing characters	3.114%	33,361	High
	p3ograde	8b - Person 3: Grade Level	OMR	Extra check-box	3.057%	26,789	High
	p1actv_1	27b - Person 1: Industry	KFI	Wrong character	3.016%	24,306	High
	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Wrong character	3.001%	24,527	High
	p1ospkwl	11c - Person 1: Speak English Well	OMR	Extra check-box	2.920%	22,228	High
	p3odegre	9 - Person 3: Highest Degree Completed	OMR	Extra check-box	2.899%	40,433	High
	rilast	R1 - Respondent's Last Name	OCR	Missing characters	2.843%	166,529	High
	p1empl_1	27a - Person 1: Employer	OCR	Wrong character	2.828%	25,598	High
	p1stx16a	H2 - Street Name	KFI	Wrong character	2.794%	33,361	High
	a_status	Summary - A: Status	KFI	Wrong character	2.647%	21,233	High
	p2olayof	25a - Person 2: Last Week Layoff	OMR	Extra check-box	2.548%	30,569	High
	p4octzn	13 - Person 4: Citizen	OMR	Extra check-box	2.537%	25,781	High
	p1ovalue	51 - Household: Property Value	OMR	Extra check-box	2.242%	67,225	High

Form Name	Field Name	Description	Mode	Manner of Misinterpretation	Nonblank Misinterpretation %	Total Nonblank Records	Outliers
d2e	p2olvcty	15b - Person 2: Live Inside City Limits	OMR	Extra check-box	2.199%	26,372	High
	p2oetype	29 - Person 2: Class of Worker	OMR	Extra check-box	2.087%	41,967	High
	c_osumma	Summary - C: Vacant	OMR	Extra check-box	2.082%	48,805	High
	p1otrans	23a - Person 1: Work Vehicle	OMR	Extra check-box	2.007%	59,801	High
	p1ethn_1	10 - Person 1: Ancestry	KFI	Missing characters	1.918%	24,765	High
	rifirst	R1 - Respondent's First Name	OCR	Missing characters	1.889%	168,443	High
	p1oagric	44c - Household: Agricultural Products	OMR	Extra check-box	1.871%	40,449	High
d2u	p1stx16a	H2 - Street Name	KFI	Missing characters	11.713%	29,874	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Missing characters	10.973%	31,150	Very High
	p2addr_1	22a - Person 2: Work Address	KFI	Missing characters	10.142%	21,475	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Missing characters	5.719%	26,981	Very High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Missing characters	5.417%	20,197	Very High
	p1stab2a	H2 - State	OCR	Wrong character	5.312%	20,481	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Wrong character	3.680%	26,981	Very High
	p1hsn10a	H2 - House Number	KFI	Missing characters	3.593%	20,818	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Wrong character	3.339%	31,150	High
	p2addr_1	22a - Person 2: Work Address	KFI	Wrong character	2.710%	21,475	High
	p1stx16a	H2 - Street Name	KFI	Wrong character	2.467%	29,874	High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Wrong character	2.431%	20,197	High
	p1olayof	25a - Person 1: Last Week Layoff	OMR	Extra check-box	2.215%	24,378	High
	p1oabsnt	25b - Person 1: Last Week Absent	OMR	Extra check-box	2.058%	21,867	High
	p2olayof	25a - Person 2: Last Week Layoff	OMR	Extra check-box	1.873%	20,283	High

Table 48. Field Category Error Rates by Manner of Misinterpretation

Field Category	Manner of Misintepretation	Nonblank Misinterpretation %	Outlier
Coverage	Extra check-box	0.128%	
	Wrong check-box	0.007%	
	Missing check-box	0.006%	
Form Management	Wrong character	7.173%	Very High
	Extra check-box	0.404%	
	Missing characters	0.368%	
	Added response	0.145%	
	Extra characters	0.105%	
	Blanked response	0.014%	
	Missing check-box	0.013%	
	Wrong check-box	0.009%	
	Position reversed	0.004%	
Housing Profile	Wrong character	0.879%	High
	Extra check-box	0.500%	
	Missing characters	0.342%	
	Added response	0.140%	
	Extra characters	0.124%	
	Blanked response	0.096%	
	Wrong check-box	0.049%	
	Position reversed	0.034%	
	Missing check-box	0.027%	

Field Category	Manner of Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Demographic	Wrong character	0.746%	
	Extra check-box	0.382%	
	Missing characters	0.287%	
	Extra characters	0.060%	
	Wrong check-box	0.052%	
	Position reversed	0.050%	
	Blanked response	0.049%	
	Added response	0.037%	
	Missing check-box	0.024%	
POP--Disability	Extra check-box	0.498%	
	Wrong check-box	0.025%	
	Missing check-box	0.007%	
POP--Education	Extra check-box	0.971%	High
	Missing check-box	0.113%	
	Wrong check-box	0.067%	
POP--Ethnic	Missing characters	1.730%	Very High
	Wrong character	1.604%	Very High
	Extra characters	0.591%	
	Added response	0.236%	
	Position reversed	0.189%	
	Extra check-box	0.167%	
	Blanked response	0.087%	
	Missing check-box	0.017%	
	Wrong check-box	0.009%	
POP--Income	Wrong character	1.236%	High
	Added response	0.678%	
	Extra check-box	0.551%	
	Missing characters	0.483%	
	Blanked response	0.198%	
	Extra characters	0.191%	
	Wrong check-box	0.036%	
	Position reversed	0.023%	
	Missing check-box	0.011%	
POP--Military	Extra check-box	1.211%	High
	Missing check-box	0.224%	
	Wrong check-box	0.043%	
POP--Name	Wrong character	2.322%	Very High
	Missing characters	0.481%	
	Extra characters	0.156%	
	Blanked response	0.075%	
	Position reversed	0.064%	
	Added response	0.031%	

Field Category	Manner of Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Occupation	Missing characters	2.391%	Very High
	Wrong character	1.665%	Very High
	Extra check-box	1.248%	High
	Extra characters	0.402%	
	Position reversed	0.174%	
	Blanked response	0.087%	
	Wrong check-box	0.051%	
	Added response	0.045%	
	Missing check-box	0.033%	
POP--Race	Wrong character	4.105%	Very High
	Missing characters	2.506%	Very High
	Added response	1.802%	Very High
	Extra characters	0.780%	
	Position reversed	0.255%	
	Blanked response	0.214%	
	Extra check-box	0.171%	
	Missing check-box	0.063%	
	Wrong check-box	0.008%	
Special Housing	Blanked response	0.996%	High
	Added response	0.252%	
	Wrong character	0.159%	
	Missing characters	0.107%	
	Extra characters	0.049%	

As Table 47 shows, at the level of field, the error Wrong character dominates(124 of 195 outliers in table). At the more general level of Table 48, the errors Extra check-box and Wrong character are in one of the top three positions for nine of the 13 categories. Missing characters appears in one of the top three positions for seven of the 13 categories. All these reach to the heart of possible problems with the automated technology. If it misses characters, adds characters that are not there, or substitutes characters, our ability to discern the intent of the respondent decreases. Tables 47 and 48 suggest these problems are not confined to a particular field or field category but rather exist across a wide swath. For more specific comments beyond the general need to improve performance in these areas, we have to look for trends in the reasons for these errors.

4.11.6 Outlier Rates by Reason for Misinterpretation

As mentioned in section 4.11.1, when we calculate the misinterpretation rate for all the combinations of variables relevant to Table 49, we have 37,303 rates by the time we are done. Some of these rates–almost 6,900--are high or very high outliers according to the procedure discussed in section 4.3. How do we communicate what these outliers have to say without forcing the reader to wade through a 6,900 line table?

We think a fair compromise is to restrict the table to the outliers that are based on a reasonably large number of records. It is hard to conclude much when the data behind an outlier consist of two, three, or some other small number of records. After experimenting with different possibilities, we believe 50,000 records is a reasonable minimum to require. This results in Table 49. It consists of 149 outliers. It provides insight into the highest 0.4 percent of the nonblank misinterpretation rates. We believe this emphasizes problem fields that occur often enough to be a priority for investigation and improvement.

Table 49. Field Nonblank Error Rates that are High and Very High Outliers, Broken Out by Mode of Data Capture and Reason for Misinterpretation And Based on at Least 50,000 Blank and Nonblank Data Records

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d1	p2last	1 - Person 2: Last Name	KFI	Wrong character	Poor handwriting	2.343%	64,740	Very High
	p1last	3 - Person 1: Last Name	KFI	Wrong character	Poor handwriting	1.890%	85,962	Very High
	p1first	3 - Person 1: First Name	KFI	Wrong character	Poor handwriting	1.812%	50,770	Very High
	p1last	3 - Person 1: Last Name	KFI	Missing characters	No reason found	0.824%	85,962	High
	p2last	1 - Person 2: Last Name	KFI	Missing characters	No reason found	0.726%	64,740	High
	p1first	3 - Person 1: First Name	KFI	Missing characters	No reason found	0.691%	50,770	High
	p1last	3 - Person 1: Last Name	KFI	Wrong character	No reason found	0.580%	85,962	High
	p1last	3 - Person 1: Last Name	OCR	Wrong character	Poor handwriting	0.549%	148,090	High
	p2last	1 - Person 2: Last Name	KFI	Wrong character	No reason found	0.548%	64,740	High
	p2last	1 - Person 2: Last Name	OCR	Wrong character	Poor handwriting	0.523%	109,321	High
	p1phext	2 - Person 1: Phone Number Digits	OCR	Wrong character	Poor handwriting	0.518%	200,597	High
	p3last	1 - Person 3: Last Name	OCR	Wrong character	Poor handwriting	0.507%	59,951	High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d1e	rilast	R1 - Respondent's Last Name	OCR	Wrong character	Poor handwriting	8.643%	131,961	Very High
	rifirst	R1 - Respondent's First Name	OCR	Wrong character	Poor handwriting	6.296%	133,156	Very High
	rilast	R1 - Respondent's Last Name	OCR	Missing characters	No reason found	1.733%	131,961	Very High
	rifirst	R1 - Respondent's First Name	OCR	Missing characters	No reason found	1.080%	133,156	Very High
	p1phext	2 - Person 1: Phone Number Digits	OCR	Wrong character	Poor handwriting	0.930%	103,022	Very High
	rilast	R1 - Respondent's Last Name	OCR	Wrong character	No reason found	0.805%	131,961	High
	p1pharea	2 - Person 1: Phone Number Area Code	OCR	Wrong character	Poor handwriting	0.775%	107,554	High
	p1phpre	2 - Person 1: Phone Number Exchange	OCR	Wrong character	Poor handwriting	0.680%	107,167	High
	p1last	3 - Person 1: Last Name	OCR	Wrong character	Poor handwriting	0.601%	54,208	High
	rilast	R1 - Respondent's Last Name	OCR	Missing characters	Poor handwriting	0.558%	131,961	High
	rifirst	R1 - Respondent's First Name	OCR	Wrong character	No reason found	0.535%	133,156	High
d2	p1addr_1	22a - Person 1: Work Address	KFI	Missing characters	Rules not followed	12.240%	91,310	Very High
	p2addr_1	22a - Person 2: Work Address	KFI	Missing characters	Rules not followed	11.522%	56,468	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Missing characters	Rules not followed	4.943%	78,439	Very High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Missing characters	No reason found	4.521%	60,098	Very High
	p2empl_1	27a - Person 2: Employer	KFI	Missing characters	Rules not followed	4.366%	51,441	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Wrong character	Poor handwriting	4.041%	78,439	Very High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d2	p2addr_1	22a - Person 2: Work Address	KFI	Missing characters	No reason found	3.974%	56,468	Very High
	p1actv_1	27b - Person 1: Industry	KFI	Missing characters	Rules not followed	3.956%	52,455	Very High
	p2empl_1	27a - Person 2: Employer	KFI	Wrong character	Poor handwriting	3.736%	51,441	Very High
	r1last	Roster: Person 1 Last Name	KFI	Wrong character	Poor handwriting	3.485%	58,706	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Wrong character	Poor handwriting	3.457%	91,310	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Missing characters	No reason found	3.436%	91,310	Very High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Missing characters	Rules not followed	3.419%	60,098	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Missing characters	No reason found	3.287%	78,439	Very High
	p2empl_1	27a - Person 2: Employer	KFI	Missing characters	No reason found	2.996%	51,441	Very High
	p2addr_1	22a - Person 2: Work Address	KFI	Wrong character	Poor handwriting	2.941%	56,468	Very High
	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Missing characters	Rules not followed	2.864%	52,833	Very High
	p1last	3 - Person 1: Last Name	KFI	Wrong character	Poor handwriting	2.856%	60,464	Very High
	p1ethn_1	10 - Person 1: Ancestry	KFI	Missing characters	No reason found	2.470%	50,779	Very High
	p1actv_1	27b - Person 1: Industry	KFI	Missing characters	No reason found	2.377%	52,455	Very High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Wrong character	Poor handwriting	2.363%	60,098	Very High
	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Missing characters	No reason found	2.165%	52,833	Very High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d2	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Wrong character	Poor handwriting	2.110%	52,833	Very High
	p1actv_1	27b - Person 1: Industry	KFI	Wrong character	Poor handwriting	2.072%	52,455	Very High
	p1olayof	25a - Person 1: Last Week Layoff	OMR	Extra check-box	Box is crossed out	2.048%	64,926	Very High
	p1oabsnt	25b - Person 1: Last Week Absent	OMR	Extra check-box	Box is crossed out	1.857%	57,247	Very High
	p2olayof	25a - Person 2: Last Week Layoff	OMR	Extra check-box	Box is crossed out	1.814%	54,031	Very High
	p1olook	25d - Person 1: Looking for Work	OMR	Extra check-box	Box is crossed out	1.490%	54,159	Very High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Extra characters	No reason found	1.298%	60,098	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Missing characters	Character goes out field	1.232%	91,310	Very High
	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Extra characters	No reason found	1.179%	52,833	Very High
	p1empl_1	27a - Person 1: Employer	KFI	Extra characters	No reason found	1.177%	78,439	Very High
	p1ethn_1	10 - Person 1: Ancestry	KFI	Wrong character	Poor handwriting	1.176%	50,779	Very High
	p2last	1 - Person 2: Last Name	OCR	Wrong character	Poor handwriting	1.126%	72,904	Very High
	r1last	Roster: Person 1 Last Name	KFI	Missing characters	No reason found	1.088%	58,706	Very High
	p1actv_1	27b - Person 1: Industry	KFI	Extra characters	No reason found	1.071%	52,455	Very High
	p2empl_1	27a - Person 2: Employer	KFI	Extra characters	No reason found	1.036%	51,441	Very High
	p1last	3 - Person 1: Last Name	OCR	Wrong character	Poor handwriting	0.993%	101,436	Very High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d2	p1last	3 - Person 1: Last Name	KFI	Missing characters	No reason found	0.963%	60,464	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Missing characters	Truncated	0.937%	91,310	Very High
	r2last	Roster: Person 2 Last Name	OCR	Wrong character	Poor handwriting	0.899%	75,513	Very High
	p1olstwk	26 - Person 1: Last Worked	OMR	Extra check-box	Box is crossed out	0.893%	56,465	Very High
	p1oabsnt	25b - Person 1: Last Week Absent	OMR	Extra check-box	Stray mark or spot	0.886%	57,247	Very High
	p1olayof	25a - Person 1: Last Week Layoff	OMR	Extra check-box	Stray mark or spot	0.855%	64,926	Very High
	p1addr_1	22a - Person 1: Work Address	KFI	Extra characters	No reason found	0.847%	91,310	Very High
	p2addr_1	22a - Person 2: Work Address	KFI	Extra characters	No reason found	0.841%	56,468	High
	r1last	Roster: Person 1 Last Name	OCR	Wrong character	Poor handwriting	0.832%	99,939	High
	p1lvzip	15b - Person 1: Migration Zip Code	OCR	Wrong character	Poor handwriting	0.831%	56,299	High
	p1oabsnt	25b - Person 1: Last Week Absent	OMR	Extra check-box	Big X through person	0.805%	57,247	High
	p1zip	22f - Person 1: Work Zip Code	OCR	Wrong character	Poor handwriting	0.785%	65,616	High
	p1owages	31a - Person 1: Wages	OMR	Extra check-box	Box is crossed out	0.777%	115,064	High
	p1ospkwl	11c - Person 1: Speak English Well	OMR	Extra check-box	Box is crossed out	0.774%	53,123	High
	p1actv_1	27b - Person 1: Industry	OCR	Wrong character	Poor handwriting	0.755%	60,104	High
	p1olook	25d - Person 1: Looking for Work	OMR	Extra check-box	Stray mark or spot	0.751%	54,159	High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d2	p1phext	2 - Person 1: Phone Number Digits	OCR	Wrong character	Poor handwriting	0.733%	137,827	High
	p1olook	25d - Person 1: Looking for Work	OMR	Extra check-box	Big X through person	0.727%	54,159	High
	p1ooffce	44a - Household: Business	OMR	Extra check-box	Box is crossed out	0.725%	124,205	High
	p1addr_1	22a - Person 1: Work Address	KFI	Wrong character	No reason found	0.712%	91,310	High
	p1olayof	25a - Person 1: Last Week Layoff	OMR	Extra check-box	Big X through person	0.698%	64,926	High
	p1total	32 - Person 1: Total Income Amount	OCR	Wrong character	Poor handwriting	0.690%	75,101	High
	p1oagric	44c - Household: Agricultural Products	OMR	Extra check-box	Box is crossed out	0.676%	51,605	High
	p2empl_1	27a - Person 2: Employer	KFI	Wrong character	No reason found	0.665%	51,441	High
	p1kind_1	28a - Person 1: Occupation Kind of Work	OCR	Missing characters	No reason found	0.665%	63,873	High
	p1kind_1	28a - Person 1: Occupation Kind of Work	OCR	Wrong character	Poor handwriting	0.664%	63,873	High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Wrong character	No reason found	0.641%	60,098	High
	p1empl_1	27a - Person 1: Employer	KFI	Wrong character	No reason found	0.640%	78,439	High
	p3age	4 - Person 3: Age	OCR	Wrong character	Rules not followed	0.616%	56,206	High
	r1last	Roster: Person 1 Last Name	KFI	Wrong character	No reason found	0.600%	58,706	High
	p2first	1 - Person 2: First Name	OCR	Wrong character	Poor handwriting	0.598%	87,106	High
	p1ethn_1	10 - Person 1: Ancestry	KFI	Extra characters	No reason found	0.597%	50,779	High
	p1pharea	2 - Person 1: Phone Number Area Code	OCR	Wrong character	Poor handwriting	0.595%	142,451	High
	p1phpre	2 - Person 1: Phone Number Exchange	OCR	Wrong character	Poor handwriting	0.590%	141,675	High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d2	p2olayof	25a - Person 2: Last Week Layoff	OMR	Extra check-box	Stray mark or spot	0.587%	54,031	High
	p1city	22b - Person 1: Work City	OCR	Wrong character	Poor handwriting	0.580%	56,246	High
	p2addr_1	22a - Person 2: Work Address	KFI	Wrong character	No reason found	0.579%	56,468	High
	p1ospkwl	11c - Person 1: Speak English Well	OMR	Extra check-box	Stray mark or spot	0.578%	53,123	High
	p2ethn_1	10 - Person 2: Ancestry	OCR	Extra characters	Rules not followed	0.569%	60,795	High
	r1last	Roster: Person 1 Last Name	KFI	Extra characters	No reason found	0.566%	58,706	High
	p1last	3 - Person 1: Last Name	KFI	Wrong character	No reason found	0.564%	60,464	High
	p1ethn_1	10 - Person 1: Ancestry	OCR	Extra characters	Rules not followed	0.562%	88,317	High
	p1ethn_1	10 - Person 1: Ancestry	OCR	Missing characters	No reason found	0.551%	88,317	High
	p2ethn_1	10 - Person 2: Ancestry	OCR	Wrong character	Poor handwriting	0.551%	60,795	High
	p2ethn_1	10 - Person 2: Ancestry	OCR	Missing characters	No reason found	0.526%	60,795	High
	p1wages	31a - Person 1: Wages Amount	OCR	Wrong character	Poor handwriting	0.523%	66,692	High
	p2owages	31a - Person 2: Wages	OMR	Extra check-box	Box is crossed out	0.516%	77,289	High
	p2olayof	25a - Person 2: Last Week Layoff	OMR	Extra check-box	Big X through person	0.516%	54,031	High
	p1last	3 - Person 1: Last Name	KFI	Extra characters	No reason found	0.513%	60,464	High
	p1esttax	49 - Household: Real Estate Tax Amount	OCR	Wrong character	Poor handwriting	0.484%	63,651	High
	p1kind_1	28a - Person 1: Occupation Kind of Work	KFI	Wrong character	No reason found	0.483%	52,833	High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d2	p1actv_1	27b - Person 1: Industry	KFI	Wrong character	No reason found	0.482%	52,455	High
	p1flood	50 - Household: Insurance Payment	OCR	Wrong character	Poor handwriting	0.476%	59,705	High
	p1duty_1	28b - Person 1: Occupation Duties	KFI	Position reversed	No reason found	0.473%	60,098	High
	p1first	3 - Person 1: First Name	OCR	Wrong character	Poor handwriting	0.472%	125,718	High
	p1oagric	44c - Household: Agricultural Products	OMR	Extra check-box	Stray mark or spot	0.471%	51,605	High
	p1minute	24b - Person 1: Minutes to Work	OCR	Wrong character	Poor handwriting	0.470%	79,368	High
	p1water	45c - Household: Water and Sewer Cost	OCR	Wrong character	Poor handwriting	0.464%	74,853	High
	p1olstwk	26 - Person 1: Last Worked	OMR	Extra check-box	Stray mark or spot	0.460%	56,465	High
	p1ethn_1	10 - Person 1: Ancestry	KFI	Position reversed	No reason found	0.457%	50,779	High
	r2first	Roster: Person 2 First Name	OCR	Wrong character	Poor handwriting	0.457%	89,527	High
	p1odegre	9 - Person 1: Highest Degree Completed	OMR	Missing check-box	No reason found	0.454%	159,646	High
	p1odeed	47a - Household: Mortgage	OMR	Extra check-box	Box is crossed out	0.453%	110,786	High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d2e	rilast	R1 - Respondent's Last Name	OCR	Wrong character	Poor handwriting	15.575%	166,529	Very High
	rifirst	R1 - Respondent's First Name	OCR	Wrong character	Poor handwriting	10.579%	168,443	Very High
	rilast	R1 - Respondent's Last Name	OCR	Missing characters	No reason found	2.395%	166,529	Very High
	p1ovalue	51 - Household: Property Value	OMR	Extra check-box	Poor image	1.859%	67,225	Very High
	p1otrans	23a - Person 1: Work Vehicle	OMR	Extra check-box	Poor image	1.848%	59,801	Very High
	p1phext	2 - Person 1: Phone Number Digits	OCR	Wrong character	Poor handwriting	1.571%	129,893	Very High
	rifirst	R1 - Respondent's First Name	OCR	Missing characters	No reason found	1.568%	168,443	Very High
	s4ointro	S4 - Vacant or Occupied	OMR	Extra check-box	Stray mark or spot	1.345%	50,179	Very High
	p1pharea	2 - Person 1: Phone Number Area Code	OCR	Wrong character	Poor handwriting	1.251%	134,961	Very High
	p1phpre	2 - Person 1: Phone Number Exchange	OCR	Wrong character	Poor handwriting	1.163%	134,911	Very High
	p2last	1 - Person 2: Last Name	OCR	Wrong character	Poor handwriting	1.155%	52,203	Very High
	p1lasta	7 - Person 1: Last Name	OCR	Wrong character	Poor handwriting	1.133%	64,356	Very High
	p1odegre	9 - Person 1: Highest Degree Completed	OMR	Extra check-box	Poor image	1.124%	84,670	Very High
	p1last	3 - Person 1: Last Name	OCR	Wrong character	Poor handwriting	1.104%	71,488	Very High
	rilast	R1 - Respondent's Last Name	OCR	Wrong character	No reason found	0.957%	166,529	Very High
	p1odeed	47a - Household: Mortgage	OMR	Extra check-box	Poor image	0.798%	51,140	High
	p1oride	23b - Person 1: Carpool	OMR	Extra check-box	Poor image	0.683%	51,244	High
	p2first	1 - Person 2: First Name	OCR	Wrong character	Poor handwriting	0.644%	57,722	High
	s3ointro	S3 - Seasonal Home	OMR	Extra check-box	Stray mark or spot	0.634%	118,922	High

Form Name	Field Name	Description	Mode	Manner of Misinter-pretation	Reason for Misinter-pretation	Nonblank Misinter-pretation %	Total Nonblank Records	Outlier
d2e	p1dob_d	6 - Person 1: Day of Birth	OCR	Wrong character	Poor handwriting	0.609%	83,628	High
	p1first	3 - Person 1: First Name	OCR	Wrong character	Poor handwriting	0.584%	83,387	High
	rifirst	R1 - Respondent's First Name	OCR	Wrong character	No reason found	0.582%	168,443	High
	p2dob_d	4 - Person 2: Day of Birth	OCR	Wrong character	Poor handwriting	0.555%	62,003	High
	p2firsta	7 - Person 2: First Name	OCR	Wrong character	Poor handwriting	0.544%	53,524	High
	p1odeed	47a - Household: Mortgage	OMR	Extra check-box	Stray mark or spot	0.542%	51,140	High
	p1elec	45a - Household: Electricity Cost	OCR	Missing characters	No reason found	0.523%	53,303	High
	p1ooffce	44a - Household: Business	OMR	Extra check-box	Stray mark or spot	0.522%	111,898	High
	rifirst	R1 - Respondent's First Name	OCR	Wrong character	Poor image	0.515%	168,443	High
	p1ethn_1	10 - Person 1: Ancestry	OCR	Wrong character	Poor handwriting	0.509%	55,244	High
	p1hours	30c - Person 1: Hours Worked per Week	OCR	Wrong character	Poor handwriting	0.472%	51,265	High
	p1firsta	7 - Person 1: First Name	OCR	Wrong character	Poor handwriting	0.465%	76,352	High

For Table 50, we show only the field category rates that are high or very high outliers. The total number of field category error rates, 713, is too large to be readable. Instead we place the entire list in Appendix I for easier reference.

Table 50. Field Category Misinterpretation Rates that are High or Very High Outliers, Broken Out by Reason For Misinterpretation

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
Coverage	Extra check-box	Poor image	0.088%	High
Form Management	Added response	Poor handwriting	0.120%	High
		Stray mark or spot	0.211%	Very High
		No reason found	0.131%	High
		Poor image	0.093%	High
		No reason found	0.289%	Very High
		Poor handwriting	6.127%	Very High
		Rules not followed	0.647%	Very High
		No reason found	0.287%	Very High
Housing Profile	Added response	Rules not followed	0.151%	High
		Poor image	0.170%	Very High
		Stray mark or spot	0.163%	Very High
		Box is crossed out	0.138%	High
		No reason found	0.239%	Very High
		Poor image	0.091%	High
		Poor handwriting	0.637%	Very High
		Spanish accents	0.196%	Very High
		Mixed upper case & lower case	0.110%	High
		Rules not followed	0.092%	High
POP--Demographic	Added response	Spanish accents	0.923%	Very High
		Spanish accents	1.010%	Very High
		Poor image	0.171%	Very High
		Box is crossed out	0.093%	High
		Stray mark or spot	0.086%	High
		No reason found	0.194%	Very High
		Rules not followed	0.193%	Very High
		Poor handwriting	0.550%	Very High
		Spanish accents	0.265%	Very High
		Box is crossed out	0.149%	High
		Poor image	0.147%	High
		Stray mark or spot	0.145%	High
POP--Education	Extra check-box	Poor image	0.450%	Very High
		Box is crossed out	0.303%	Very High
		Stray mark or spot	0.191%	Very High
		No reason found	0.110%	High
POP--Ethnic	Added response	Response crossed out	0.395%	Very High
		Spanish accents	0.106%	High
		Poor handwriting	0.093%	High
		Rules not followed	0.281%	Very High
		No reason found	0.253%	Very High
		No reason found	1.422%	Very High
		Truncated	0.144%	High
		Character goes out field	0.085%	High
		Spanish accents	0.654%	Very High
		No reason found	0.181%	Very High
		Poor handwriting	1.157%	Very High
		No reason found	0.198%	Very High
		Spanish accents	0.154%	High

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Income	Added response	Rules not followed	0.858%	Very High
		Response crossed out	0.147%	High
		Poor handwriting	0.085%	High
		No reason found	0.156%	High
		Box is crossed out	0.195%	Very High
		Stray mark or spot	0.146%	High
		Poor image	0.144%	High
		No reason found	0.360%	Very High
		Response written over	0.121%	High
		Poor handwriting	0.753%	Very High
		Rules not followed	0.318%	Very High
		Response written over	0.167%	Very High
		No reason found	0.098%	High
POP--Military	Extra check-box	Poor image	0.889%	Very High
		Stray mark or spot	0.223%	Very High
		Big X through person	0.145%	High
		Box is crossed out	0.138%	High
		No reason found	0.224%	Very High
POP--Name	Extra characters	No reason found	0.137%	High
		No reason found	0.340%	Very High
		Truncated	0.102%	High
		Poor handwriting	1.848%	Very High
		No reason found	0.228%	Very High
		Mixed upper case & lower case	0.124%	High
POP--Occupation	Extra characters	No reason found	0.328%	Very High
		Rules not followed	0.100%	High
		Poor image	0.385%	Very High
		Box is crossed out	0.364%	Very High
		Stray mark or spot	0.329%	Very High
		Big X through person	0.194%	Very High
		Rules not followed	2.096%	Very High
		No reason found	0.935%	Very High
		Character goes out field	0.166%	Very High
		Truncated	0.128%	High
		Poor handwriting	0.095%	High
		No reason found	0.170%	Very High
		Poor handwriting	1.303%	Very High
		No reason found	0.188%	Very High
POP--Race	Added response	Response crossed out	1.961%	Very High
		Poor handwriting	0.976%	Very High
		Big X through person	0.228%	Very High
		Rules not followed	0.183%	Very High
POP--Race	Blanked response	No reason found	0.184%	Very High

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Race	Extra characters	Response crossed out	0.404%	Very High
		Rules not followed	0.339%	Very High
		No reason found	0.314%	Very High
		Poor handwriting	0.166%	Very High
POP--Race	Extra check-box	Big X through person	0.086%	High
POP--Race	Missing characters	No reason found	1.602%	Very High
		Truncated	0.891%	Very High
		Poor handwriting	0.269%	Very High
		Character goes out field	0.228%	Very High
		Characters too close	0.222%	Very High
POP--Race	Position reversed	No reason found	0.247%	Very High
		Poor image	0.141%	High
POP--Race	Wrong character	Poor handwriting	3.047%	Very High
		No reason found	0.537%	Very High
		Spanish accents	0.252%	Very High
		Mixed upper case & lower case	0.207%	Very High
		Characters too close	0.161%	Very High
		Response written over	0.129%	High
		Truncated	0.105%	High
		Rules not followed	0.091%	High
Special Housing	Added response	Poor handwriting	0.231%	Very High
		Character goes out field	0.098%	High
Special Housing	Blanked response	No reason found	0.916%	Very High
Special Housing	Missing characters	No reason found	0.104%	High
		Rules not followed	0.101%	High
Special Housing	Wrong character	Poor handwriting	0.135%	High

The three themes of Table 49 are Poor handwriting (82 out of 195 outliers in the table), No reason found (56 out of 195 outliers in the table), and Rules not followed (23 out of 195 outliers in the table). These reasons cut across the most forms and fields. At the field category level in Table 50, the picture is the same. Of the 117 outliers in Table 50, the reasons poor handwriting, no reason found, and rules not followed account for 58–almost one-half of the cases.

We see two options in light of these findings. One is to review the rules used by the analysts to judge the intent of the respondent. Were these rules too strict for adequately capturing intent? Did the analysts and supervisors apply them too conservatively? In either case, it is possible the error results make the picture worse than it really is.

If we do not review the rules, or if we think their application was reasonable, then we have to rely on the data as is. When Poor handwriting or No reason found are a plurality of the reasons for the most frequent errors, we cannot count on high technology alone for major improvements. The most obvious solution, reducing some write-in fields to check-boxes or using enumerators more often to get long form data, raise prospects of higher cost or more limited information.

Our course of action is highly dependent on strategic decisions about the decennial census. If one-sixth of the nation's households continue to supply long form data, the resulting sea of handwritten responses will continue to limit our ability to capture intent via automated technology. If the long form data collection is dropped, or if a more check-box oriented, reduced set of questions can be substituted for the present one, then it will be much easier to use the automated technology to better capture respondent intent.

5. POSSIBLE QUESTIONS FOR FUTURE RESEARCH

5.1 Questions vs. Recommendations

While the usual title for this section is Recommendations, we believe our choice describes its content more precisely. At the start of our extensive examination of the quality of automated data capture, we hoped to produce recommendations in such areas as system hardware, software logic, form processing, and form design.

Our ideal recommendations would be of sufficient detail to suggest directly how they could be implemented, how much they would cost, and what the broad economic and technical benefits would be. Reluctantly, we end our examination short of this ideal. Despite our in depth understanding of how data capture errors behave, we cannot in any concrete way provide detailed guidance on how to make the data capture algorithm more intelligent or how to design decennial census forms that better leverage the capabilities of the automated technology.

We started this evaluation with a set of questions. The best way to end it is with a different set of questions. Throughout the evaluation, we have commented about patterns and trends that struck us as worth a more extensive look. Now these comments are brought together here.

At several points in this section, we refer to "fields filled out for multiple persons on a form." These are fields like name, age, and sex which appear more than once on a decennial census form. They are repeated so information can be recorded for every member of a household. For all other fields, we use the phrase "fields filled out for only one person on a form." For definitions of common or special terms in this section, see the glossary in Appendix M.

We close by framing our comments as questions. Perhaps if these questions are pondered by the specialists who design the relevant software, hardware, or census forms, the marriage of their reflection and knowledge may help bring about the next advance in how the Census Bureau uses automated data capture and imaging technology. Within the limits of our specialty, quality assurance, we hope what we have said so far contributes to vigorous and fruitful investigation.

5.2 Should the Census Bureau expand existing efforts to make certain groups of fields easier for respondents to understand and fill out?

From section 4.2, we see evidence the enumerator-returned forms had significantly lower soft and hard match error rates compared to the respondent-returned forms for these categories of fields:

- POP–Ethnic,
- POP–Name, and
- POP–Race.

The difference in the median nonblank error rate for POP–Ethnic is 2.6 percent. The difference for POP–Name is 1.3 percent. The difference for POP–Race is 2.4 percent. Is this gap large enough to justify more efforts to improve the layout and readability of these field categories for respondent-returned forms?

The Census 2000 Questionnaire Design Study suggests some specific ways to enhance readability in the context of possible improvements for the short form. While the discussion there does not distinguish respondent-returned vs. enumerator-returned forms, the recommendations can clearly apply to either.

- consider having the person information for household members be filled out from left to right across the page instead of up and down,
- consider allowing the use of pencil so respondents can correct mistakes more easily,
- change the sizes, fonts, appearance, and so forth of the instruction icons so they are easier to spot,
- allow more spaces for the last name field,
- include instructions for filling out or correcting write-in fields,
- include more detailed instructions for the race and ethnicity questions,
- try to make the instructions to the head of household for filling out the form more concise,
- consider including headers to separate the Asian ethnicity options from the ones for Pacific Islander,
- do not spread the choices for check-box fields over more than one row or column on a page, and
- choose a background color with better visual contrast to the human eye.

5.3 Do the outlier rates for the d2ur or the POP–Name outliers on the d1e, d1s, d2e, and d2ur suggest challenges to the automated technology that are great enough to require increased attention?

The forms mentioned in the question are

- d1e, the English enumerator short form,
- d1s, the Spanish mailout/mailback short form,
- d2e, the English enumerator long form, and
- d2ur, the English update/leave form for Puerto Rico.

From section 4.3, we see evidence the d2ur, poses a challenge to the automated technology in terms of hard or soft match errors, at least for name and ethnicity fields on long forms. When the error rates are calculated at the field category level, d2ur has more categories that are high or very high outliers than any other form. The outlier error rates range from 2.7 percent to 7.9

percent.

POP–Name is the field category that is a high or very high outlier on the largest number of forms. It is a high or very high outlier on the d1e, d1s, d2e, and d2ur forms. The error rates for POP–Name over these four forms range from 4.2 percent to 7.1 percent.

Are the outlier rates for the d2ur or the POP–Name outliers for the four forms listed above high enough to require increased efforts to improve them?

5.4 Is the disproportionately higher number of outlier error rates on the d2 an issue?

From section 4.4 and Appendix H, we see evidence the d2, the English mailout/mailback long form, has a disproportionately greater number of high or very high outliers for hard and soft match error rates when compared with the forms

- d1, the English mailout/mailback short form,
- d1e, the English enumerator short form,
- d1s, the Spanish mailout/mailback short form,
- d1u, the English update/leave form,
- d2e, the English enumerator long form,
- d2u, the English update/leave long form, and
- d2ur, the English update/leave form for Puerto Rico.

Based on the number of fields on the d2, we expect 44 high or very high outliers. The actual number is 69. The difference, 25, is statistically significant at the 10 percent level. Is the difference large enough to support increased efforts to redesign or simplify the d2?

5.5 Does the difference in significant factors for nonperson and person fields when the raw data are broken out by data capture mode require explanation?

From section 4.5, we see that when the raw data are broken out by data capture mode, the factors significantly affecting the nonblank hard or soft match error rate are not constant over field type. For fields filled out for only one person on a form, the only significant factor is form.

When fields that are filled out for multiple persons are considered, the significant factors are form, field category, mode, and the interaction of field with mode. Interaction means that the effect of field will change depending on the mode. The field and mode do not operate independently in their affect on the nonblank error rate. The last factor is the most significant.

Is this difference in significant factors for nonperson and person fields something important enough to be explained? Does this difference offer any clues about how to improve the performance of the automated technology?

5.6 Is the appearance of the categories Form Management and POP–Name as the top two error rates in all four data capture centers something that requires explanation?

The field categories Form Management and POP–Name have the highest nonblank error rates in all four data capture centers. Form Management covers the person added and person canceled fields on the enumerator forms. It is encouraging to note that only one of the 52 outlier rates shown for Form Management was for adding or canceling persons.

An interesting follow up question is "What specifically is there about the nature of the Form Management and POP–Name categories that leads them to occupy the top two positions in all four data capture centers?"

5.7 Is the appearance of the POP–Name category as an outlier in Census 2000 RCCs containing areas of traditional immigrant concentration something that requires more detailed investigation?

The immigrant populations concentrated in regional census centers 22, 23, 27, 29, and 32 could account at least partly for high error rates in POP–Name fields.

From section 4.7, we see evidence that when the error rates are calculated at the field category level, the category POP–Name appears as a high outlier for soft match errors in Census 2000 regional census centers 22, 23, 27, 29, and 32. The error rates range from 3.9 percent to 4.4 percent. RCCs 22, 23, 27, 29, and 32 cover Chicago, Los Angeles, New York City, and Texas. These areas have concentrations of immigrant populations where problems with name fields are not a surprise. Are name field outliers in these RCCs high enough to merit more detailed investigation?

5.8 Is the difference in the largest significant factor for nonperson and person fields when the raw data are broken out by KFI impact an issue that should be explained?

From section 4.8, we see evidence that when the raw data are broken out by KFI impact, the factors significantly affecting the nonblank hard or soft match error rate are not constant over field type. KFI impact refers to how well we capture the respondent's intent after

- content is rejected by the automated technology during Census 2000 processing, and
- the rejected content is sent to a human operator for Key From Image.

When we look at fields that are filled out for only one person on a form, those with a data capture mode of KFI have their nonblank soft match error rate significantly affected by form and field category. Of the two, form is the larger contributor.

When fields that are filled out for multiple persons are considered, there are four significant factors: form, KFI Impact, the interaction of field with KFI impact, and field category. The largest contributor is the interaction of field with KFI impact. Interaction means that the effect of field will change depending on the impact of KFI. Field and KFI impact do not operate independently in their affect on the nonblank soft match error rate.

For fields filled out for only one person on a form, the largest significant factor is form. The largest significant factor for person fields is the interaction of field by KFI impact. Is this difference something important enough to be explained? Does explaining this difference offer any clues about how to improve the performance of the automated technology?

There are some clues to partly explain the interaction of field and KFI impact on the nonblank soft match error rate. First, the most frequent category of KFI impact is "Cannot be determined". The automated technology rejected the content, and the entry keyed by the human operator was ultimately not judged to reflect the intent of the respondent, character for character. These are examples of content that tend to be especially difficult to interpret.

Second, there are name field nonblank error rates on the d1s form that are outliers. The d1s is the Spanish mailout/mailback short form. It is possible these outliers reflect limits on the capability of the automated technology to understand special Spanish language characters.

Third, many of the outliers on the d2, the English mailout/mailback long form and d2u, the English update/leave long form, are for fields in which respondents write in a race or ethnicity other than the ones provided. This might reflect the increased challenge of interpreting characters written by hand instead of checked off in a box, especially when the handwriting is poor.

5.9 Is the concentration of redundant KFI cases in the POP–Name category something that requires explanation?

From section 4.9, we see the field category POP–Name is the only one flagged a high or very high outlier. Specific fields in the POP–Name category appear as high or very high outliers for forms d1s and d2u, specifically the middle initial for higher numbered persons. The d1s is the Spanish mailout/mailback short . The d2u is the English update/leave long form.

While we do not propose it as the only explanation, respondent fatigue is a possible one for the POP–Name outliers. By the time respondents supply name information for the fifth or sixth person in a household, it is reasonable to suppose accuracy or neatness in the middle initial is not a high priority. Ideally, no field should be sent to KFI redundantly. In the case of a field consisting of single character, however, it is not clear to us the benefits of achieving the ideal is worth the cost.

5.10 Should certain fields sent automatically to KFI be allowed to go through the automated technology for processing?

From section 4.5.2, we note some fields automatically went to KFI regardless of how well the technology thought it could process them. These were check-box fields where more than one box could be selected and still count as a valid response. Recognizing that KFI is subject to error from factors not affecting the technology, e.g. human fatigue and inattention, a possible future test for the automated technology is to allow it to process multiple response check-box fields. It would be helpful to find out if the technology can be adjusted to accept such fields without the errors of keying.

5.11 If the present long form data collection process is retained for the 2010 census is it worthwhile to improve the quality performance of the automated technology?

According to section 4.11, the three most commonly assigned reasons for failure to capture respondent intent were

- Poor handwriting (82 out of 195 outliers shown in Table 49),
- No reason found (56 out of 195 outliers shown in Table 49), and
- Rules not followed (23 out of 195 outliers shown in Table 49).

If we assume the analysts and supervisor properly applied the rules for determining respondent intent, then we have to rely on the data as we have them. When Poor handwriting or No reason found are a plurality of the reasons for the most frequent errors, we cannot count on high technology by itself for significant improvement.

The most obvious solution, reducing more write-in fields to check-boxes or using enumerators more frequently to gather long form data, raise prospects of higher cost or more limited information. If one-sixth of the nation's households continue to supply long form data, the resulting sea of handwritten responses will limit our ability to capture intent via automated technology.

If the long form data collection is dropped, or if a more check-box oriented, streamlined set of questions can be substituted for the present one, then it will be much easier to use the automated technology to better capture intent. Is it better to accept the present performance of the automated technology and invest more effort to simplify or redesign the decennial census forms?

References

[1] Graybill, F. A., 1961, *An Introduction to Linear Statistical Models*, McGraw-Hill.

[2] Hopkins, Will G., 2002, *A New View of Statistics, Square-root and Arcsine-root Transformation*, http://www.sportsci.org/resource/stats/counts.html#squarc..

[3] Neugebauer, Randall J., <randall.j.neugebauer@census.gov>, "KFI write-up", May 9, 2002, office communication (May 9, 2002).

[4] Reichert, Jennifer W., <jenniferl.w.reichert@census.gov>, "Points to confirm for evaluation K.1.B", December 18, 2002, office communication (December 18, 2002).

[5] SAS Institute, Inc., 1990, *SAS/STAT User's Guide, Version 6.0*, 4th ed., Cary, NC.

[6] Tukey, J.W., 1977, *Exploratory Data Analysis*, Addison-Wesley.

[7] University of New Brunswick, Canada, *Confidence Intervals for the Median, Two sided Symmetric, 95% or Better*, http://www.math.unb.ca/~knight/utility/MedInt95.htm.

[8] U.S. Bureau of the Census, 1999, *Data Capture System Quality Evaluation, Project Description and Procedures*, Decennial Statistical Studies Division, Quality Assurance branch internal document, May, 1999.

[9] U.S. Bureau of the Census, 1999, *H3: Quality of the Data Capture System*, Census 2000 Dress Rehearsal Evaluation Results Memorandum Series # H3, July, 1999.

[10] U.S. Bureau of the Census, 2000, *Study Plan for Evaluation of the Quality of the Data Capture System and the Impact of the Data Capture Mode on the Data Quality, K.1.B*, DSSD Census 2000 Procedures and Operations Memorandum Series #JJ-13, December 21, 2000.

[11] U.S. Bureau of the Census, 2002, *Census 2000 Questionnaire Design Study*, Planning Research and Evaluation Division, December 20, 2002.

[12] Wallis, W.A. and Roberts, H.W., 1957, *Statistics: A New Approach*, The Free Press, Glencoe, IL.

Appendix A: List of Census 2000 Forms

In this appendix we list the Census 2000 form names included in the raw data for this evaluation. We also give the abbreviations of these form names as they appear in the tables of the body of the
evaluation.

Table A1. List of Form Name

Form Name	Abbreviation
Short Form, Mailout/Mailback	d1
Short Form, Enumerator	d1e
Short Form, Enumerator, Puerto Rico	d1er
Short Form, Mailout/Mailback, Spanish	d1s
Short Form, Update/Leave	d1u
Short Form, Update/Leave, Puerto Rico	d1ur
Long Form, Mailout/Mailback	d2
Long Form, Enumerator	d2e
Long Form, Enumerator, Puerto	d2er
Long Form, Mailout/Mailback, Spanish	d2s
Long Form, Update/Leave	d2u
Long Form, Update/Leave, Puerto Rico	d2ur

Appendix B: List of Census 2000 Field Categories

In this appendix, we list the categories of fields that were used to analyze and summarize the data in this evaluation. We also give a short description of each category.

Table B1. List of Field Categories

Field Category	Description
Coverage	Household coverage questions on enumerator form
Form Management	Contact data, persons added or canceled on enumerator form
POP–Demographic	Age, marital status, ancestry, and similar demographic data
POP–Disability	Existence and extent of personal disability of household members
POP–Education	Educational attainment of household members
POP–Ethnic	Ethnic data of household members, including Hispanic origin
POP–Income	Income characteristics of household members
POP–Military	Military service characteristics of household members
POP–Name	First, middle, and last names of household members
POP–Occupation	Occupational characteristics of household members
POP–Race	Racial data of household members
Residential Profile	Features, expenses, age and similar data of residential structure
Special Housing	Special Place, Usual Home Elsewhere, and related designations

Appendix C: List of Census 2000 Field Names

In this appendix, we list the 810 field names with categories and descriptions.

Table C1. List of Field Names With Categories and Descriptions

Field Name	Description	Category
1 a_status	Summary - A: Status	Residential Profile
2 b_pop	Summary - B: Pop	POP--Demographic
3 c_osumma	Summary - C: Vacant	Residential Profile
4 c1ocover	C1 - Coverage	Coverage
5 c2ocover	C2 - Coverage	Coverage
6 d_sp	Summary - D: SP	Special Housing
7 e_oconti	Continuation Forms	Form Management
8 e_sheets	Number of Continuation Forms	Form Management
9 e_uhe	Summary - E: UHE	Special Housing
10 f_mov	Summary - F: MOV	Special Housing
11 g_pi	Summary - G: PI	Special Housing
12 h_ref	Summary - H: REF	Special Housing
13 i_rep	Summary - I: REP	Special Housing
14 j_co	Summary - J: CO	Special Housing
15 jic1	Summary - L: JIC1	Special Housing
16 jic2	Summary - M: JIC2	Special Housing
17 jic3	Summary - N: JIC3	Special Housing
18 jic4	Summary - O: JIC4	Special Housing
19 k_tc	Summary - K: TC	Special Housing
20 p1_oil	45d - Household: Oil Cost	Residential Profile
21 p1_other	31h - Person 1: Other Income Amount	POP--Income
22 p10first	Person 10: First Name	POP--Name
23 p10last	Person 10: Last Name	POP--Name
24 p10mi	Person 10: Middle Initial	POP--Name
25 p11first	Person 11: First Name	POP--Name
26 p11last	Person 11: Last Name	POP--Name
27 p11mi	Person 11: Middle Initial	POP--Name
28 p12first	Person 12: First Name	POP--Name
29 p12last	Person 12: Last Name	POP--Name
30 p12mi	Person 12: Middle Initial	POP--Name
31 p1actv_1	27b - Person 1: Industry	POP--Occupation
32 p1addr_1	22a - Person 1: Work Address	POP--Occupation
33 p1age	6 - Person 1: Age	POP--Demographic
34 p1apt16a	H2 - Apartment Number	Residential Profile
35 p1asia_1	6 - Person 1: Other Asian	POP--Ethnic
36 p1asia19	8 - Person 1: Other Asian	POP--Ethnic
37 p1auto	44 - Household: Number of Automobiles	Residential Profile
38 p1bnoth	12 - Person 1: Name of Country	POP--Demographic
39 p1bnus	12 - Person 1: Name of State	POP--Demographic
40 p1city	22b - Person 1: Work City	POP--Occupation
41 p1cntry	15a - Person 1: Migration Country	POP--Demographic
42 p1condo	52 - Household: Condo Fee	Residential Profile
43 p1cost	53b - Household: Mobile Home Payment	Residential Profile
44 p1county	22d - Person 1: Work County	POP--Occupation
45 p1cty16a	H2 - City	Residential Profile
46 p1dob_d	6 - Person 1: Day of Birth	POP--Demographic

Field Name	Description	Category
47 p1dob_m	6 - Person 1: Month of Birth	POP--Demographic
48 p1dob_y	6 - Person 1: Year of Birth	POP--Demographic
49 p1duty_1	28b - Person 1: Occupation Duties	POP--Occupation
50 p1elec	45a - Household: Electricity Cost	Residential Profile
51 p1empl_1	27a - Person 1: Employer	POP--Occupation
52 p1esttax	49 - Household: Real Estate Tax Amount	Residential Profile
53 p1ethn_1	10 - Person 1: Ancestry	POP--Ethnic
54 p1first	3 - Person 1: First Name	POP--Name
55 p1firsta	7 - Person 1: First Name	POP--Name
56 p1flood	50 - Household: Insurance Payment	Residential Profile
57 p1gas	45b - Household: Gas Cost	Residential Profile
58 p1hisp_1	5 - Person 1: Other Hispanic Origin	POP--Ethnic
59 p1hisp19	7 - Person 1: Other Hispanic Origin	POP--Ethnic
60 p1hours	30c - Person 1: Hours Worked per Week	POP--Occupation
61 p1hsn10a	H2 - House Number	Residential Profile
62 p1int	31c - Person 1: Interest Amount	POP--Income
63 p1kind_1	28a - Person 1: Occupation Kind of Work	POP--Occupation
64 p1lang	11b - Person 1: Language	POP--Demographic
65 p1last	3 - Person 1: Last Name	POP--Name
66 p1lasta	7 - Person 1: Last Name	POP--Name
67 p1lvcity	15b - Person 1: Migration City	POP--Demographic
68 p1lvcnty	15b - Person 1: Migration County	POP--Demographic
69 p1lvstat	15b - Person 1: Migration State	POP--Demographic
70 p1lvzip	15b - Person 1: Migration Zip Code	POP--Demographic
71 p1mi	3 - Person 1: Middle Initial	POP--Name
72 p1mia	7 - Person 1: Middle Initial	POP--Name
73 p1minute	24b - Person 1: Minutes to Work	POP--Occupation
74 p1mort	47b - Household: Mortgage Amount	Residential Profile
75 p1o15age	19 - Person 1: Under 15 Interviewer Instruction	Form Management
76 p1o2mort	48a - Household: Second Mortgage	Residential Profile
77 p1o5ago	15a - Person 1: Live Here 5 Years Ago	POP--Demographic
78 p1oabsnt	25b - Person 1: Last Week Absent	POP--Occupation
79 p1oacres	44b - Household: Acreage	Residential Profile
80 p1oadd	1 - Person 1: Add	Form Management
81 p1oagric	44c - Household: Agricultural Products	Residential Profile
82 p1oalone	17c - Person 1: Difficulty Shopping	POP--Disability
83 p1oam_pm	24a - Person 1: Time to Work am/pm	POP--Occupation
84 p1oarmed	27a - Person 1: Armed Forces	POP--Military
85 p1oauto	43 - Household: Number of Automobiles	Residential Profile
86 p1obdrm	38 - Household: Number of Bedrooms	Residential Profile
87 p1obldg	34 - Household: Building Type	Residential Profile
88 p1oblind	16a - Person 1: Blind or Deaf	POP--Disability
89 p1oborn	18 - Person 1: Under 15	POP--Demographic
90 p1obuilt	35 - Household: Building Age	Residential Profile
91 p1ocancl	1 - Person 1: Cancel	Form Management
92 p1ocondo	57a - Household: Condo	Residential Profile
93 p1octlmt	22c - Person 1: Work Inside City Limits	POP--Occupation
94 p1octzn	13 - Person 1: Citizen	POP--Demographic
95 p1odeed	47a - Household: Mortgage	Residential Profile
96 p1odegre	9 - Person 1: Highest Degree Completed	POP--Education
97 p1odress	17b - Person 1: Difficulty Dressing	POP--Disability
98 p1oelec	45a - Household: Electricity	Residential Profile
99 p1oesttx	49 - Household: No Real Estate Taxes	Residential Profile

Field Name	Description	Category
100 p1oetype	29 - Person 1: Class of Worker	POP--Occupation
101 p1oflood	50 - Household: No Insurance	Residential Profile
102 p1ofuel	42 - Household: Fuel for Heating	Residential Profile
103 p1ogas	45b - Household: Gas	Residential Profile
104 p1ograde	8b - Person 1: Grade Level	POP--Education
105 p1ogrand	19a - Person 1: Grandchildren	POP--Demographic
106 p1ohisp	7 - Person 1: Hispanic Origin	POP--Ethnic
107 p1ohouse	33 - Household: Ownership Status	Residential Profile
108 p1oins	47d - Household: Insurance	Residential Profile
109 p1oint	31c - Person 1: Interest	POP--Income
110 p1ointls	31c - Person 1: Interest Loss	POP--Income
111 p1ojob	17d - Person 1: Difficulty Working	POP--Disability
112 p1oktchn	40 - Household: Kitchen	Residential Profile
113 p1olayof	25a - Person 1: Last Week Layoff	POP--Occupation
114 p1olimit	16b - Person 1: Limits Physical Activities	POP--Disability
115 p1oloan	53a - Household: Mobile Home Loan	Residential Profile
116 p1olook	25d - Person 1: Looking for Work	POP--Occupation
117 p1olstwk	26 - Person 1: Last Worked	POP--Occupation
118 p1olvcty	15b - Person 1: Live Inside City Limits	POP--Demographic
119 p1omarry	7 - Person 1: Marital Status	POP--Demographic
120 p1omentl	17a - Person 1: Difficulty Learning	POP--Disability
121 p1omilit	20a - Person 1: Active Duty	POP--Military
122 p1omort	47b - Household: No Payment	Residential Profile
123 p1omoven	36 - Household: Year Moved In	Residential Profile
124 p1oneeds	19b - Person 1: Responsible for Needs	POP--Disability
125 p1ooffce	44a - Household: Business	Residential Profile
126 p1ooil	45d - Household: Oil	Residential Profile
127 p1oother	31h - Person 1: Other Income	POP--Income
128 p1ophone	41 - Household: Telephone	Residential Profile
129 p1oplumb	39 - Household: Plumbing	Residential Profile
130 p1oproft	21 - Person 1: Work Last Week	POP--Occupation
131 p1orace	8 - Person 1: Race	POP--Race
132 p1orecal	25c - Person 1: Will Be Recalled	POP--Occupation
133 p1orent	46b - Household: Meals with Rent	Residential Profile
134 p1oresp	19c - Person 1: How Long	Residential Profile
135 p1oretax	47c - Household: Real Estate Taxes	Residential Profile
136 p1oretir	31g - Person 1: Retirement Income	POP--Income
137 p1oride	23b - Person 1: Carpool	POP--Occupation
138 p1orooms	37 - Household: Number of Rooms	Residential Profile
139 p1oscool	8a - Person 1: Attend School	POP--Education
140 p1osecpy	48b - Household: No Payment	Residential Profile
141 p1oselfe	31b - Person 1: Self- Person 1:employment Income	POP--Income
142 p1oserve	20b - Person 1: When on Active Duty	POP--Military
143 p1osex	5 - Person 1: Sex	POP--Demographic
144 p1oslfls	31b - Person 1: Self- Person 1:employment Loss	POP--Income
145 p1osocl	31d - Person 1: Social Security, Railroad Retirement	POP--Income
146 p1ospeak	11a - Person 1: Home Language	POP--Demographic
147 p1ospkwl	11c - Person 1: Speak English Well	POP--Demographic
148 p1ossi	31e - Person 1: SSI	POP--Income
149 p1ostart	25e - Person 1: Could Start Last Week	POP--Occupation
150 p1ototal	32 - Person 1: Total Income None	POP--Income
151 p1ototls	32 - Person 1: Total Income Loss	POP--Income
152 p1otrans	23a - Person 1: Work Vehicle	POP--Occupation

Field Name	Description	Category
153 p1otype	27c - Person 1: Business Type	POP--Occupation
154 p1ovalue	51 - Household: Property Value	Residential Profile
155 p1owages	31a - Person 1: Wages	POP--Income
156 p1owater	45c - Household: Water and Sewer	Residential Profile
157 p1owelfr	31f - Person 1: Welfare	POP--Income
158 p1owhrbn	12 - Person 1: Place of Birth	POP--Demographic
159 p1owork	30a - Person 1: Work Last Year	POP--Occupation
160 p1oyears	20c - Person 1: Years on Active Duty	POP--Military
161 p1pharea	2 - Person 1: Phone Number Area Code	POP--Demographic
162 p1phext	2 - Person 1: Phone Number Digits	POP--Demographic
163 p1phpre	2 - Person 1: Phone Number Exchange	POP--Demographic
164 p1race_1	6 - Person 1: Other Race	POP--Race
165 p1race19	8 - Person 1: Other Race	POP--Race
166 p1rent	46a - Household: Monthly Rent Amount	Residential Profile
167 p1retir	31g - Person 1: Retirement Income Amount	POP--Income
168 p1rooms	37 - Household: Number of Rooms	Residential Profile
169 p1secpay	48b - Household: Second Mortgage Amount	Residential Profile
170 p1selfe	31b - Person 1: Self Employment Income Amount	POP--Income
171 p1socl	31d - Person 1: Social Security, Railroad Retirement Amount	POP--Income
172 p1ssi	31e - Person 1: SSI Amount	POP--Income
173 p1stab2a	H2 - State	POP--Demographic
174 p1state	22e - Person 1: Work State	POP--Occupation
175 p1stx16a	H2 - Street Name	POP--Demographic
176 p1time	24a - Person 1: Time Leave for Work	POP--Occupation
177 p1total	32 - Person 1: Total Income Amount	POP--Income
178 p1trib_1	6 - Person 1: Am Indian, Alaska Native Tribe	POP--Race
179 p1trib19	8 - Person 1: Am. Indian, AK Native Tribe	POP--Race
180 p1wages	31a - Person 1: Wages Amount	POP--Income
181 p1water	45c - Household: Water and Sewer Cost	Residential Profile
182 p1weeks	30b - Person 1: Weeks Worked	POP--Occupation
183 p1welfr	31f - Person 1: Welfare Amount	POP--Income
184 p1yrmvus	14 - Person 1: Migration Year	POP--Demographic
185 p1zip	22f - Person 1: Work Zip Code	POP--Occupation
186 p1zip5a	H1- Zip Code	POP--Demographic
187 p2_other	31h - Person 2: Other Income Amount	POP--Income
188 p2_relo	2 - Person 2: Other Relative	POP--Demographic
189 p2actv_1	27b - Person 2: Industry	POP--Occupation
190 p2addr_1	22a - Person 2: Work Address	POP--Occupation
191 p2age	4 - Person 2: Age	POP--Demographic
192 p2asia_1	6 - Person 2: Other Asian	POP--Ethnic
193 p2asia19	6 - Person 2: Other Asian	POP--Ethnic
194 p2bnoth	12 - Person 2: Name of Country	POP--Demographic
195 p2bnus	12 - Person 2: Name of State	POP--Demographic
196 p2city	22b - Person 2: Work City	POP--Occupation
197 p2cntry	15a - Person 2: Migration Country	POP--Demographic
198 p2county	22d - Person 2: Work County	POP--Occupation
199 p2dob_d	4 - Person 2: Day of Birth	POP--Demographic
200 p2dob_m	4 - Person 2: Month of Birth	POP--Demographic
201 p2dob_y	4 - Person 2: Year of Birth	POP--Demographic
202 p2duty_1	28b - Person 2: Occupation Duties	POP--Occupation
203 p2empl_1	27a - Person 2: Employer	POP--Occupation
204 p2ethn_1	10 - Person 2: Ancestry	POP--Ethnic
205 p2first	1 - Person 2: First Name	POP--Name

Field Name	Description	Category
206 p2firsta	7 - Person 2: First Name	POP--Name
207 p2hisp_1	5 - Person 2: Other Hispanic Origin	POP--Ethnic
208 p2hisp19	5 - Person 2: Other Hispanic Origin	POP--Ethnic
209 p2hours	30c - Person 2: Hours Worked per Week	POP--Occupation
210 p2int	31c - Person 2: Interest Amount	POP--Income
211 p2kind_1	28a - Person 2: Occupation Kind of Work	POP--Occupation
212 p2lang	11b - Person 2: Language	POP--Demographic
213 p2last	1 - Person 2: Last Name	POP--Name
214 p2lasta	7 - Person 1: Last Name	POP--Name
215 p2lvcity	15b - Person 2: Migration City	POP--Demographic
216 p2lvcnty	15b - Person 2: Migration County	POP--Demographic
217 p2lvstat	15b - Person 2: Migration State	POP--Demographic
218 p2lvzip	15b - Person 2: Migration Zip Code	POP--Demographic
219 p2mi	1 - Person 2: Middle Initial	POP--Name
220 p2mia	7 - Person 1: Middle Initial	POP--Name
221 p2minute	24b - Person 2: Minutes to Work	POP--Occupation
222 p2o15age	19 - Person 2: Under 15 Interviewer Instruction	Form Management
223 p2o5ago	15a - Person 2: Live Here 5 Years Ago	POP--Demographic
224 p2oabsnt	25b - Person 2: Last Week Absent	POP--Occupation
225 p2oadd	1 - Person 2: Add	Form Management
226 p2oalone	17c - Person 2: Difficulty Shopping	POP--Disability
227 p2oam_pm	24a - Person 2: Time to Work am/pm	POP--Occupation
228 p2oarmed	27a - Person 2: Armed Forces	POP--Military
229 p2oblind	16a - Person 2: Blind or Deaf	POP--Disability
230 p2oborn	18 - Person 2: Under 16	POP--Demographic
231 p2ocancl	1 - Person 2: Cancel	Form Management
232 p2octlmt	22c - Person 2: Work Inside City Limits	POP--Occupation
233 p2octzn	13 - Person 2: Citizen	POP--Demographic
234 p2odegre	9 - Person 2: Highest Degree Completed	POP--Education
235 p2odress	17b - Person 2: Difficulty Dressing	POP--Disability
236 p2oetype	29 - Person 2: Class of Worker	POP--Occupation
237 p2ograde	8b - Person 2: Grade Level	POP--Education
238 p2ogrand	19a - Person 2: Grandchildren	POP--Demographic
239 p2ohisp	5 - Person 2: Hispanic Origin	POP--Ethnic
240 p2oint	31c - Person 2: Interest	POP--Income
241 p2ointls	31c - Person 2: Interest Loss	POP--Income
242 p2ojob	17d - Person 2: Difficulty Working	POP--Disability
243 p2olayof	25a - Person 2: Last Week Layoff	POP--Occupation
244 p2olimit	16b - Person 2: Limits Physical Activities	POP--Disability
245 p2olook	25d - Person 2: Looking for Work	POP--Occupation
246 p2olstwk	26 - Person 2: Last Worked	POP--Occupation
247 p2olvcty	15b - Person 2: Live Inside City Limits	POP--Demographic
248 p2omarry	7 - Person 2: Marital Status	POP--Demographic
249 p2omentl	17a - Person 2: Difficulty Learning	POP--Disability
250 p2omilit	20a - Person 2: Active Duty	POP--Military
251 p2oneeds	19b - Person 2: Responsible for Needs	POP--Disability
252 p2oother	31h - Person 2: Other Income	POP--Income
253 p2oproft	21 - Person 2: Work Last Week	POP--Occupation
254 p2orace	6 - Person 2: Race	POP--Race
255 p2orecal	25c - Person 2: Will Be Recalled	POP--Occupation
256 p2orel	2 - Person 2: Relationship	POP--Demographic
257 p2oresp	19c - Person 2: How Long	Residential Profile
258 p2oretir	31g - Person 2: Retirement Income	POP--Income

Field Name	Description	Category
259 p2oride	23b - Person 2: Carpool	POP--Occupation
260 p2oscool	8a - Person 2: Attend School	POP--Education
261 p2oselfe	31b - Person 2: Self- Person 2:employment Income	POP--Income
262 p2oserve	20b - Person 2: When on Active Duty	POP--Military
263 p2osex	3 - Person 2: Sex	POP--Demographic
264 p2oslfls	31b - Person 2: Self- Person 2:employment Loss	POP--Income
265 p2osocl	31d - Person 2: Social Security, Railroad Retirement	POP--Income
266 p2ospeak	11a - Person 2: Home Language	POP--Demographic
267 p2ospkwl	11c - Person 2: Speak English Well	POP--Demographic
268 p2ossi	31e - Person 2: SSI	POP--Income
269 p2ostart	25e - Person 2: Could Start Last Week	POP--Occupation
270 p2ototal	32 - Person 2: Total Income None	POP--Income
271 p2ototls	32 - Person 2: Total Income Loss	POP--Income
272 p2otrans	23a - Person 2: Work Vehicle	POP--Occupation
273 p2otype	27c - Person 2: Business Type	POP--Occupation
274 p2owages	31a - Person 2: Wages	POP--Income
275 p2owelfr	31f - Person 2: Welfare	POP--Income
276 p2owhrbn	12 - Person 2: Place of Birth	POP--Demographic
277 p2owork	30a - Person 2: Work Last Year	POP--Occupation
278 p2oyears	20c - Person 2: Years on Active Duty	POP--Military
279 p2race_1	6 - Person 2: Other Race	POP--Race
280 p2race19	6 - Person 2: Other Race	POP--Race
281 p2retir	31g - Person 2: Retirement Income Amount	POP--Income
282 p2selfe	31b - Person 2: Self Employment Income Amount	POP--Income
283 p2socl	31d - Person 2: Social Security, Railroad Retirement Amount	POP--Income
284 p2ssi	31e - Person 2: SSI Amount	POP--Income
285 p2state	22e - Person 2: Work State	POP--Occupation
286 p2time	24a - Person 2: Time Leave for Work	POP--Occupation
287 p2total	32 - Person 2: Total Income Amount	POP--Income
288 p2trib_1	6 - Person 2: Am Indian, Alaska Native Tribe	POP--Race
289 p2trib19	6 - Person 2: Am. Indian, AK Native - Tribe	POP--Race
290 p2wages	31a - Person 2: Wages Amount	POP--Income
291 p2weeks	30b - Person 2: Weeks Worked	POP--Occupation
292 p2welfr	31f - Person 2: Welfare Amount	POP--Income
293 p2yrmvus	14 - Person 2: Migration Year	POP--Demographic
294 p2zip	22f - Person 2: Work Zip Code	POP--Occupation
295 p3_other	31h - Person 3: Other Income Amount	POP--Income
296 p3_relo	2 - Person 3: Other Relative	POP--Demographic
297 p3actv_1	27b - Person 3: Industry	POP--Occupation
298 p3addr_1	22a - Person 3: Work Address	POP--Occupation
299 p3age	4 - Person 3: Age	POP--Demographic
300 p3asia_1	6 - Person 3: Other Asian	POP--Ethnic
301 p3asia19	6 - Person 3: Other Asian	POP--Ethnic
302 p3bnoth	12 - Person 3: Name of Country	POP--Demographic
303 p3bnus	12 - Person 3: Name of State	POP--Demographic
304 p3city	22b - Person 3: Work City	POP--Occupation
305 p3cntry	15a - Person 3: Migration Country	POP--Demographic
306 p3county	22d - Person 3: Work County	POP--Occupation
307 p3dob_d	4 - Person 3: Day of Birth	POP--Demographic
308 p3dob_m	4 - Person 3: Month of Birth	POP--Demographic
309 p3dob_y	4 - Person 3: Year of Birth	POP--Demographic
310 p3duty_1	28b - Person 3: Occupation Duties	POP--Occupation
311 p3empl_1	27a - Person 3: Employer	POP--Occupation

Field Name	Description	Category
312 p3ethn_1	10 - Person 3: Ancestry	POP--Ethnic
313 p3first	1 - Person 3: First Name	POP--Name
314 p3firsta	7 - Person 3: First Name	POP--Name
315 p3hisp_1	5 - Person 3: Other Hispanic Origin	POP--Ethnic
316 p3hisp19	5 - Person 3: Other Hispanic Origin	POP--Ethnic
317 p3hours	30c - Person 3: Hours Worked per Week	POP--Occupation
318 p3int	31c - Person 3: Interest Amount	POP--Income
319 p3kind_1	28a - Person 3: Occupation Kind of Work	POP--Occupation
320 p3lang	11b - Person 3: Language	POP--Demographic
321 p3last	1 - Person 3: Last Name	POP--Name
322 p3lasta	7 - Person 3: Last Name	POP--Name
323 p3lvcity	15b - Person 3: Migration City	POP--Demographic
324 p3lvcnty	15b - Person 3: Migration County	POP--Demographic
325 p3lvstat	15b - Person 3: Migration State	POP--Demographic
326 p3lvzip	15b - Person 3: Migration Zip Code	POP--Demographic
327 p3mi	1 - Person 3: Middle Initial	POP--Name
328 p3mia	7 - Person 3: Middle Initial	POP--Name
329 p3minute	24b - Person 3: Minutes to Work	POP--Occupation
330 p3o15age	19 - Person 3: Under 15 Interviewer Instruction	Form Management
331 p3o5ago	15a - Person 3: Live Here 5 Years Ago	POP--Demographic
332 p3oabsnt	25b - Person 3: Last Week Absent	POP--Occupation
333 p3oadd	1 - Person 3: Add	Form Management
334 p3oalone	17c - Person 3: Difficulty Shopping	POP--Disability
335 p3oam_pm	24a - Person 3: Time to Work am/pm	POP--Occupation
336 p3oarmed	27a - Person 3: Armed Forces	POP--Military
337 p3oblind	16a - Person 3: Blind or Deaf	POP--Disability
338 p3oborn	18 - Person 3: Under 17	POP--Demographic
339 p3ocancl	1 - Person 3: Cancel	Form Management
340 p3octlmt	22c - Person 3: Work Inside City Limits	POP--Occupation
341 p3octzn	13 - Person 3: Citizen	POP--Demographic
342 p3odegre	9 - Person 3: Highest Degree Completed	POP--Education
343 p3odress	17b - Person 3: Difficulty Dressing	POP--Disability
344 p3oetype	29 - Person 3: Class of Worker	POP--Occupation
345 p3ograde	8b - Person 3: Grade Level	POP--Education
346 p3ogrand	19a - Person 3: Grandchildren	POP--Demographic
347 p3ohisp	5 - Person 3: Hispanic Origin	POP--Ethnic
348 p3oint	31c - Person 3: Interest	POP--Income
349 p3ointls	31c - Person 3: Interest Loss	POP--Income
350 p3ojob	17d - Person 3: Difficulty Working	POP--Disability
351 p3olayof	25a - Person 3: Last Week Layoff	POP--Occupation
352 p3olimit	16b - Person 3: Limits Physical Activities	POP--Disability
353 p3olook	25d - Person 3: Looking for Work	POP--Occupation
354 p3olstwk	26 - Person 3: Last Worked	POP--Occupation
355 p3olvcty	15b - Person 3: Live Inside City Limits	POP--Demographic
356 p3omarry	7 - Person 3: Marital Status	POP--Demographic
357 p3omentl	17a - Person 3: Difficulty Learning	POP--Disability
358 p3omilit	20a - Person 3: Active Duty	POP--Military
359 p3oneeds	19b - Person 3: Responsible for Needs	POP--Disability
360 p3oother	31h - Person 3: Other Income	POP--Income
361 p3oproft	21 - Person 3: Work Last Week	POP--Occupation
362 p3orace	6 - Person 3: Race	POP--Race
363 p3orecal	25c - Person 3: Will Be Recalled	POP--Occupation
364 p3orel	2 - Person 3: Relationship	POP--Demographic

Field Name	Description	Category
365 p3oresp	19c - Person 3: How Long	Residential Profile
366 p3oretir	31g - Person 3: Retirement Income	POP--Income
367 p3oride	23b - Person 3: Carpool	POP--Occupation
368 p3oscool	8a - Person 3: Attend School	POP--Education
369 p3oselfe	31b - Person 3: Self- Person 3:employment Income	POP--Income
370 p3oserve	20b - Person 3: When on Active Duty	POP--Military
371 p3osex	3 - Person 3: Sex	POP--Demographic
372 p3oslfls	31b - Person 3: Self- Person 3:employment Loss	POP--Income
373 p3osocl	31d - Person 3: Social Security, Railroad Retirement	POP--Income
374 p3ospeak	11a - Person 3: Home Language	POP--Demographic
375 p3ospkwl	11c - Person 3: Speak English Well	POP--Demographic
376 p3ossi	31e - Person 3: SSI	POP--Income
377 p3ostart	25e - Person 3: Could Start Last Week	POP--Occupation
378 p3ototal	32 - Person 3: Total Income None	POP--Income
379 p3ototls	32 - Person 3: Total Income Loss	POP--Income
380 p3otrans	23a - Person 3: Work Vehicle	POP--Occupation
381 p3otype	27c - Person 3: Business Type	POP--Occupation
382 p3owages	31a - Person 3: Wages	POP--Income
383 p3owelfr	31f - Person 3: Welfare	POP--Income
384 p3owhrbn	12 - Person 3: Place of Birth	POP--Demographic
385 p3owork	30a - Person 3: Work Last Year	POP--Occupation
386 p3oyears	20c - Person 3: Years on Active Duty	POP--Military
387 p3race_1	6 - Person 3: Other Race	POP--Race
388 p3race19	6 - Person 3: Other Race	POP--Race
389 p3retir	31g - Person 3: Retirement Income Amount	POP--Income
390 p3selfe	31b - Person 3: Self Employment Income Amount	POP--Income
391 p3socl	31d - Person 3: Social Security, Railroad Retirement Amount	POP--Income
392 p3ssi	31e - Person 3: SSI Amount	POP--Income
393 p3state	22e - Person 3: Work State	POP--Occupation
394 p3time	24a - Person 3: Time Leave for Work	POP--Occupation
395 p3total	32 - Person 3: Total Income Amount	POP--Income
396 p3trib_1	6 - Person 3: Am Indian, Alaska Native Tribe	POP--Race
397 p3trib19	6 - Person 3: Am. Indian, AK Native - Tribe	POP--Race
398 p3wages	31a - Person 3: Wages Amount	POP--Income
399 p3weeks	30b - Person 3: Weeks Worked	POP--Occupation
400 p3welfr	31f - Person 3: Welfare Amount	POP--Income
401 p3yrmvus	14 - Person 3: Migration Year	POP--Demographic
402 p3zip	22f - Person 3: Work Zip Code	POP--Occupation
403 p4_other	31h - Person 4: Other Income Amount	POP--Income
404 p4_relo	2 - Person 4: Other Relative	POP--Demographic
405 p4actv_1	27b - Person 4: Industry	POP--Occupation
406 p4addr_1	22a - Person 4: Work Address	POP--Occupation
407 p4age	4 - Person 4: Age	POP--Demographic
408 p4asia_1	6 - Person 4: Other Asian	POP--Ethnic
409 p4asia19	6 - Person 4: Other Asian	POP--Ethnic
410 p4bnoth	12 - Person 4: Name of Country	POP--Demographic
411 p4bnus	12 - Person 4: Name of State	POP--Demographic
412 p4city	22b - Person 4: Work City	POP--Occupation
413 p4cntry	15a - Person 4: Migration Country	POP--Demographic
414 p4county	22d - Person 4: Work County	POP--Occupation
415 p4dob_d	4 - Person 4: Day of Birth	POP--Demographic
416 p4dob_m	4 - Person 4: Month of Birth	POP--Demographic
417 p4dob_y	4 - Person 4: Year of Birth	POP--Demographic

Field Name	Description	Category
418 p4duty_1	28b - Person 4: Occupation Duties	POP--Occupation
419 p4empl_1	27a - Person 4: Employer	POP--Occupation
420 p4ethn_1	10 - Person 4: Ancestry	POP--Ethnic
421 p4first	1 - Person 4: First Name	POP--Name
422 p4firsta	7 - Person 4: First Name	POP--Name
423 p4hisp_1	5 - Person 4: Other Hispanic Origin	POP--Ethnic
424 p4hisp19	5 - Person 4: Other Hispanic Origin	POP--Ethnic
425 p4hours	30c - Person 4: Hours Worked per Week	POP--Occupation
426 p4int	31c - Person 4: Interest Amount	POP--Income
427 p4kind_1	28a - Person 4: Occupation Kind of Work	POP--Occupation
428 p4lang	11b - Person 4: Language	POP--Demographic
429 p4last	1 - Person 4: Last Name	POP--Name
430 p4lasta	7 - Person 4: Last Name	POP--Name
431 p4lvcity	15b - Person 4: Migration City	POP--Demographic
432 p4lvcnty	15b - Person 4: Migration County	POP--Demographic
433 p4lvstat	15b - Person 4: Migration State	POP--Demographic
434 p4lvzip	15b - Person 4: Migration Zip Code	POP--Demographic
435 p4mi	1 - Person 4: Middle Initial	POP--Name
436 p4mia	7 - Person 4: Middle Initial	POP--Name
437 p4minute	24b - Person 4: Minutes to Work	POP--Occupation
438 p4o15age	19 - Person 4: Under 15 Interviewer Instruction	Form Management
439 p4o5ago	15a - Person 4: Live Here 5 Years Ago	POP--Demographic
440 p4oabsnt	25b - Person 4: Last Week Absent	POP--Occupation
441 p4oadd	1 - Person 4: Add	Form Management
442 p4oalone	17c - Person 4: Difficulty Shopping	POP--Disability
443 p4oam_pm	24a - Person 4: Time to Work am/pm	POP--Occupation
444 p4oarmed	27a - Person 4: Armed Forces	POP--Military
445 p4oblind	16a - Person 4: Blind or Deaf	POP--Disability
446 p4oborn	18 - Person 4: Under 18	POP--Demographic
447 p4ocancl	1 - Person 4: Cancel	Form Management
448 p4octlmt	22c - Person 4: Work Inside City Limits	POP--Occupation
449 p4octzn	13 - Person 4: Citizen	POP--Demographic
450 p4odegre	9 - Person 4: Highest Degree Completed	POP--Education
451 p4odress	17b - Person 4: Difficulty Dressing	POP--Disability
452 p4oetype	29 - Person 4: Class of Worker	POP--Occupation
453 p4ograde	8b - Person 4: Grade Level	POP--Education
454 p4ogrand	19a - Person 4: Grandchildren	POP--Demographic
455 p4ohisp	5 - Person 4: Hispanic Origin	POP--Ethnic
456 p4oint	31c - Person 4: Interest	POP--Income
457 p4ointls	31c - Person 4: Interest Loss	POP--Income
458 p4ojob	17d - Person 4: Difficulty Working	POP--Disability
459 p4olayof	25a - Person 4: Last Week Layoff	POP--Occupation
460 p4olimit	16b - Person 4: Limits Physical Activities	POP--Disability
461 p4olook	25d - Person 4: Looking for Work	POP--Occupation
462 p4olstwk	26 - Person 4: Last Worked	POP--Occupation
463 p4olvcty	15b - Person 4: Live Inside City Limits	POP--Demographic
464 p4omarry	7 - Person 4: Marital Status	POP--Demographic
465 p4omentl	17a - Person 4: Difficulty Learning	POP--Disability
466 p4omilit	20a - Person 4: Active Duty	POP--Military
467 p4oneeds	19b - Person 4: Responsible for Needs	POP--Disability
468 p4oother	31h - Person 4: Other Income	POP--Income
469 p4oproft	21 - Person 4: Work Last Week	POP--Occupation
470 p4orace	6 - Person 4: Race	POP--Race

Field Name	Description	Category
471 p4orecal	25c - Person 4: Will Be Recalled	POP--Occupation
472 p4orel	2 - Person 4: Relationship	POP--Demographic
473 p4oresp	19c - Person 4: How Long	Residential Profile
474 p4oretir	31g - Person 4: Retirement Income	POP--Income
475 p4oride	23b - Person 4: Carpool	POP--Occupation
476 p4oscool	8a - Person 4: Attend School	POP--Education
477 p4oselfe	31b - Person 4: Self- Person 4:employment Income	POP--Income
478 p4oserve	20b - Person 4: When on Active Duty	POP--Military
479 p4osex	3 - Person 4: Sex	POP--Demographic
480 p4oslfls	31b - Person 4: Self- Person 4:employment Loss	POP--Income
481 p4osocl	31d - Person 4: Social Security, Railroad Retirement	POP--Income
482 p4ospeak	11a - Person 4: Home Language	POP--Demographic
483 p4ospkwl	11c - Person 4: Speak English Well	POP--Demographic
484 p4ossi	31e - Person 4: SSI	POP--Income
485 p4ostart	25e - Person 4: Could Start Last Week	POP--Occupation
486 p4ototal	32 - Person 4: Total Income None	POP--Income
487 p4ototls	32 - Person 4: Total Income Loss	POP--Income
488 p4otrans	23a - Person 4: Work Vehicle	POP--Occupation
489 p4otype	27c - Person 4: Business Type	POP--Occupation
490 p4owages	31a - Person 4: Wages	POP--Income
491 p4owelfr	31f - Person 4: Welfare	POP--Income
492 p4owhrbn	12 - Person 4: Place of Birth	POP--Demographic
493 p4owork	30a - Person 4: Work Last Year	POP--Occupation
494 p4oyears	20c - Person 4: Years on Active Duty	POP--Military
495 p4race_1	6 - Person 4: Other Race	POP--Race
496 p4race19	6 - Person 4: Other Race	POP--Race
497 p4retir	31g - Person 4: Retirement Income Amount	POP--Income
498 p4selfe	31b - Person 4: Self Employment Income Amount	POP--Income
499 p4socl	31d - Person 4: Social Security, Railroad Retirement Amount	POP--Income
500 p4ssi	31e - Person 4: SSI Amount	POP--Income
501 p4state	22e - Person 4: Work State	POP--Occupation
502 p4time	24a - Person 4: Time Leave for Work	POP--Occupation
503 p4total	32 - Person 4: Total Income Amount	POP--Income
504 p4trib_1	6 - Person 4: Am Indian, Alaska Native Tribe	POP--Race
505 p4trib19	6 - Person 4: Am. Indian, AK Native - Tribe	POP--Race
506 p4wages	31a - Person 4: Wages Amount	POP--Income
507 p4weeks	30b - Person 4: Weeks Worked	POP--Occupation
508 p4welfr	31f - Person 4: Welfare Amount	POP--Income
509 p4yrmvus	14 - Person 4: Migration Year	POP--Demographic
510 p4zip	22f - Person 4: Work Zip Code	POP--Occupation
511 p5_other	31h - Person 5: Other Income Amount	POP--Income
512 p5_relo	2 - Person 5: Other Relative	POP--Demographic
513 p5actv_1	27b - Person 5: Industry	POP--Occupation
514 p5addr_1	22a - Person 5: Work Address	POP--Occupation
515 p5age	4 - Person 5: Age	POP--Demographic
516 p5asia_1	6 - Person 5: Other Asian	POP--Ethnic
517 p5asia19	6 - Person 5: Other Asian	POP--Ethnic
518 p5bnoth	12 - Person 5: Name of Country	POP--Demographic
519 p5bnus	12 - Person 5: Name of State	POP--Demographic
520 p5city	22b - Person 5: Work City	POP--Occupation
521 p5cntry	15a - Person 5: Migration Country	POP--Demographic
522 p5county	22d - Person 5: Work County	POP--Occupation
523 p5dob_d	4 - Person 5: Day of Birth	POP--Demographic

Field Name	Description	Category
524 p5dob_m	4 - Person 5: Month of Birth	POP--Demographic
525 p5dob_y	4 - Person 5: Year of Birth	POP--Demographic
526 p5duty_1	28b - Person 5: Occupation Duties	POP--Occupation
527 p5empl_1	27a - Person 5: Employer	POP--Occupation
528 p5ethn_1	10 - Person 5: Ancestry	POP--Ethnic
529 p5first	1 - Person 5: First Name	POP--Name
530 p5firsta	7 - Person 5: First Name	POP--Name
531 p5hisp_1	5 - Person 5: Other Hispanic Origin	POP--Ethnic
532 p5hisp19	5 - Person 5: Other Hispanic Origin	POP--Ethnic
533 p5hours	30c - Person 5: Hours Worked per Week	POP--Occupation
534 p5int	31c - Person 5: Interest Amount	POP--Income
535 p5kind_1	28a - Person 5: Occupation Kind of Work	POP--Occupation
536 p5lang	11b - Person 5: Language	POP--Demographic
537 p5last	1 - Person 5: Last Name	POP--Name
538 p5lasta	7 - Person 5: Last Name	POP--Name
539 p5lvcity	15b - Person 5: Migration City	POP--Demographic
540 p5lvcnty	15b - Person 5: Migration County	POP--Demographic
541 p5lvstat	15b - Person 5: Migration State	POP--Demographic
542 p5lvzip	15b - Person 5: Migration Zip Code	POP--Demographic
543 p5mi	1 - Person 5: Middle Initial	POP--Name
544 p5mia	7 - Person 5: Middle Initial	POP--Name
545 p5minute	24b - Person 5: Minutes to Work	POP--Occupation
546 p5o15age	19 - Person 5: Under 15 Interviewer Instruction	Form Management
547 p5o5ago	15a - Person 5: Live Here 5 Years Ago	POP--Demographic
548 p5oabsnt	25b - Person 5: Last Week Absent	POP--Occupation
549 p5oadd	1 - Person 5: Add	Form Management
550 p5oalone	17c - Person 5: Difficulty Shopping	POP--Disability
551 p5oam_pm	24a - Person 5: Time to Work am/pm	POP--Occupation
552 p5oarmed	27a - Person 5: Armed Forces	POP--Military
553 p5oblind	16a - Person 5: Blind or Deaf	POP--Disability
554 p5oborn	18 - Person 5: Under 19	POP--Demographic
555 p5ocancl	1 - Person 5: Cancel	Form Management
556 p5octlmt	22c - Person 5: Work Inside City Limits	POP--Occupation
557 p5octzn	13 - Person 5: Citizen	POP--Demographic
558 p5odegre	9 - Person 5: Highest Degree Completed	POP--Education
559 p5odress	17b - Person 5: Difficulty Dressing	POP--Disability
560 p5oetype	29 - Person 5: Class of Worker	POP--Occupation
561 p5ograde	8b - Person 5: Grade Level	POP--Education
562 p5ogrand	19a - Person 5: Grandchildren	POP--Demographic
563 p5ohisp	5 - Person 5: Hispanic Origin	POP--Ethnic
564 p5oint	31c - Person 5: Interest	POP--Income
565 p5ointls	31c - Person 5: Interest Loss	POP--Income
566 p5ojob	17d - Person 5: Difficulty Working	POP--Disability
567 p5olayof	25a - Person 5: Last Week Layoff	POP--Occupation
568 p5olimit	16b - Person 5: Limits Physical Activities	POP--Disability
569 p5olook	25d - Person 5: Looking for Work	POP--Occupation
570 p5olstwk	26 - Person 5: Last Worked	POP--Occupation
571 p5olvcty	15b - Person 5: Live Inside City Limits	POP--Demographic
572 p5omarry	7 - Person 5: Marital Status	POP--Demographic
573 p5omentl	17a - Person 5: Difficulty Learning	POP--Disability
574 p5omilit	20a - Person 5: Active Duty	POP--Military
575 p5oneeds	19b - Person 5: Responsible for Needs	POP--Disability
576 p5oother	31h - Person 5: Other Income	POP--Income

Field Name	Description	Category
577 p5oproft	21 - Person 5: Work Last Week	POP--Occupation
578 p5orace	6 - Person 5: Race	POP--Race
579 p5orecal	25c - Person 5: Will Be Recalled	POP--Occupation
580 p5orel	2 - Person 5: Relationship	POP--Demographic
581 p5oresp	19c - Person 5: How Long	Residential Profile
582 p5oretir	31g - Person 5: Retirement Income	POP--Income
583 p5oride	23b - Person 5: Carpool	POP--Occupation
584 p5oscool	8a - Person 5: Attend School	POP--Education
585 p5oselfe	31b - Person 5: Self- Person 5:employment Income	POP--Income
586 p5oserve	20b - Person 5: When on Active Duty	POP--Military
587 p5osex	3 - Person 5: Sex	POP--Demographic
588 p5oslfls	31b - Person 5: Self- Person 5:employment Loss	POP--Income
589 p5osocl	31d - Person 5: Social Security, Railroad Retirement	POP--Income
590 p5ospeak	11a - Person 5: Home Language	POP--Demographic
591 p5ospkwl	11c - Person 5: Speak English Well	POP--Demographic
592 p5ossi	31e - Person 5: SSI	POP--Income
593 p5ostart	25e - Person 5: Could Start Last Week	POP--Occupation
594 p5ototal	32 - Person 5: Total Income None	POP--Income
595 p5ototls	32 - Person 5: Total Income Loss	POP--Income
596 p5otrans	23a - Person 5: Work Vehicle	POP--Occupation
597 p5otype	27c - Person 5: Business Type	POP--Occupation
598 p5owages	31a - Person 5: Wages	POP--Income
599 p5owelfr	31f - Person 5: Welfare	POP--Income
600 p5owhrbn	12 - Person 5: Place of Birth	POP--Demographic
601 p5owork	30a - Person 5: Work Last Year	POP--Occupation
602 p5oyears	20c - Person 5: Years on Active Duty	POP--Military
603 p5race_1	6 - Person 5: Other Race	POP--Race
604 p5race19	6 - Person 5: Other Race	POP--Race
605 p5retir	31g - Person 5: Retirement Income Amount	POP--Income
606 p5selfe	31b - Person 5: Self Employment Income Amount	POP--Income
607 p5socl	31d - Person 5: Social Security, Railroad Retirement Amount	POP--Income
608 p5ssi	31e - Person 5: SSI Amount	POP--Income
609 p5state	22e - Person 5: Work State	POP--Occupation
610 p5time	24a - Person 5: Time Leave for Work	POP--Occupation
611 p5total	32 - Person 5: Total Income Amount	POP--Income
612 p5trib_1	6 - Person 5: Am Indian, Alaska Native Tribe	POP--Race
613 p5trib19	6 - Person 5: Am. Indian, AK Native - Tribe	POP--Race
614 p5wages	31a - Person 5: Wages Amount	POP--Income
615 p5weeks	30b - Person 5: Weeks Worked	POP--Occupation
616 p5welfr	31f - Person 5: Welfare Amount	POP--Income
617 p5yrmvus	14 - Person 5: Migration Year	POP--Demographic
618 p5zip	22f - Person 5: Work Zip Code	POP--Occupation
619 p6_other	31h - Person 6: Other Income Amount	POP--Income
620 p6_relo	2 - Person 6: Other Relative	POP--Demographic
621 p6actv_1	27b - Person 6: Industry	POP--Occupation
622 p6addr_1	22a - Person 6: Work Address	POP--Occupation
623 p6age	4 - Person 6: Age	POP--Demographic
624 p6asia_1	6 - Person 6: Other Asian	POP--Ethnic
625 p6asia19	6 - Person 6: Other Asian	POP--Ethnic
626 p6bnoth	12 - Person 6: Name of Country	POP--Demographic
627 p6bnus	12 - Person 6: Name of State	POP--Demographic
628 p6city	22b - Person 6: Work City	POP--Occupation
629 p6cntry	15a - Person 6: Migration Country	POP--Demographic

Field Name	Description	Category
630 p6county	22d - Person 6: Work County	POP--Occupation
631 p6dob_d	4 - Person 6: Day of Birth	POP--Demographic
632 p6dob_m	4 - Person 6: Month of Birth	POP--Demographic
633 p6dob_y	4 - Person 6: Year of Birth	POP--Demographic
634 p6duty_1	28b - Person 6: Occupation Duties	POP--Occupation
635 p6empl_1	27a - Person 6: Employer	POP--Occupation
636 p6ethn_1	10 - Person 6: Ancestry	POP--Ethnic
637 p6first	1 - Person 6: First Name	POP--Name
638 p6hisp_1	5 - Person 6: Other Hispanic Origin	POP--Ethnic
639 p6hisp19	5 - Person 6: Other Hispanic Origin	POP--Ethnic
640 p6hours	30c - Person 6: Hours Worked per Week	POP--Occupation
641 p6int	31c - Person 6: Interest Amount	POP--Income
642 p6kind_1	28a - Person 6: Occupation Kind of Work	POP--Occupation
643 p6lang	11b - Person 6: Language	POP--Demographic
644 p6last	1 - Person 6: Last Name	POP--Name
645 p6lvcity	15b - Person 6: Migration City	POP--Demographic
646 p6lvcnty	15b - Person 6: Migration County	POP--Demographic
647 p6lvstat	15b - Person 6: Migration State	POP--Demographic
648 p6lvzip	15b - Person 6: Migration Zip Code	POP--Demographic
649 p6mi	1 - Person 6: Middle Initial	POP--Name
650 p6minute	24b - Person 6: Minutes to Work	POP--Occupation
651 p6o5ago	15a - Person 6: Live Here 5 Years Ago	POP--Demographic
652 p6oabsnt	25b - Person 6: Last Week Absent	POP--Occupation
653 p6oalone	17c - Person 6: Difficulty Shopping	POP--Disability
654 p6oam_pm	24a - Person 6: Time to Work am/pm	POP--Occupation
655 p6oarmed	27a - Person 6: Armed Forces	POP--Military
656 p6oblind	16a - Person 6: Blind or Deaf	POP--Disability
657 p6oborn	18 - Person 6: Under 20	POP--Demographic
658 p6octlmt	22c - Person 6: Work Inside City Limits	POP--Occupation
659 p6octzn	13 - Person 6: Citizen	POP--Demographic
660 p6odegre	9 - Person 6: Highest Degree Completed	POP--Education
661 p6odress	17b - Person 6: Difficulty Dressing	POP--Disability
662 p6oetype	29 - Person 6: Class of Worker	POP--Occupation
663 p6ograde	8b - Person 6: Grade Level	POP--Education
664 p6ogrand	19a - Person 6: Grandchildren	POP--Demographic
665 p6ohisp	5 - Person 6: Hispanic Origin	POP--Ethnic
666 p6oint	31c - Person 6: Interest	POP--Income
667 p6ointls	31c - Person 6: Interest Loss	POP--Income
668 p6ojob	17d - Person 6: Difficulty Working	POP--Disability
669 p6olayof	25a - Person 6: Last Week Layoff	POP--Occupation
670 p6olimit	16b - Person 6: Limits Physical Activities	POP--Disability
671 p6olook	25d - Person 6: Looking for Work	POP--Occupation
672 p6olstwk	26 - Person 6: Last Worked	POP--Occupation
673 p6olvcty	15b - Person 6: Live Inside City Limits	POP--Demographic
674 p6omarry	7 - Person 6: Marital Status	POP--Demographic
675 p6omentl	17a - Person 6: Difficulty Learning	POP--Disability
676 p6omilit	20a - Person 6: Active Duty	POP--Military
677 p6oneeds	19b - Person 6: Responsible for Needs	POP--Disability
678 p6oother	31h - Person 6: Other Income	POP--Income
679 p6oproft	21 - Person 6: Work Last Week	POP--Occupation
680 p6orace	6 - Person 6: Race	POP--Race
681 p6orecal	25c - Person 6: Will Be Recalled	POP--Occupation
682 p6orel	2 - Person 6: Relationship	POP--Demographic

Field Name	Description	Category
683 p6oresp	19c - Person 6: How Long	Residential Profile
684 p6oretir	31g - Person 6: Retirement Income	POP--Income
685 p6oride	23b - Person 6: Carpool	POP--Occupation
686 p6oscool	8a - Person 6: Attend School	POP--Education
687 p6oselfe	31b - Person 6: Self- Person 6:employment Income	POP--Income
688 p6oserve	20b - Person 6: When on Active Duty	POP--Military
689 p6osex	3 - Person 6: Sex	POP--Demographic
690 p6oslfls	31b - Person 6: Self- Person 6:employment Loss	POP--Income
691 p6osocl	31d - Person 6: Social Security, Railroad Retirement	POP--Income
692 p6ospeak	11a - Person 6: Home Language	POP--Demographic
693 p6ospkwl	11c - Person 6: Speak English Well	POP--Demographic
694 p6ossi	31e - Person 6: SSI	POP--Income
695 p6ostart	25e - Person 6: Could Start Last Week	POP--Occupation
696 p6ototal	32 - Person 6: Total Income None	POP--Income
697 p6ototls	32 - Person 6: Total Income Loss	POP--Income
698 p6otrans	23a - Person 6: Work Vehicle	POP--Occupation
699 p6otype	27c - Person 6: Business Type	POP--Occupation
700 p6owages	31a - Person 6: Wages	POP--Income
701 p6owelfr	31f - Person 6: Welfare	POP--Income
702 p6owhrbn	12 - Person 6: Place of Birth	POP--Demographic
703 p6owork	30a - Person 6: Work Last Year	POP--Occupation
704 p6oyears	20c - Person 6: Years on Active Duty	POP--Military
705 p6race_1	6 - Person 6: Other Race	POP--Race
706 p6race19	6 - Person 6: Other Race	POP--Race
707 p6retir	31g - Person 6: Retirement Income Amount	POP--Income
708 p6selfe	31b - Person 6: Self Employment Income Amount	POP--Income
709 p6socl	31d - Person 6: Social Security, Railroad Retirement Amount	POP--Income
710 p6ssi	31e - Person 6: SSI Amount	POP--Income
711 p6state	22e - Person 6: Work State	POP--Occupation
712 p6time	24a - Person 6: Time Leave for Work	POP--Occupation
713 p6total	32 - Person 6: Total Income Amount	POP--Income
714 p6trib_1	6 - Person 6: Am Indian, Alaska Native Tribe	POP--Race
715 p6trib19	6 - Person 6: Am. Indian, AK Native - Tribe	POP--Race
716 p6wages	31a - Person 6: Wages Amount	POP--Income
717 p6weeks	30b - Person 6: Weeks Worked	POP--Occupation
718 p6welfr	31f - Person 6: Welfare Amount	POP--Income
719 p6yrmvus	14 - Person 6: Migration Year	POP--Demographic
720 p6zip	22f - Person 6: Work Zip Code	POP--Occupation
721 p7first	Person 7: First Name	POP--Name
722 p7last	Person 7: Last Name	POP--Name
723 p7mi	Person 7: Middle Initial	POP--Name
724 p8first	Person 8: First Name	POP--Name
725 p8last	Person 8: Last Name	POP--Name
726 p8mi	Person 8: Middle Initial	POP--Name
727 p9first	Person 9: First Name	POP--Name
728 p9last	Person 9: Last Name	POP--Name
729 p9mi	Person 9: Middle Initial	POP--Name
730 r10first	Roster: Person 10 First Name	POP--Name
731 r10last	Roster: Person 10 Last Name	POP--Name
732 r10mi	Roster: Person 10 Middle Initial	POP--Name
733 r11first	Roster: Person 11 First Name	POP--Name
734 r11last	Roster: Person 11 Last Name	POP--Name
735 r11mi	Roster: Person 11 Middle Initial	POP--Name

Field Name	Description	Category
736 r12first	Roster: Person 12 First Name	POP--Name
737 r12last	Roster: Person 12 Last Name	POP--Name
738 r12mi	Roster: Person 12 Middle Initial	POP--Name
739 r1first	Roster: Person 1 First Name	POP--Name
740 r1last	Roster: Person 1 Last Name	POP--Name
741 r1mi	Roster: Person 1 Middle Initial	POP--Name
742 r2first	Roster: Person 2 First Name	POP--Name
743 r2last	Roster: Person 2 Last Name	POP--Name
744 r2mi	Roster: Person 2 Middle Initial	POP--Name
745 r2odayev	R2 - Time to Call	Form Management
746 r3first	Roster: Person 3 First Name	POP--Name
747 r3last	Roster: Person 3 Last Name	POP--Name
748 r3mi	Roster: Person 3 Middle Initial	POP--Name
749 r3orespo	R3 - Respondent Status	Form Management
750 r4first	Roster: Person 4 First Name	POP--Name
751 r4last	Roster: Person 4 Last Name	POP--Name
752 r4mi	Roster: Person 4 Middle Initial	POP--Name
753 r5first	Roster: Person 5 First Name	POP--Name
754 r5last	Roster: Person 5 Last Name	POP--Name
755 r5mi	Roster: Person 5 Middle Initial	POP--Name
756 r6first	Roster: Person 6 First Name	POP--Name
757 r6last	Roster: Person 6 Last Name	POP--Name
758 r6mi	Roster: Person 6 Middle Initial	POP--Name
759 r7first	Roster: Person 7 First Name	POP--Name
760 r7last	Roster: Person 7 Last Name	POP--Name
761 r7mi	Roster: Person 7 Middle Initial	POP--Name
762 r8first	Roster: Person 8 First Name	POP--Name
763 r8last	Roster: Person 8 Last Name	POP--Name
764 r8mi	Roster: Person 8 Middle Initial	POP--Name
765 r9first	Roster: Person 9 First Name	POP--Name
766 r9last	Roster: Person 9 Last Name	POP--Name
767 r9mi	Roster: Person 9 Middle Initial	POP--Name
768 rc_d1	Record of Contact 1 - Day	Form Management
769 rc_d2	Record of Contact 2 - Day	Form Management
770 rc_d3	Record of Contact 3 - Day	Form Management
771 rc_d4	Record of Contact 4 - Day	Form Management
772 rc_d5	Record of Contact 5 - Day	Form Management
773 rc_d6	Record of Contact 6 - Day	Form Management
774 rc_m1	Record of Contact 1 - Month	Form Management
775 rc_m2	Record of Contact 2 - Month	Form Management
776 rc_m3	Record of Contact 3 - Month	Form Management
777 rc_m4	Record of Contact 4 - Month	Form Management
778 rc_m5	Record of Contact 5 - Month	Form Management
779 rc_m6	Record of Contact 6 - Month	Form Management
780 rc_oc1	Record of Contact 1 - Outcome	Form Management
781 rc_oc2	Record of Contact 2 - Outcome	Form Management
782 rc_oc3	Record of Contact 3 - Outcome	Form Management
783 rc_oc4	Record of Contact 4 - Outcome	Form Management
784 rc_oc5	Record of Contact 5 - Outcome	Form Management
785 rc_oc6	Record of Contact 6 - Outcome	Form Management
786 rc_t1	Record of Contact 1 - Time	Form Management
787 rc_t2	Record of Contact 2 - Time	Form Management
788 rc_t3	Record of Contact 3 - Time	Form Management

Field Name	Description	Category
789 rc_t4	Record of Contact 4 - Time	Form Management
790 rc_t5	Record of Contact 5 - Time	Form Management
791 rc_t6	Record of Contact 6 - Time	Form Management
792 rco_ap1	Record of Contact 1 - am/pm	Form Management
793 rco_ap2	Record of Contact 2 - am/pm	Form Management
794 rco_ap3	Record of Contact 3 - am/pm	Form Management
795 rco_ap4	Record of Contact 4 - am/pm	Form Management
796 rco_ap5	Record of Contact 5 - am/pm	Form Management
797 rco_ap6	Record of Contact 6 - am/pm	Form Management
798 rco_typ2	Record of Contact 2 - Type	Form Management
799 rco_typ3	Record of Contact 3 - Type	Form Management
800 rco_typ4	Record of Contact 4 - Type	Form Management
801 rco_typ5	Record of Contact 5 - Type	Form Management
802 rco_typ6	Record of Contact 6 - Type	Form Management
803 rifirst	R1 - Respondent's First Name	POP--Name
804 rilast	R1 - Respondent's Last Name	POP--Name
805 rn_pop	1 - Household: Number of People	POP--Demographic
806 rnohouse	2 - Household: Ownership Status	Residential Profile
807 s1ointro	S1 - Introduction	Form Management
808 s2ointro	S2 - Live Here April 1	Form Management
809 s3ointro	S3 - Seasonal Home	Form Management
810 s4ointro	S4 - Vacant or Occupied	Form Management

Appendix D: Record Counts Before and After Unduplication

In this appendix, we show the count of records in the raw data files before and after unduplication. A duplicate is a repeated combination of form, field, and Census ID number in a file. We include this information for anyone concerned about the reduction due to unduplication. The reduction is slight. We believe it is not enough to skew the analysis in this evaluation.

Table D1. Record Counts Before and After Duplication

Data File	Record Count Before Unduplication	Record Count After Unduplication
RCC 21	5,951,010	5,839,840
RCC 22	3,835,616	3,751,466
RCC 23	5,467,382	5,372,883
RCC 24	5,943,969	5,853,332
RCC 25	6,365,741	6,279,896
RCC 26	6,714,557	6,581,710
RCC 27	5,075,565	5,001,248
RCC 28	7,140,822	7,012,029
RCC 29	6,315,054	6,198,035
RCC 30	6,664,514	6,533,146
RCC 31	5,263,145	5,166,440
RCC 32	4,963,912	4,891,749
Total	69,701,287	68,481,774
File of Disagreements between Methods	1,725,518	1,715,967

Appendix E: Approximate 90 Percent Confidence Intervals for the Median

In this appendix, we describe the distribution free method used in this evaluation to approximate 90 percent confidence intervals for the median data capture error rate. For cases where we felt there were too few data points, we did not compute a confidence interval.

- Let n be the number of observations in the data set
- Compute the square root of n. Multiply the square root of n by 0.8. Call the result s
- Find integer nearest ((n+1)/2) - s. Call the result L.
- Find the integer nearest ((n+1)/2) + s. Call the result U.
- Sort the observations from lowest to highest.
- After sorting, find the observations at positions L and U**.**
- The values at observations L and U are the boundaries of the approximate confidence interval.

We modify this procedure for the confidence intervals shown in section 4.1.1. We conclude the median rates for the data capture modes are significantly different if they do not overlap. With three modes of data capture, there are three possible pairwise comparisons.

To test in this manner whether the medians differ significantly at the 90 percent level of confidence, the confidence levels for each individual median must be higher than 90 percent to account for multiple pairwise comparisons. A conservative estimate of the higher confidence is available by taking the nth root of 90 percent, where n is the number of comparisons. With three comparisons, this leads to the cube root of 90 percent, 96.5 percent.

In discussing nonparametric confidence intervals for the median, the Wallis text in the reference list says the multiple in step 2 of the above procedure should be 1.0 for the 95 percent level and 1.3 for the 99 percent level. Interpolating between 1.0 and 1.3, we select 1.2 for the multiplier more appropriate to 96.5 percent. We substitute 1.2 for 0.8 in step 2 in arriving at the confidence intervals shown in section 4.1.1.

Appendix F: Formulas for Median, Quartiles, and Outliers

In this appendix, we demonstrate with an example the formulas we used to computerize the calculation of the medians, quartiles, and outliers in this evaluation.

Item A. Raw data for example

1. 74
2. 86
3. 88
4. 89
5. 89
6. 91
7. 91
8. 91
9. 94
10. 95
11. 95
12. 96
13. 97

Item B. Finding the Median (M)

1. There are 13 data points.
2. Divide 13 by 2. Obtain 6.5. Round to the nearest integer greater than or equal to 6.5, 7.
3. Find the data point with a rank of 7. This is 91.
4. The median is 91.

If there are an even number of data points, the procedure works differently. We repeat it to show how to find the median considering only the first twelve data points.

1. There are twelve data points.
2. Divide twelve by 2. Obtain 6. Round to the nearest integer less than or equal to 6, 6.
3. Find the data point with a rank of 6. This is 91.
4. Go up one more observation. Take the one with a rank of 7. This is 91.
5. Average the observations with ranks 6 and 7. This is (91 + 91)/2 = 91.
6. The median is 91.

Item C. Finding the First Quartile (Q1)

1. There are 13 data points. Divide 13 by 4. Obtain 3.25.
2. Round 3.25 to nearest integer less than or equal to 3.25, 3.
3. Take the difference between 3.25 and 3. This is 0.25.
4. Find the observation with a rank of 3. This is 88.

5. Go up one more observation. Take the one with a rank of 4. This is 89.
6. Take the difference between the two observations. This is 89 - 88 = 1.
7. Multiply the difference in step 3 by the difference in step 6. This is 0.25 x 1 = 0.25.
8. Add the result in step 7 to the value with a rank of 3. This is 88 + 0.25 = 88.25.
9. The first quartile for these 13 data points is 88.25.

Item D. Finding the Third Quartile (Q3)

1. There are 13 data points. Divide 13 by 4. Multiply by 3. Obtain 9.75.
2. Round 9.75 to nearest integer less than or equal to 9.75, 9.
3. Take the difference between 9.75 and 9. This is 0.75.
4. Find the observation with a rank of 9. This is 94.
5. Go up one more observation. Take the one with a rank of 10. This is 95.
6. Take the difference between the two observations. This is 95 - 94 = 1.
7. Multiply the difference in step 3 by the difference in step 6. This is 0.75 x 1 = 0.75.
8. Add the result in step 7 to the value with a rank of 9. This is 94 + 0.75 = 94.75.
9. The third quartile for these 13 data points is 94.75.

Item E. Finding the Interquartile Range (IQR)

1. Take the value for the first quartile, 88.25.
2. Take the value for the third quartile, 94.75.
3. Find the difference. 94.75 - 88.25 = 6.50.
4. The interquartile range is 6.50.

Item F. Finding Very Low Outliers

1. Multiply the interquartile range by 3. 6.5 x 3 = 19.5.
2. Subtract the result from the median. 91 - 19.5 = 71.5.
3. Any values below 71.5 are very low outliers.

Item G. Finding Low Outliers

1. Multiply the interquartile range by 1.5. 6.5 x 1.5 = 9.75.
2. Subtract the result from the median. 91 - 9.75 = 81.25.
3. Any values at or above 71.5 and below 81.25 are low outliers.

Item H. Finding Very High Outliers

1. Multiply the interquartile range by 3. 6.5 x 3 = 19.5.
2. Add the result to the median. 91 + 19.5 = 110.5.
3. Any values above 110.5 are very high outliers.

Item I. Finding High Outliers

1. Multiply the interquartile range by 1.5. 6.5 x 1.5 = 9.75.
2. Add the result to the median. 91 + 9.75 = 100.5.
3. Any values above 100.5 and at or below 110.5 are high outliers.

For our example data set, only one value, 74, is an outlier, and it is classified as a low outlier.

Appendix G: Pseudocode for the Soft Match Algorithm

In this appendix, we show pseudocode for the soft match algorithm. The soft match algorithm compares the characters read by the automated technology and by KFI for a given field. It measures how much the readings from each method diverge and assigns a score. If the score is high enough, the reading from the automated technology is classified as a soft match error.

For the captured field do a tally TA(I), (I = 0, 1, 2, 3), of characters as follows:

- TA(0) = # non-alphanumerics
- TA(1) = # characters in set {b d f h k l t 6}
- TA(2) = # characters in set {g j p q y z 3 9}
- TA(3) = # characters in set {a c e I m n o r s u v w x 0 1 2 4 5 7 8}

NOTE: Upper and lowercase letters are interchangeable.

Do a similar tally, TB(j), (j = 0, 1, 2, 3), for all characters in the truth value field.

Let

- NA = TA(0) + TA(1) + TA(2) + TA(3)

- NB = TB(0) + TB(1) + TB(2) + TB(3)

- DIFF = ABS(TA(0)-TB(0)) + ABS(TA(1)-TB(1)) + ABS(TA(2)-TB(2)) + ABS(TA(3)- TB(3)), where ABS is the absolute value function.

Define DIFFALL(k) as

- 0 if $k \leq 5$,
- 1 if $6 \leq k \leq 12$,
- 2 if $13 \leq k \leq 21$, and
- 3 if $22 \leq k \leq 32$.

Then a soft match error occurs when

- the maximum of NA and NB > 0 and

- DIFF > DIFFALL(the minimum of NA and NB).

Appendix H: Distribution of Form Type, Form Name, and Person Number in Table 8

We analyze the distribution of form type, form name, and person number through contingency tables. Our first step is to compare the distribution of short and long form types in Table 8 versus the same distribution in the entire group of 2,996 error rates discussed in section 4.4.5.

Table H1. Distribution of Short and Long Form Types in Table 8 and In Entire Group of 2,996 Error Rates

Form Type	Number in Entire Group of 2,996 Error Rates	Number in Table 8
Long	2,460	162
Short	536	22

The table we would expect if the distributions were perfectly equal is below.

Table H2. Expected Distribution of Short and Long Form Types in Table 8 and In Entire Group of 2,996 Error Rates

Form Type	Expected Number in Entire Group of 2,996 Error Rates	Expected Number in Table 8
Long	2,470	152
Short	526	32

We compute the expected values by the formula from contingency table analysis. If a contingency table is of dimension r rows and c columns, the expected value for the ij-th cell is (Total for row I x Total for column j) / Total of all values in the table.

To test for statistical equality between the distributions of the Table 8 figures and the ones for all 2,996 error rates, we generate the chi square components for each cell in the table. For an r x c contingency table, the chi square component for cell ij is (Actual value - Expected value)^2 / Expected value. The chi square components we need are below.

Table H3. Chi Square Components for Short and Long Form Types in Table 8 and In Entire Group of 2,996 Error Rates

Form Type	Chi Square Component for Number in Entire Group of 2,996 Error Rates	Chi Square Component for Number in Table 8
Long	0.043	0.697
Short	0.201	3.278

After carrying more decimal places than we show in Table H3, the sum of the chi square components is 4.219. To test at the 10 percent level of significance whether the distributions are equal, we compare the sum of our chi square components with the upper ten percent tail value of a chi square distribution with the proper number of degrees of freedom.

The proper degrees of freedom for an r x c contingency table is (r - 1) x (c - 1). For Table H3, the degrees of freedom is (2 - 1) x (2 - 1) or 1. The upper ten percent tail value for a chi square distribution with one degree of freedom is 2.706. Since 4.219 exceeds this, we have evidence the two distributions are not the same. The largest chi square component is generated in the cell for the short form count in Table 8. Comparing the actual value of 22 with the expected value of 32, we conclude the short form error rates are disproportionately underrepresented in Table 8.

We use the same procedure for our second step. Here we compare the distribution of form names in Table 8 with their distribution in the entire group of 2,996 error rates. The three tables we need follow.

Table H4. Distribution of Short and Long Form Names in Table 8 and In Entire Group of 2,996 Error Rates

Form Name	Number in Table 8	Number in Entire Group of 2,996 Error Rates
d1	1	117
d1e	10	151
d1s	10	117
d1u	1	121
d2	69	666
d2e	51	621
d2u	24	671
d2ur	18	447

Table H5. Expected Distribution of Short and Long Form Names in Table 8 and In Entire Group of 2,996 Error Rates

Form Name	Expected Number in Table 8	Expected Number in Entire Group of 2,996 Error Rate
d1	7.015	110.985
d1e	9.572	151.429
d1s	7.550	119.450
d1u	7.253	114.747
d2	43.696	691.304
d2e	39.951	632.049
d2u	41.318	653.682
d2ur	27.645	437.355

Table H6. Chi Square Components for Short and Long Form Names in Table 8 and In Entire Group of 2,996 Error Rates

Form Name	Chi Square Component for Number in Table 8	Chi Square Component for Number in Entire Group of 2,996 Error Rates
d1	5.158	0.326
d1e	0.019	0.001
d1s	0.795	0.050
d1u	5.391	0.341
d2	14.653	0.926
d2e	3.056	0.193
d2u	7.259	0.459
d2ur	3.365	0.213

After carrying more decimal places than we show in Table H6, the sum of the chi square components is 42.204. For Table H6, the degrees of freedom is (8 - 1) x (2 - 1) or 7. The upper 10 percent tail value for a chi square distribution with seven degrees of freedom is 12.017. Since 42.204 exceeds this, the two distributions are not the same. The largest chi square components are generated in the cells for d1, d1u, d2, and d2u counts in Table 8. Comparing the actual values with the expected values, we see form d2 has a disproportionately greater presence in Table 8. The other three have disproportionately less. The most natural form to investigate further is d2.

For our third and last step, we compare the distribution of person number in Table 8 with its distribution in the entire group of 2,996 error rates. The three tables we need follow.

Table H7. Distribution of Person Number in Table 8 and In Entire Group of 2,996 Error Rates

Person Number	Number in Table 8	Number in Entire Group of 2,996 Error Rates
0	18	155
1	47	664
2	32	461
3	29	451
4	18	438
5	23	437
6	17	293

Table H8. Expected Distribution Person Number in Table 8 and In Entire Group of 2,996 Error Rates

Person Number	Expected Number in Table 8	Expected Number in Entire Group of 2,996 Error
0	10.325	162.675
1	42.434	668.566
2	29.423	463.577
3	28.647	451.353
4	27.215	428.785
5	27.454	432.546
6	18.501	291.499

Table H9. Chi Square Components for Person Number in Table 8 and In Entire Group of 2,996 Error Rates

Person Number	Chi Square Component for Number in Table 8	Chi Square Component for Number in Entire Group of 2,996 Error Rates
0	5.705	0.362
1	0.491	0.031
2	0.226	0.014
3	0.004	0.000
4	3.120	0.198
5	0.723	0.046
6	0.122	0.008

The sum of the chi square components is 11.051. The degrees of freedom is six. The upper 10 percent tail value for a chi square distribution with six degrees of freedom is 10.645. Since 11.051 exceeds 10.645, the two distributions are not the same. The largest chi square component is generated for person number 0 in Table 8. Comparing the actual with the expected values, we see person number 0 has a disproportionately greater presence there. Comparing the three steps, the most logical thing to investigate is the disproportionately greater presence of outliers on form d2.

Appendix I: Field Category Nonblank Misinterpretation Rates By Reason

In this appendix, we show by field category the nonblank error rates for each combination of error type and error reason. The rates are for errors in determining the most likely intent of the respondent. The intent of the respondent was defined by the judgement of analysts examining and comparing the contents of fields captured by both the automated and technology and by independent keying. We discuss the limits of this procedure in section 3.4. The outliers shown in Table I1 are computed according to the procedure in Appendix F.

Table I1. Field Category Nonblank Misinterpretation Rates by Error Type and Error Reason

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
Coverage	Extra check-box	Poor image	0.088%	High
		Stray mark or spot	0.053%	
		Box is crossed out	0.007%	
		Mark touches another box	0.001%	
		No reason found	0.001%	
Coverage	Missing check-box	No reason found	0.006%	
Coverage	Wrong check-box	Poor image	0.003%	
		Stray mark or spot	0.003%	
		Mark Outside Box	0.001%	
		Mark touches another box	0.001%	
Form Management	Added response	Poor handwriting	0.120%	High
		Rules not followed	0.013%	
		No reason found	0.011%	
		Big X through person	0.003%	
		Response crossed out	0.003%	
		Character goes out field	0.002%	
		Poor image	0.002%	
		Characters too close	0.001%	
		Response written over	0.001%	
Form Management	Blanked response	No reason found	0.012%	
		Response written over	0.005%	
		Poor handwriting	0.004%	
		Rules not followed	0.003%	
		Character goes out field	0.001%	
		Response crossed out	0.001%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
Form Management	Extra characters	Poor handwriting	0.079%	
		No reason found	0.026%	
		Character goes out field	0.003%	
		Rules not followed	0.003%	
		Poor image	0.002%	
		Response crossed out	0.002%	
		Response written over	0.002%	
		Big X through person	0.001%	
		Characters too close	0.001%	
Form Management	Extra check-box	Stray mark or spot	0.211%	Very High
		No reason found	0.131%	High
		Poor image	0.093%	High
		Box is crossed out	0.009%	
		Mark touches another box	0.005%	
		Big X through person	0.004%	
		Mark Outside Box	0.003%	
Form Management	Missing characters	No reason found	0.289%	Very High
		Poor handwriting	0.053%	
		Characters too close	0.015%	
		Character goes out field	0.014%	
		Response written over	0.003%	
		Truncated	0.003%	
		Mixed upper case & lower case	0.002%	
		Poor image	0.002%	
		Rules not followed	0.002%	
		Decimal point	0.001%	
		Response crossed out	0.001%	
Form Management	Missing check-box	No reason found	0.012%	
		Box is crossed out	0.011%	
		Poor image	0.011%	
		Stray mark or spot	0.002%	
		Mark Outside Box	0.001%	
Form Management	Position reversed	Response written over	0.006%	
		Poor handwriting	0.003%	
		No reason found	0.002%	
		Character goes out field	0.001%	
		Characters too close	0.001%	
		Rules not followed	0.001%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
Form Management	Wrong character	Poor handwriting	6.127%	Very High
		Rules not followed	0.647%	Very High
		No reason found	0.287%	Very High
		Response written over	0.050%	
		Character goes out field	0.027%	
		Characters too close	0.024%	
		Poor image	0.019%	
		Mixed upper case & lower case	0.015%	
		Response crossed out	0.002%	
		Big X through person	0.001%	
		Spanish accents	0.001%	
		Truncated	0.001%	
Form Management	Wrong check-box	No reason found	0.004%	
		Stray mark or spot	0.004%	
		Box is crossed out	0.002%	
		Mark touches another box	0.002%	
		Poor image	0.002%	
Housing Profile	Added response	Rules not followed	0.151%	High
		Response crossed out	0.040%	
		Poor handwriting	0.027%	
		Poor image	0.024%	
		Character goes out field	0.022%	
		Big X through person	0.016%	
		No reason found	0.006%	
		Decimal point	0.004%	
		Response written over	0.002%	
Housing Profile	Blanked response	No reason found	0.069%	
		Response crossed out	0.039%	
		Rules not followed	0.031%	
		Character goes out field	0.022%	
		Response written over	0.016%	
		Poor handwriting	0.011%	
		Poor image	0.010%	
		Truncated	0.007%	
Housing Profile	Extra characters	Decimal point	0.069%	
		No reason found	0.045%	
		Response crossed out	0.038%	
		Response written over	0.020%	
		Rules not followed	0.020%	
		Poor handwriting	0.016%	
		Character goes out field	0.007%	
		Poor image	0.006%	
		Big X through person	0.004%	
		Mixed upper case & lower case	0.004%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
Housing Profile	Extra check-box	Poor image	0.170%	Very High
		Stray mark or spot	0.163%	Very High
		Box is crossed out	0.138%	High
		Big X through person	0.049%	
		Mark touches another box	0.014%	
		No reason found	0.013%	
		Mark Outside Box	0.002%	
Housing Profile	Missing characters	No reason found	0.239%	Very High
		Poor image	0.091%	High
		Rules not followed	0.076%	
		Response written over	0.064%	
		Mixed upper case & lower case	0.061%	
		Character goes out field	0.045%	
		Poor handwriting	0.027%	
		Truncated	0.024%	
		Response crossed out	0.019%	
		Decimal point	0.009%	
		Big X through person	0.005%	
		Characters too close	0.005%	
		No reason found	0.026%	
		Mark Outside Box	0.002%	
		Stray mark or spot	0.002%	
		Big X through person	0.001%	
		Box is crossed out	0.001%	
		Mark touches another box	0.001%	
		Poor image	0.001%	
Housing Profile	Position reversed	No reason found	0.045%	
		Poor handwriting	0.008%	
		Response written over	0.002%	
		Rules not followed	0.001%	
Housing Profile	Wrong character	Poor handwriting	0.637%	Very High
		Spanish accents	0.196%	Very High
		Mixed upper case & lower case	0.110%	High
		Rules not followed	0.092%	High
		Response written over	0.078%	
		No reason found	0.065%	
		Poor image	0.018%	
		Characters too close	0.010%	
		Response crossed out	0.010%	
		Character goes out field	0.009%	
		Decimal point	0.006%	
		Truncated	0.005%	
		Big X through person	0.003%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
Housing Profile	Wrong check-box	Box is crossed out	0.029%	
		Mark touches another box	0.011%	
		Stray mark or spot	0.010%	
		No reason found	0.008%	
		Poor image	0.003%	
		Mark Outside Box	0.002%	
		Big X through person	0.001%	
POP--Demographic	Added response	Spanish accents	0.923%	Very High
		Big X through person	0.021%	
		Rules not followed	0.021%	
		Poor handwriting	0.014%	
		Response crossed out	0.010%	
		Response written over	0.009%	
		Mixed upper case & lower case	0.004%	
		No reason found	0.004%	
		Poor image	0.004%	
		Character goes out field	0.003%	
POP--Demographic	Blanked response	No reason found	0.038%	
		Response crossed out	0.026%	
		Mixed upper case & lower case	0.022%	
		Response written over	0.019%	
		Character goes out field	0.016%	
		Poor image	0.016%	
		Spanish accents	0.016%	
		Poor handwriting	0.015%	
		Truncated	0.013%	
		Decimal point	0.011%	
		Rules not followed	0.011%	
		Characters too close	0.005%	
		Big X through person	0.002%	
POP--Demographic	Extra characters	Spanish accents	1.010%	Very High
		No reason found	0.073%	
		Decimal point	0.023%	
		Rules not followed	0.021%	
		Response crossed out	0.011%	
		Poor handwriting	0.009%	
		Response written over	0.008%	
		Characters too close	0.007%	
		Big X through person	0.006%	
		Character goes out field	0.004%	
		Mixed upper case & lower case	0.004%	
		Poor image	0.002%	
		Truncated	0.001%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Demographic	Extra check-box	Poor image	0.171%	Very High
		Box is crossed out	0.093%	High
		Stray mark or spot	0.086%	High
		Big X through person	0.071%	
		No reason found	0.021%	
		Mark touches another box	0.013%	
		Mark Outside Box	0.002%	
POP--Demographic	Missing characters	No reason found	0.194%	Very High
		Rules not followed	0.193%	Very High
		Spanish accents	0.065%	
		Character goes out field	0.057%	
		Truncated	0.038%	
		Poor handwriting	0.023%	
		Response written over	0.017%	
		Big X through person	0.011%	
		Response crossed out	0.009%	
		Characters too close	0.007%	
		Decimal point	0.006%	
		Mixed upper case & lower case	0.005%	
		Poor image	0.003%	
POP--Demographic	Missing check-box	No reason found	0.024%	
		Poor image	0.003%	
		Big X through person	0.002%	
		Box is crossed out	0.002%	
		Mark Outside Box	0.002%	
		Mark touches another box	0.002%	
		Stray mark or spot	0.001%	
POP--Demographic	Position reversed	No reason found	0.056%	
		Spanish accents	0.036%	
		Mixed upper case & lower case	0.009%	
		Response written over	0.009%	
		Truncated	0.008%	
		Poor handwriting	0.006%	
		Rules not followed	0.005%	
		Response crossed out	0.004%	
		Character goes out field	0.002%	
		Poor image	0.001%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Demographic	Wrong character	Poor handwriting	0.550%	Very High
		Spanish accents	0.265%	Very High
		Mixed upper case & lower case	0.070%	
		No reason found	0.070%	
		Rules not followed	0.058%	
		Decimal point	0.054%	
		Response written over	0.044%	
		Character goes out field	0.025%	
		Poor image	0.010%	
		Characters too close	0.006%	
		Response crossed out	0.005%	
		Big X through person	0.003%	
		Truncated	0.003%	
POP--Demographic	Wrong check-box	Box is crossed out	0.033%	
		Mark touches another box	0.013%	
		Stray mark or spot	0.012%	
		No reason found	0.008%	
		Mark Outside Box	0.004%	
		Poor image	0.004%	
		Big X through person	0.002%	
POP--Disability	Extra check-box	Box is crossed out	0.149%	High
		Poor image	0.147%	High
		Stray mark or spot	0.145%	High
		Big X through person	0.078%	
		No reason found	0.038%	
		Mark touches another box	0.003%	
		Mark Outside Box	0.002%	
POP--Disability	Missing check-box	No reason found	0.007%	
		Box is crossed out	0.002%	
		Mark Outside Box	0.002%	
		Stray mark or spot	0.002%	
		Mark touches another box	0.001%	
POP--Disability	Wrong check-box	Box is crossed out	0.021%	
		Big X through person	0.007%	
		Mark touches another box	0.006%	
		No reason found	0.006%	
		Stray mark or spot	0.006%	
		Poor image	0.003%	
		Mark Outside Box	0.002%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Education	Extra check-box	Poor image	0.450%	Very High
		Box is crossed out	0.303%	Very High
		Stray mark or spot	0.191%	Very High
		Big X through person	0.078%	
		No reason found	0.026%	
		Mark touches another box	0.016%	
		Mark Outside Box	0.003%	
POP--Education	Missing check-box	No reason found	0.110%	High
		Stray mark or spot	0.005%	
		Box is crossed out	0.002%	
		Mark Outside Box	0.002%	
		Mark touches another box	0.002%	
		Poor image	0.002%	
POP--Education	Wrong check-box	Box is crossed out	0.046%	
		Mark touches another box	0.013%	
		Stray mark or spot	0.013%	
		No reason found	0.007%	
		Poor image	0.003%	
		Mark Outside Box	0.002%	
POP--Ethnic	Added response	Response crossed out	0.395%	Very High
		Spanish accents	0.106%	High
		Poor handwriting	0.093%	High
		Poor image	0.079%	
		Rules not followed	0.073%	
		Response written over	0.044%	
		Characters too close	0.043%	
		Big X through person	0.032%	
		No reason found	0.026%	
		Character goes out field	0.004%	
POP--Ethnic	Blanked response	No reason found	0.074%	
		Poor handwriting	0.023%	
		Response crossed out	0.023%	
		Rules not followed	0.010%	
		Character goes out field	0.006%	
		Poor image	0.005%	
POP--Ethnic	Extra characters	Rules not followed	0.281%	Very High
		No reason found	0.253%	Very High
		Response crossed out	0.052%	
		Poor handwriting	0.038%	
		Mixed upper case & lower case	0.034%	
		Poor image	0.020%	
		Character goes out field	0.014%	
		Big X through person	0.012%	
		Response written over	0.008%	
		Truncated	0.008%	
		Characters too close	0.004%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Ethnic	Extra check-box	Mark touches another box	0.071%	
		Big X through person	0.064%	
		Box is crossed out	0.054%	
		Poor image	0.036%	
		Stray mark or spot	0.030%	
		No reason found	0.006%	
		Mark Outside Box	0.001%	
POP--Ethnic	Missing characters	No reason found	1.422%	Very High
		Truncated	0.144%	High
		Character goes out field	0.085%	High
		Poor handwriting	0.079%	
		Characters too close	0.033%	
		Rules not followed	0.022%	
		Response written over	0.020%	
		Mixed upper case & lower case	0.015%	
		Spanish accents	0.014%	
		Poor image	0.007%	
		Response crossed out	0.005%	
POP--Ethnic	Missing check-box	No reason found	0.050%	
		Big X through person	0.011%	
		Stray mark or spot	0.011%	
		Mark touches another box	0.006%	
		Mark Outside Box	0.004%	
		Box is crossed out	0.001%	
POP--Ethnic	Position reversed	Spanish accents	0.654%	Very High
		No reason found	0.181%	Very High
		Response crossed out	0.023%	
		Poor handwriting	0.011%	
		Rules not followed	0.008%	
		Mixed upper case & lower case	0.007%	
		Character goes out field	0.005%	
		Response written over	0.005%	
		Characters too close	0.002%	
POP--Ethnic	Wrong character	Poor handwriting	1.157%	Very High
		No reason found	0.198%	Very High
		Spanish accents	0.154%	High
		Big X through person	0.071%	
		Mixed upper case & lower case	0.071%	
		Truncated	0.061%	
		Response written over	0.034%	
		Decimal point	0.027%	
		Rules not followed	0.026%	
		Characters too close	0.022%	
		Response crossed out	0.018%	
		Character goes out field	0.011%	
		Poor image	0.009%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Ethnic	Wrong check-box	Big X through person	0.032%	
		Mark touches another box	0.005%	
		Box is crossed out	0.004%	
		No reason found	0.003%	
		Stray mark or spot	0.003%	
		Mark Outside Box	0.002%	
		Poor image	0.002%	
POP--Income	Added response	Rules not followed	0.858%	Very High
		Response crossed out	0.147%	High
		Poor handwriting	0.085%	High
		Big X through person	0.063%	
		Response written over	0.047%	
		Characters too close	0.039%	
		Poor image	0.025%	
		No reason found	0.017%	
		Character goes out field	0.006%	
		Truncated	0.003%	
POP--Income	Blanked response	No reason found	0.156%	High
		Rules not followed	0.040%	
		Big X through person	0.027%	
		Response crossed out	0.027%	
		Truncated	0.020%	
		Poor image	0.016%	
		Character goes out field	0.010%	
		Poor handwriting	0.009%	
		Response written over	0.007%	
POP--Income	Extra characters	Decimal point	0.083%	
		No reason found	0.046%	
		Poor handwriting	0.036%	
		Response crossed out	0.031%	
		Rules not followed	0.031%	
		Poor image	0.024%	
		Big X through person	0.018%	
		Response written over	0.009%	
		Mixed upper case & lower case	0.005%	
		Character goes out field	0.003%	
		Spanish accents	0.001%	
POP--Income	Extra check-box	Box is crossed out	0.195%	Very High
		Stray mark or spot	0.146%	High
		Poor image	0.144%	High
		Big X through person	0.069%	
		No reason found	0.049%	
		Mark touches another box	0.008%	
		Mark Outside Box	0.005%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Income	Missing characters	No reason found	0.360%	Very High
		Response written over	0.121%	High
		Character goes out field	0.040%	
		Poor handwriting	0.038%	
		Rules not followed	0.023%	
		Poor image	0.018%	
		Decimal point	0.017%	
		Response crossed out	0.008%	
		Truncated	0.007%	
		Characters too close	0.004%	
POP--Income	Missing check-box	No reason found	0.010%	
		Poor image	0.003%	
		Big X through person	0.002%	
		Box is crossed out	0.002%	
		Stray mark or spot	0.002%	
		Mark Outside Box	0.001%	
POP--Income	Position reversed	Poor handwriting	0.040%	
		No reason found	0.022%	
		Character goes out field	0.017%	
		Rules not followed	0.009%	
		Response written over	0.003%	
POP--Income	Wrong character	Poor handwriting	0.753%	Very High
		Rules not followed	0.318%	Very High
		Response written over	0.167%	Very High
		No reason found	0.098%	High
		Big X through person	0.043%	
		Character goes out field	0.019%	
		Characters too close	0.015%	
		Response crossed out	0.014%	
		Decimal point	0.010%	
		Poor image	0.006%	
		Mixed upper case & lower case	0.005%	
		Truncated	0.003%	
POP--Income	Wrong check-box	Box is crossed out	0.031%	
		Stray mark or spot	0.007%	
		No reason found	0.006%	
		Mark Outside Box	0.002%	
		Mark touches another box	0.002%	
		Poor image	0.002%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Military	Extra check-box	Poor image	0.889%	Very High
		Stray mark or spot	0.223%	Very High
		Big X through person	0.145%	High
		Box is crossed out	0.138%	High
		No reason found	0.042%	
		Mark touches another box	0.016%	
		Mark Outside Box	0.005%	
POP--Military	Missing check-box	No reason found	0.224%	Very High
		Poor image	0.018%	
		Box is crossed out	0.009%	
		Stray mark or spot	0.006%	
		Mark Outside Box	0.004%	
POP--Military	Wrong check-box	Box is crossed out	0.029%	
		Stray mark or spot	0.014%	
		Mark touches another box	0.005%	
		No reason found	0.004%	
		Mark Outside Box	0.003%	
		Poor image	0.002%	
		Big X through person	0.001%	
POP--Name	Added response	Spanish accents	0.016%	
		Big X through person	0.015%	
		Poor handwriting	0.014%	
		Response crossed out	0.014%	
		Characters too close	0.010%	
		Character goes out field	0.007%	
		Poor image	0.006%	
		Rules not followed	0.006%	
		No reason found	0.003%	
		Response written over	0.003%	
		Mixed upper case & lower case	0.002%	
		Truncated	0.002%	
POP--Name	Blanked response	No reason found	0.063%	
		Poor handwriting	0.013%	
		Character goes out field	0.011%	
		Poor image	0.009%	
		Response crossed out	0.009%	
		Response written over	0.009%	
		Rules not followed	0.007%	
		Truncated	0.005%	
		Big X through person	0.004%	
		Mixed upper case & lower case	0.002%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Name	Extra characters	No reason found	0.137%	High
		Poor handwriting	0.034%	
		Poor image	0.028%	
		Rules not followed	0.016%	
		Response crossed out	0.014%	
		Mixed upper case & lower case	0.007%	
		Big X through person	0.006%	
		Character goes out field	0.005%	
		Response written over	0.005%	
		Truncated	0.004%	
		Characters too close	0.003%	
		Spanish accents	0.002%	
POP--Name	Missing characters	No reason found	0.340%	Very High
		Truncated	0.102%	High
		Poor handwriting	0.066%	
		Rules not followed	0.065%	
		Character goes out field	0.016%	
		Characters too close	0.014%	
		Response written over	0.011%	
		Mixed upper case & lower case	0.009%	
		Spanish accents	0.009%	
		Poor image	0.008%	
		Big X through person	0.007%	
		Response crossed out	0.004%	
POP--Name	Position reversed	No reason found	0.062%	
		Mixed upper case & lower case	0.007%	
		Poor handwriting	0.006%	
		Response written over	0.005%	
		Characters too close	0.003%	
		Poor image	0.003%	
		Rules not followed	0.003%	
		Character goes out field	0.002%	
		Truncated	0.002%	
POP--Name	Wrong character	Poor handwriting	1.848%	Very High
		No reason found	0.228%	Very High
		Mixed upper case & lower case	0.124%	High
		Spanish accents	0.073%	
		Poor image	0.062%	
		Response written over	0.032%	
		Character goes out field	0.028%	
		Characters too close	0.017%	
		Rules not followed	0.009%	
		Truncated	0.007%	
		Big X through person	0.004%	
		Response crossed out	0.004%	
		Decimal point	0.001%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Occupation	Added response	Poor image	0.029%	
		Rules not followed	0.028%	
		Big X through person	0.020%	
		Poor handwriting	0.012%	
		Response crossed out	0.011%	
		Response written over	0.010%	
		No reason found	0.008%	
POP--Occupation	Blanked response	No reason found	0.074%	
		Poor handwriting	0.013%	
		Poor image	0.012%	
		Response crossed out	0.011%	
		Rules not followed	0.011%	
		Big X through person	0.010%	
		Response written over	0.008%	
		Character goes out field	0.004%	
		Truncated	0.001%	
POP--Occupation	Extra characters	No reason found	0.328%	Very High
		Rules not followed	0.100%	High
		Poor handwriting	0.024%	
		Spanish accents	0.023%	
		Response crossed out	0.017%	
		Character goes out field	0.012%	
		Decimal point	0.007%	
		Big X through person	0.006%	
		Response written over	0.005%	
		Characters too close	0.004%	
		Truncated	0.004%	
		Mixed upper case & lower case	0.003%	
		Poor image	0.003%	
POP--Occupation	Extra check-box	Poor image	0.385%	Very High
		Box is crossed out	0.364%	Very High
		Stray mark or spot	0.329%	Very High
		Big X through person	0.194%	Very High
		No reason found	0.052%	
		Mark touches another box	0.018%	
		Mark Outside Box	0.004%	
POP--Occupation	Missing characters	Rules not followed	2.096%	Very High
		No reason found	0.935%	Very High
		Character goes out field	0.166%	Very High
		Truncated	0.128%	High
		Poor handwriting	0.095%	High
		Response written over	0.033%	
		Characters too close	0.024%	
		Poor image	0.008%	
		Response crossed out	0.005%	
		Mixed upper case & lower case	0.004%	
		Decimal point	0.003%	
		Big X through person	0.002%	
		Spanish accents	0.002%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Occupation	Missing check-box	No reason found	0.033%	
		Poor image	0.006%	
		Stray mark or spot	0.003%	
		Big X through person	0.002%	
		Box is crossed out	0.002%	
		Mark Outside Box	0.002%	
		Mark touches another box	0.002%	
POP--Occupation	Position reversed	No reason found	0.170%	Very High
		Poor handwriting	0.011%	
		Poor image	0.006%	
		Rules not followed	0.005%	
		Character goes out field	0.003%	
		Mixed upper case & lower case	0.003%	
		Characters too close	0.002%	
		Response crossed out	0.002%	
		Response written over	0.002%	
		Truncated	0.002%	
POP--Occupation	Wrong character	Poor handwriting	1.303%	Very High
		No reason found	0.188%	Very High
		Rules not followed	0.084%	
		Mixed upper case & lower case	0.082%	
		Response written over	0.052%	
		Spanish accents	0.016%	
		Characters too close	0.012%	
		Character goes out field	0.011%	
		Poor image	0.008%	
		Response crossed out	0.008%	
		Truncated	0.005%	
		Big X through person	0.002%	
		Decimal point	0.002%	
POP--Occupation	Wrong check-box	Box is crossed out	0.036%	
		Mark touches another box	0.013%	
		No reason found	0.009%	
		Stray mark or spot	0.009%	
		Mark Outside Box	0.005%	
		Poor image	0.005%	
		Big X through person	0.004%	
POP--Race	Added response	Response crossed out	1.961%	Very High
		Poor handwriting	0.976%	Very High
		Big X through person	0.228%	Very High
		Rules not followed	0.183%	Very High
		No reason found	0.070%	
		Poor image	0.052%	
		Character goes out field	0.049%	
		Response written over	0.028%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Race	Blanked response	No reason found	0.184%	Very High
		Poor handwriting	0.060%	
		Poor image	0.041%	
		Rules not followed	0.034%	
		Response written over	0.031%	
		Character goes out field	0.029%	
		Response crossed out	0.028%	
POP--Race	Extra characters	Response crossed out	0.404%	Very High
		Rules not followed	0.339%	Very High
		No reason found	0.314%	Very High
		Poor handwriting	0.166%	Very High
		Big X through person	0.080%	
		Characters too close	0.063%	
		Mixed upper case & lower case	0.058%	
		Character goes out field	0.055%	
		Response written over	0.039%	
		Poor image	0.036%	
		Truncated	0.033%	
POP--Race	Extra check-box	Big X through person	0.086%	High
		Box is crossed out	0.051%	
		Stray mark or spot	0.035%	
		Poor image	0.022%	
		Mark touches another box	0.015%	
		No reason found	0.007%	
		Mark Outside Box	0.005%	
POP--Race	Missing characters	No reason found	1.602%	Very High
		Truncated	0.891%	Very High
		Poor handwriting	0.269%	Very High
		Character goes out field	0.228%	Very High
		Characters too close	0.222%	Very High
		Response crossed out	0.056%	
		Rules not followed	0.056%	
		Mixed upper case & lower case	0.055%	
		Spanish accents	0.048%	
		Response written over	0.047%	
		Poor image	0.039%	
POP--Race	Missing check-box	No reason found	0.065%	
		Stray mark or spot	0.040%	
		Mark touches another box	0.026%	
		Poor image	0.023%	
		Box is crossed out	0.004%	
		Big X through person	0.003%	
		Mark Outside Box	0.003%	

Field Category	Manner of Misinterpretation	Reason for Misinterpretation	Nonblank Misinterpretation %	Outlier
POP--Race	Position reversed	No reason found	0.247%	Very High
		Poor image	0.141%	High
		Poor handwriting	0.069%	
		Mixed upper case & lower case	0.052%	
		Truncated	0.029%	
POP--Race	Wrong character	Poor handwriting	3.047%	Very High
		No reason found	0.537%	Very High
		Spanish accents	0.252%	Very High
		Mixed upper case & lower case	0.207%	Very High
		Characters too close	0.161%	Very High
		Response written over	0.129%	High
		Truncated	0.105%	High
		Rules not followed	0.091%	High
		Character goes out field	0.060%	
		Decimal point	0.059%	
		Big X through person	0.047%	
		Response crossed out	0.045%	
		Poor image	0.043%	
POP--Race	Wrong check-box	No reason found	0.008%	
		Mark touches another box	0.005%	
		Box is crossed out	0.003%	
		Mark Outside Box	0.003%	
		Stray mark or spot	0.003%	
Special Housing	Added response	Poor handwriting	0.231%	Very High
		Character goes out field	0.098%	High
		No reason found	0.066%	
		Rules not followed	0.036%	
		Response crossed out	0.031%	
		Poor image	0.015%	
Special Housing	Blanked response	No reason found	0.916%	Very High
		Character goes out field	0.082%	
		Rules not followed	0.067%	
		Poor handwriting	0.027%	
		Poor image	0.018%	
Special Housing	Extra characters	Poor handwriting	0.047%	
		Poor image	0.044%	
		No reason found	0.032%	
		Response crossed out	0.012%	
Special Housing	Missing characters	No reason found	0.104%	High
		Rules not followed	0.101%	High
Special Housing	Wrong character	Poor handwriting	0.135%	High
		Rules not followed	0.070%	
		No reason found	0.048%	
		Character goes out field	0.030%	
		Response crossed out	0.030%	

Appendix J: Further Details on Significance Testing

In this appendix, we cover further details of how we test the factors in the various models for statistical significance. Since they are not needed to support the discussion in the results section, it is more appropriate to discuss them here. There are five questions we anticipate.

J.1 What theory does SAS PROC GLM use to produce the ANOVA tables?

SAS PROC GLM uses linear models theory. To understand this theory, we recommend the Graybill text in the reference list. To understand how SAS PROC GLM implements linear models theory, we recommend the SAS Institute text in the reference list.

J.2 Why are the factors called fixed?

The factors in an ANOVA table may be fixed or random. Fixed means all the possible values of a factor, or some constant subset of values that are particularly relevant, are allowed in the analysis. Random means a randomly chosen subset of the possible values is allowed.

Fixed factors are appropriate when the possible or relevant values are all known and the number of them is considered manageable. When the possible or relevant values are not all known, or exist in an unmanageably large number, random factors are more appropriate.

J.3 What does it mean to say one factor is nested inside another?

The factors in an ANOVA table may be crossed or nested. It depends on whether the values of one factor can exist or be set without first specifying the values of the other. If the values can exist or be set independently, the two factors are said to be crossed if some or all of the possible combinations of their values are included in the analysis. If they cannot exist or be set independently, the factor set last is said to be nested inside the factor set first.

An example of two factors that could be crossed is a person's height and weight. The factors form and field are nested. The field has no meaning without first knowing what the form is. So field is said to be nested inside form.

The crossed and nested factors must be appropriately identified to SAS so PROC GLM produces the correct ANOVA table.

J.4 Why do Type III sum of squares identify if individual factors are significant?

The answer depends on the theory of estimable functions, a concept within the theory of linear models. We recommend the SAS Institute text in the reference list for a discussion of how this concept works in SAS PROC GLM. Broadly speaking, the sums of squares reflect how much of the variation in the response variable can be associated with a factor.
There are four types of estimable functions. These lead to four possible sums of squares. The

differences between the four types depend on two things. One is whether we want to know a variable's net contribution after other factors are accounted for. The other is whether the combinations of the factor values occur in equal numbers in the analysis.

In our analysis, we want to know a factor's contribution without first accounting for any other factor. Also, the factor values occur in unequal numbers of combinations. Given these two conditions, Type III sums of squares are the most appropriate of the four types.

J.5 What exactly is the response variable in the ANOVA table?

The results in an ANOVA table assume the response variable approximates a traditional set of assumptions. In our analysis, we are interested in error rates. The error rates are in the form of percents. Percents do not follow the traditional assumptions.

The traditional assumptions tend to be better met if the percents are converted using the arcsine root transformation. The Hopkins item in the reference list provides details. We applied this transformation to our error rates. The values resulting from the transformation are the response variable in the ANOVA tables.

J.6 What is the way to walk through an ANOVA table?

Study the following two tables. Our example is based on an imaginary experiment to understand what factors affect the finished weight of a loaf of bread. In our experiment, we have tried different combinations of flour, water, oven temperature, and baking time. The results in the ANOVA tables are simulated for purposes of illustration.

Table J6a. Sample ANOVA For Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	7200	400.00	20.00	**0.0002**
Error	7	140	20.00		
Corrected Total	25	7340			

In Table J6a, we are testing whether the combination of flour, water, oven temperature, and baking time as a group have a significant effect on the finished weight of a loaf of bread. The finished weight is the response variable. The flour, water, oven temperature, and baking time are factors. Significant means that when one or more of the factors changes, a real change in the response variable tends to follow. By real, we mean a change too large to be considered a coincidence.

Table J6a has three rows: model, error, and corrected total. As we vary the flour, water, temperature, and time, we create different loaves, each with their own finished weight. If we

write down the finished weights after all the loaves are baked, we will see they will vary from some minimum to some maximum value.

What do the various columns mean? We have just explained the terms under the column labeled source. The column labeled DF stands for degrees of freedom. The degrees of freedom is associated with how many different ways we manipulate the factors in our experiment. The more types of flour, quantities of water, number of baking times, and so on that we use the more the degrees of freedom go up. If we use fewer types of flour, fewer quantities of water, and so on, the degrees of freedom will go down. We prefer more degrees of freedom to fewer because that means we are using a larger, more complex experiment to understand our response variable.

The column labeled sum of squares is designed to measure how much the finished weights vary from lightest to heaviest. The more they vary the higher the sum of squares will be. The calculation of the sums of squares depends on a complex mathematical formula. More details can be found in the Graybill item in the reference list. We do not need to know them here for our purposes.

The column labeled mean square is derived from the DF and sum of squares columns. To obtain the mean square for a row, we divide the sum of squares for that row by its DF or degrees of freedom. Only the rows for model and error will generate a mean square in Table J6a.

Mean square for model row = Sum of squares for model row / Degrees of freedom for model row =

7200 / 18 = 400.00.

Mean square for error row = Sum of squares for error row / Degrees of freedom for error row =

140 / 7 = 20.00.

The column labeled F value is derived from the mean square column. To obtain the F value, we divide the mean square in the model row by the mean square in the error row.

F value = mean square for model row / mean square for error row = 400.00 / 20.00 = 20.00.

The column labeled Pr > F helps us conclude whether changes in the flour, water, temperature, and time leads to a real change in the finished weight. If these factors lead to a real change, the Pr > F column will be close to zero. If the change in the finished weight is just a coincidence, the Pr > F column will be close to one.

There is no universal rule to say how close to zero we have to get before we conclude the change in the finished weight is real. The standard in our evaluation is to conclude the change in our response variable is real if the Pr > F is less than 0.10. In Table J6a, Pr > F is 0.0002. By that standard, we would say that as a group the flour, water, temperature, and time lead to a real change in the finished weight. This agrees with our common sense understanding of how to bake bread. We are now ready to walk through Table J6b. This table is designed to tell us the

individual contribution of flour, water, temperature, and time in affecting the finished weight of our loaves of bread.

Table J6b. Sample ANOVA For Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Flour	3	1000	333.33	16.67	0.0014
Water	4	3000	750.00	37.50	<0.0001
Oven Temperature	3	750	250.00	12.50	0.0034
Baking Time	2	800	400.00	20.00	0.0013
Oven Temperature x Baking Time	6	60	10.00	0.50	0.7917

We see in Table J6b a separate row for each of the four factors. The last row is something we have not discussed yet. The last row measures the interaction of oven temperature and baking time. In ANOVA, the term interaction has a precise mathematical definition. More details are available in the Graybill item already mentioned.

To translate the mathematics into more common terms, we begin with the basic observation that quite often a result requires two or more things to work together. We need heat and oxygen for fire, red and yellow to get orange, ice cream and soda to get a float, and so on. When we experiment, the factors we use can affect the response variable in one of two ways.

There can be an independent effect. That means the factor operates in a certain way regardless of what any of the other factors do. There can be an interaction effect. That means the way one factor operates depends on what some other factor does.

When a row lists two or more factors connected by a times sign, it measures the effect of all the factors interacting together. Table J6b shows only one row for an interaction, and that is all we need to illustrate the concept. In the real world, the rule is to see more than one interaction in a table like J6b.

The column DF, degrees of freedom has the same general meaning as in Table J6a. One aspect that is different is in the row for the interaction. The degrees of freedom for an interaction row is the product of the degrees of freedom for the individual factors.

In the row for oven temperature, we see three degrees of freedom. In the row for baking time, we see two degrees of freedom. So the degrees of freedom for the interaction of oven temperature and baking time is two times three, or six. The column Type III SS stands for Type III sum of squares. We have already explained this concept in the answer to question J.4. The concept of a sum of squares has the same general meaning here as in Table J6a. Since Type III SS is what we use in this evaluation, that is what we have picked for our example. In a real experiment, the sum of squares we use depends on how we design the experiment and whether all the data we planned on are actually available by the time we are done.

The column for mean squares is derived from the Type III SS and DF columns. To obtain the mean square for a row, just as in Table J6a, divide the Type III SS for that row by the degrees of freedom. A quick check will verify this is the case for Table J6b.

Since we are assessing individual factors and interactions, we need a separate F value for each one. To obtain it, we divide the mean square for a row by the mean square in the error row of Table J6a.

F value for flour row = Mean square for flour row / Mean square for error row in Table J6a =
750 / 20 = 37.50.

The remaining rows are easily checked to verify the F values.

The Pr > F column in Table J6b is interpreted the same as the Pr > F column in Table J6a. Using the same standard we applied for Table J6a, we conclude from the baking time x oven temperature row that these two factors do not interact in a way that leads to a real change in the finished weight of the loaf of bread. In other words, the interaction is not significant.
The significance of interactions affects how we plan any follow up experiments. The goal of a follow up experiment would be to understand even better what influences the finished weight of the bread. If an interaction is significant, we normally favor "an all for one" policy for a follow up experiment. That means if we want the follow up experiment to include one of the factors that make up an interaction, we have to include them all.

Since baking time and oven temperature do not interact, we have more freedom to include one but not the other in any future experiment. It is easier to plan follow up experiments when none of the interactions are significant, but in real life that is more the exception than the rule. To keep our example simple, we have allowed no significant interactions. We can focus our attention on the rows of Table J6b that list only the name of a single factor. The Pr > F values for all these rows are less than 0.10. We conclude that each one when manipulated contributes to a real change in the finished weight.

We note that the flour and water have a higher type III sum of squares than the oven temperature or baking time. We interpret this to mean that a change in the type or amount of the ingredients has a greater influence on the finished weight than how we bake the loaf. This again agrees with our common sense understanding. In a real experiment, we are free to make similar interpretations. If we do not understand at least roughly how the factors should affect the response variable, we should consider such interpretations tentative until we can confirm them in follow up experiments.

Appendix K: Significance Testing Including All 27,254 Regional Census Center Error Rates

In this appendix, we test factors for statistical significance in analyzing the nonblank hard and soft match error rates by Census 2000 regional census center. We include all 27,254 RCC error rates. As explained in section 4.7, we excluded 9,071 error rates from the analysis there. Otherwise, it would not have been possible to identify any outlying error rates.

In this section, we distinguish between person and nonperson fields as discussed in section 4.4.1.

Our factors for testing statistical significance are Census 2000 regional census center, form, field, field category, and person number. We regard these factors as fixed. For more details about the significance testing, see Appendix J.

We analyze nonperson fields for statistical significance separately from person fields. For nonperson fields, our model is

- field nested within field category,
- field category nested within form, and
- regional census center crossed with field.

For person fields, our model is

- person number nested within field,
- field nested within field category,
- field category nested within form, and
- regional census center.

We compare the findings of this analysis with the testing for significance discussed in section 4.7.3 and 4.7.4.

The notation and interpretation of the output in this section is that of an ANOVA table. PROC GLM in SAS version 8.2 was used to test for significance. The significance level for testing is 10 percent. Overall significance of all factors in the model may be judged by looking at the "Pr > F" value in the line for "Model." Values less than 0.10 indicate overall significance.

The significance of individual factors may be judged by looking at the "Pr > F" value in the line for each factor in the Type III SS section. Values less than 0.10 indicate an individual factor is significant. Significant results are highlighted in bold faced type under the "Pr > F" column. For a detailed walk through of a sample ANOVA table, see Appendix J.

Table K1a. ANOVA For Nonblank Error Rates For Nonperson Fields, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	770	99175.1843	128.7989	18.74	**<0.0001**
Error	765	5256.8075	6.8716		
Corrected Total	1535	104431.9917			

Table K1b. ANOVA For Nonblank Error Rates For Nonperson Fields, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	11	976.869493	88.806318	12.92	**<0.0001**
Field Category	12	626.705612	52.225468	7.60	**<0.0001**
Field	NA	NA			
RCC	11	322.558557	29.323505	4.27	**<0.0001**
Field*RCC	673	2320.567300	3.448094	0.50	1.0000

For nonperson fields, the largest factor significantly affecting the nonblank error rate is form. There are significant secondary contributions of field category and region. The structure of the data set did not allow SAS to test field for significance. In terms of the significant factors and their relative impact on the nonblank error rate, these results agree with the analysis excluding outliers in section 4.7.3.

Table K2a. ANOVA For Nonblank Error Rates For Person Fields, Overall Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	816	170522.4264	208.9736	12.63	**<0.0001**
Error	24901	412136.1935	16.5510		
Corrected Total	25717	582658.6198			

Table K2b. ANOVA For Nonblank Error Rates For Person Fields, Individual Factors

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Form	10	823.33204	82.33320	4.97	**<0.0001**
Field Category	50	2600.65775	52.01316	3.14	**<0.0001**
Field	NA	NA			
Person Number	NA	NA			
RCC	11	12862.19913	1169.29083	70.65	**<0.0001**

There is an overall significant relationship between the nonblank error rate and the factors included in our model. For person fields, the largest factor significantly affecting the nonblank error rate is regional census center. There are significant secondary contributions of form and field category. The structure of the data set did not allow SAS to test field and person number for significance.

We did not include a test for the interaction of regional census center and field in the person field analysis. Unlike the nonperson analysis, the memory resources available to SAS did not allow enough capacity to test the model with this interaction included.

The results do not agree with the analysis in section 4.7.4, but the same factors are significant. There field category is the largest significant contributor. Form and regional census center are the significant secondary contributors.

Including all 27,254 RCC error rates does not change the conclusions of the nonperson field analysis. The person field analysis disagrees in the relative contributions of the significant factors. It is reassuring that the more comprehensive analysis turns up the same set of significant factors, however. We prefer to follow the analysis in section 4.7.4 in terms of what is the largest significant factor for the person field analysis.

Appendix L: Field Category Nonblank Error Rates by Regional Census Center, Broken Out By Respondent-Returned vs. Enumerator-Returned Forms

In this appendix, we provide a more detailed break out of the field category nonblank error rates within the Census 2000 regional census centers. Within each category, we show the rates for respondent-returned and enumerator-returned forms. Some readers of evaluation K.1.B have requested this more detailed break out to support their own analyses. Partly because of time constraints and partly because of the scope of the study plan for evaluation K.1.B, we have not undertaken any analysis of our own. Some field categories do not appear in this table because they did not exist on both respondent-returned and enumerator-returned forms.

Table L1. Field Category Nonblank Error Rates by Regional Census Center, Broken Out By Respondent-Returned vs. Enumerator-Returned Forms

Region	Field Category	Respondent Nonblank Error Rate	Enumerator Nonblank Error Rate	Respondent Nonblank Record Count	Enumerator Nonblank Record Count
21	Housing Profile	1.641%	1.280%	432,568	203,872
	POP--Demographic	1.022%	1.066%	1,191,787	422,376
	POP--Disability	0.599%	0.887%	251,053	87,997
	POP--Education	1.235%	1.671%	106,535	36,565
	POP--Ethnic	1.607%	0.797%	138,712	48,571
	POP--Income	1.377%	1.019%	305,343	93,991
	POP--Military	1.095%	3.063%	45,656	13,941
	POP--Name	2.445%	4.308%	399,185	181,978
	POP--Occupation	2.186%	2.324%	548,641	189,480
	POP--Race	0.735%	0.575%	108,965	34,613
22	Housing Profile	1.267%	1.446%	219,072	146,464
	POP--Demographic	1.064%	1.086%	758,743	398,503
	POP--Disability	0.687%	0.779%	141,385	80,577
	POP--Education	1.624%	1.753%	58,992	31,823
	POP--Ethnic	2.320%	1.024%	101,693	52,535
	POP--Income	1.623%	0.958%	169,964	83,964
	POP--Military	1.247%	2.692%	24,142	11,739
	POP--Name	3.394%	6.173%	253,013	152,930
	POP--Occupation	2.711%	2.434%	298,127	150,830
	POP--Race	1.852%	0.879%	73,593	39,833
23	Housing Profile	1.333%	1.309%	162,427	151,668
	POP--Demographic	1.009%	1.147%	387,485	240,296
	POP--Disability	0.306%	0.556%	44,392	14,376
	POP--Education	1.226%	1.850%	13,784	17,024
	POP--Ethnic	3.034%	1.564%	39,481	9,080
	POP--Income	2.515%	0.857%	72,514	23,445
	POP--Military	2.556%	1.478%	3,599	5,141
	POP--Name	2.695%	5.675%	209,041	137,853
	POP--Occupation	3.246%	2.714%	215,077	79,999
	POP--Race	0.355%	0.675%	20,581	3,703

Region	Field Category	Respondent Nonblank Error Rate	Enumerator Nonblank Error Rate	Respondent Nonblank Record Count	Enumerator Nonblank Record Count
24	Housing Profile	1.849%	1.096%	103,762	183,019
	POP--Demographic	1.184%	0.978%	364,414	241,735
	POP--Disability	0.354%	0.807%	119,911	9,540
	POP--Education	1.835%	1.288%	17,281	14,987
	POP--Ethnic	3.460%	0.301%	37,201	3,656
	POP--Income	2.404%	0.645%	82,965	15,511
	POP--Military	0.498%	0.610%	16,266	11,151
	POP--Name	2.550%	4.342%	244,845	150,835
	POP--Occupation	2.620%	2.118%	338,412	29,975
	POP--Race	22.541%	0.529%	244	9,262
25	Housing Profile	1.431%	1.496%	499,136	240,003
	POP--Demographic	1.055%	1.108%	1,348,808	526,983
	POP--Disability	0.972%	1.321%	296,510	110,024
	POP--Education	1.804%	2.138%	123,000	45,550
	POP--Ethnic	1.747%	0.626%	159,906	60,047
	POP--Income	1.664%	1.373%	365,204	117,708
	POP--Military	1.515%	4.539%	52,943	17,824
	POP--Name	2.679%	4.412%	473,823	220,653
	POP--Occupation	2.414%	2.452%	673,830	233,572
	POP--Race	1.078%	0.455%	115,148	45,029
26	Housing Profile	1.271%	1.515%	565,027	272,520
	POP--Demographic	0.934%	1.360%	1,415,325	525,051
	POP--Disability	0.684%	0.769%	329,904	113,593
	POP--Education	1.382%	2.342%	134,266	47,531
	POP--Ethnic	1.335%	0.624%	160,542	55,494
	POP--Income	1.487%	1.061%	405,510	121,819
	POP--Military	1.148%	4.201%	59,501	18,947
	POP--Name	2.219%	4.596%	488,242	217,764
	POP--Occupation	2.185%	2.243%	741,434	244,475
	POP--Race	0.841%	0.364%	115,286	40,138
27	Housing Profile	1.412%	1.230%	185,741	159,338
	POP--Demographic	1.154%	1.077%	331,851	225,740
	POP--Disability	0.806%	0.358%	45,127	36,565
	POP--Education	1.394%	2.524%	53,798	18,663
	POP--Ethnic	3.747%	0.424%	29,252	24,769
	POP--Income	1.940%	0.546%	91,793	39,896
	POP--Name	2.831%	5.913%	233,229	115,193
	POP--Occupation	3.497%	2.787%	221,955	51,493
	POP--Race	0.964%	0.496%	21,679	6,854

Region	Field Category	Respondent Nonblank Error Rate	Enumerator Nonblank Error Rate	Respondent Nonblank Record Count	Enumerator Nonblank Record Count
28	Housing Profile	1.222%	1.224%	523,199	325,729
	POP--Demographic	0.928%	1.009%	1,476,372	687,891
	POP--Disability	0.709%	0.705%	281,337	132,016
	POP--Education	1.395%	1.695%	123,457	56,283
	POP--Ethnic	1.082%	0.528%	165,650	75,051
	POP--Income	1.731%	0.956%	351,461	150,684
	POP--Military	1.393%	3.077%	56,068	22,979
	POP--Name	2.323%	3.906%	521,761	288,089
	POP--Occupation	2.131%	1.983%	619,207	280,694
	POP--Race	0.705%	0.351%	131,311	56,359
29	Housing Profile	1.541%	1.221%	46,270	259,348
	POP--Demographic	0.846%	1.141%	111,336	479,861
	POP--Education	0.844%	2.362%	17,899	28,114
	POP--Income	2.588%	0.805%	20,752	54,518
	POP--Name	2.778%	5.115%	104,279	232,850
	POP--Occupation	4.827%	2.507%	51,321	115,324
30	Housing Profile	1.378%	1.344%	436,725	302,517
	POP--Demographic	0.969%	1.072%	1,322,472	700,481
	POP--Disability	0.721%	0.768%	250,336	135,392
	POP--Education	1.474%	1.559%	111,514	59,190
	POP--Ethnic	1.503%	0.555%	155,519	80,297
	POP--Income	1.785%	0.990%	305,814	149,462
	POP--Military	1.307%	2.985%	47,821	21,947
	POP--Name	2.638%	4.310%	462,640	282,908
	POP--Occupation	2.209%	2.073%	528,822	272,636
	POP--Race	1.204%	0.586%	118,625	61,981
31	Housing Profile	1.272%	1.378%	373,876	225,898
	POP--Demographic	0.926%	1.128%	1,067,827	498,593
	POP--Disability	0.699%	0.793%	220,161	97,212
	POP--Education	1.474%	2.321%	89,215	41,268
	POP--Ethnic	1.466%	0.544%	130,457	56,407
	POP--Income	1.440%	0.939%	268,348	107,772
	POP--Military	1.273%	4.034%	39,818	16,162
	POP--Name	2.274%	4.173%	370,324	202,000
	POP--Occupation	2.289%	2.202%	492,417	208,448
	POP--Race	1.131%	0.681%	93,696	45,799
32	Housing Profile	1.747%	1.286%	109,440	137,508
	POP--Demographic	1.269%	1.154%	421,316	174,596
	POP--Disability	0.503%	0.430%	85,706	26,757
	POP--Education	2.136%	1.989%	34,172	27,000
	POP--Ethnic	3.399%	1.269%	39,539	5,912
	POP--Income	2.131%	0.678%	82,548	17,556
	POP--Name	2.986%	6.866%	290,799	105,118
	POP--Occupation	4.201%	2.720%	116,656	33,089
	POP--Race	1.356%	0.934%	76,472	22,810

Appendix M: Glossary of Terms

In this appendix, we gather and define certain terms in this evaluation that are special purpose or frequently used.

Analysis of Variance	See ANOVA.
ANOVA	Short for Analysis of Variance. A statistical technique for determining whether change in a factor or group factors is associated with a real change in a response variable of interest. Also a short hand reference to the table in which the results of the technique for a particular application are shown.
Arcsine root transformation	A transformation recommended for raw data in the form of percents or proportions so that the traditional assumptions of ANOVA are more closely met. The transformation used in this evaluation before analyzing the nonblank error rate with ANOVA. See Appendix J.
Automated data capture	Data capture performed automatically with minimal or no human intervention beyond loading or unloading of the forms during processing.
Automated technology	A system combining some form of automated data capture with some form of image technology.
Capture	(1) To reproduce content (2) To discern intent, exactly or to a reasonable approximation.
Census form	Any of the questionnaires in paper or other media that are used by the Census Bureau to enumerate and characterize the population of the United States.
Check-box field	A field on a census form in which the respondent is forced to select from a standard set of choices. The selection is shown by a "X", check mark, or like symbol.
Chi square	The name of a statistic and a technique used to analyze Table 8 in section 4.4.5. See Appendix H.
Conditional probability	The probability of an event given that some other condition aready exists.

Confidence interval	A interval constructed in such a way that its end points can be expected to bound the true value for some population characteristic some minimum percentage of the time. Time is usually understood to be over some indefinite, long run period.
Content	The string of characters forming a response on a census form.
Context value	The content of a field as captured. In the case of automated data capture, also the content after removal of extraneous characters inserted by the data capture system.
Crossed	One of the possible relationships between two or more factors in an ANOVA. See Appendix J.
Data capture	In general, any method of transferring the responses on a census form to a medium that supports easy retrieval and analysis of the data.
Data Capture Center	See DCC.
Data capture error	Any instance of a hard match error, soft match error, or misinterpretation.
Data capture mode	The ways responses were captured during Census 2000: KFI, OCR, or OMR.
DCC	One of four locations at which responses were captured from Census 2000 forms. For the names of the locations see section 4.6.1.
Degrees of Freedom	See DF.
DF	Short for degrees of freedom. One of the possible components of an ANOVA table. See Appendix J.
Enumerator	An employee of the Census Bureau obtaining household responses to a census form by directly contacting the household.
Error	(1) A hard or soft match error. (2) In an ANOVA table, a row summarizing the impact on the response variable of all factors not included in the model row. See Appendix J.
Error rate	In this evaluation, the percentage of times a given field's or

	group of fields' captured content disagrees excessively with that on the corresponding census forms.
Evaluation file	The file containing the manually keyed responses from all the census forms included in the sample for this evaluation. This keying took place after Census 2000 processing and reproduced the entire content of the questionnaires. It is distinct and independent of any remedial keying that took place during Census 2000 processing after the automated technology rejected the content for a field.
Evaluation truth value	See truth value.
F value	One of the possible components of an ANOVA table. See Appendix J.
Factor	One of the variables manipulated in an experiment to determine its impact on the response variable. The data from such an experiment can be analyzed via ANOVA. As in this evaluation, the manipulation can be in the form of post hoc cross classification of a data set by the variables of interest.
Field	Short for field name. Any single question or request for data on a census form. Also any single part of a multiple part question or data request.
Field category	One of the thirteen groups of related fields constructed for data analysis purposes in this evaluation. A list appears in Appendix B.
Fixed	A way of classifying a factor for ANOVA. See Appendix J.
Form	See census form.
Hard match error	The failure for the content of a check-box field as reproduced in data capture to match the content as it exists on the census form.
Imaging technology	Collectively all the technical means of high speed electronic reproduction of census responses originally recorded on a physical medium such as paper.

Intent	The content of a field as the respondent or enumerator meant to put it on the form.
Intent of the respondent	See intent.
Interaction	A way two or more factors can affect a response variable. See Appendix J.
Key From Image	See KFI
Key From Paper	See KFP
KFI	Short for Key From Image. The manual keying of the responses to a census form using an electronic reproduction of the original.
KFP	Short for Key From Paper The manual keying of the responses to a census form using the original paper form.
Long form	Any of the census forms which record the information asked on the short form and in addition ask additional questions relating to education, income, occupation, housing characteristics, and similar socioeconomic characteristics of the household. A list of the long forms used in this evaluation appears in Appendix A.
Mailout/mailback	Any census form mailed to and mailed back by the people in the household providing the responses.
Manner of misinterpretation	The various ways in which a data capture process may not capture what the respondent or enumerator meant to say. This includes ways that are caused by an action or omission of the respondent or enumerator. They are described in Tables 43 and 45 of section 4.11.4.
Mean square	One of the possible components of an ANOVA table. See Appendix J.
KFI	The manual keying of responses that are rejected by the automated data capture and imaging technology. This keying takes place during census processing and is distinct from the keying used to create the evaluation file for our report.

KFI impact	The impact of KFI on the ability to correctly capture what the respondent or enumerator meant to put on a form. For an explanation of the possible impacts, see Table 27 in section 4.8.1.
KFI redundancy	A case of sending content to KFI unnecessarily. For an explanation of the different ways this can happen, see Table 27 in section 4.8.1.
KFI redundancy rate	The percentage of times a field or group of fields is sent to KFI unnecessarily.
Misinterpretation	A failure to capture what the respondent or enumerator meant to indicate. If the respondent or enumerator recorded something other than what they meant, say for example by a misspelling, it is still a misinterpretation if the content recorded on the form is accurately captured. In this evaluation, we relied on clerical evaluators using predefined rules to judge the intent of the respondent.
Misinterpretation rate	In this evaluation, the percentage of a field or group of fields whose content does not reflect the intent of the respondent or enumerator.
Model	In an ANOVA table, a row summarizing the collective impact of a group of factors on the response variable. See Appendix J.
Nested	One of the possible relationships between two or more factors in an ANOVA. See Appendix J.
Nonblank error rate	An error rate whose numerator is the number of times nonblank content was captured with a soft or hard match error. The denominator is the number of times nonblank content was captured. Generally calculated on a field or field category basis.
Nonparametric	Statistical estimation, modeling, analysis, etc. without assuming the data follow any particular probability distribution.
OCR	Short for Optical Character Recognition. The automated electronic capture of the content of a write-in field on a census form.
OMR	Short for Optical Mark Recognition. The automated

	electronic capture of the content of a check-box field on a census form.
Optical Character Recognition	See OCR.
Optical Mark Recognition	See OMR.
Outlier	A data value not typical of the others in a data set. Generally values for a data set that are much smaller or larger than usually expected. See Appendix F for how we calculate outliers in this evaluation.
Person Number	A number to indicate which person in a household a particular response is for. On census forms, the responses for separate persons are grouped into sections labeled Person 1, Person 2, and so on.
Pr > F	One of the possible components of an ANOVA table. See Appendix J.
Random	A way of classifying a factor for ANOVA. See Appendix J.
RCC	See Regional Census Center
Reason for misinterpretation	The reasons why a particular manner of misinterpretation takes place. They are described in Tables 44 and 46 of section 4.11.4.
Regional Census Center	One of the twelve offices one level below Suitland, MD, headquarters that managed Census 2000. Abbreviated RCC. For the areas covered by the regions, see section 4.1.9.
Response variable	In general, a variable we wish to understand or control. In this evaluation, usually the nonblank error rate as transformed in the manner explained in Appendix J.
SAS	Commercial statistical package used at the Census Bureau, short for Statistical Analysis System.
Short form	Any of the census forms which record only the names, ages, gender, race, and ethnicity for the members of a household. A list of the short forms used for this evaluation appears in Appendix A.

Soft match algorithm	The computer program used in Census 2000 to determine if the content of a write-in field after data capture diverged within acceptable bounds from the way it exists on the census form. See Appendix G for details.
Soft match error	The failure for the content of a write-in field as reproduced in data capture to diverge within acceptable bounds from how it exists on the census form.
Source	One of the possible components of an ANOVA table. See Appendix J.
Statistical Analysis System	See SAS.
Statistically significant	An effect on a response variable that is too large to be a coincidence according to some predefined standard. See Appendix J.
Sum of Squares	One of the possible components of an ANOVA table. See Appendix J.
Total error rate	An error rate in which the numerator is the number of times nonblank content was captured with a soft or hard match error. The denominator is the number of times any content was captured, blank or nonblank. Generally calculated on a field or field category basis.
Truth value	Also called evaluation truth value. The judgement of the clerical evaluators mentioned in section 2.1 as to what the respondent or enumerator meant to put in a field.
Type III SS	One of the possible components of an ANOVA table. See Appendix J.
Update/leave	Any census form left by an employee of the Census Bureau at a household. The household is expected to fill out and mail back the form. If it is necessary to leave a form because the household's address was not in the Census Bureau address files, the employee records the address so these files can be updated.
Write-in field	A field on a census form that permits a free form answer. The response is written, hopefully, but not always, in the space provided on the form.