UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MEMORANDUM FOR Distribution

From: Cynthia Clark
Associate Director for Methodology and Standards

Subject: Quality of the Data Capture System

I am pleased to present the executive summary of one of the evaluation studies for the Census 2000 Dress Rehearsal. The dress rehearsal was conducted in three sites — Columbia, South Carolina; Menominee County, Wisconsin; and Sacramento, California. The evaluation studies cover detailed aspects of eight broad areas related to the census dress rehearsal — census questionnaire, address list, coverage measurement, coverage improvement, promotion activities, procedures addressing multiple options for census reporting, field operations, and technology.

The executive summary for each evaluation study is also available on the Census Bureau Internet site (http://www.census.gov/census2000 and click on the link to "Evaluation"). Copies of the complete report may be obtained by contacting Carnelle Sligh at (301) 457-3525 or by e-mail at carnelle.e.sligh@ccmail.census.gov. Please note that the complete copy of the following reports will not be publically released: reports regarding procedures addressing multiple options for census reporting and the Evaluation of Housing Unit Coverage on the Master Address File.

The evaluations are distributed broadly to promote the open and thorough review of census processes and procedures. The primary purpose of the dress rehearsal is to simulate portions of the environment we anticipate for Census 2000, so we can identify and correct potential problems in the processes. Thus, the purpose of the evaluation studies is to provide analysis to support time critical review and possible refinements of Census 2000 operations and procedures.

The analysis and recommendations in the evaluation study reports are those of staff working on specific evaluations and, thus, do not represent the official position of the Census Bureau. They represent the results of an evaluation of a component of the census plan. They will be used to analyze and improve processes and procedures for Census 2000. The individual evaluation recommendations have not all yet been reviewed for incorporation in the official plan for Census 2000. These evaluation study reports will be used as input to the decision making process to refine the plans for Census 2000.

The Census Bureau will issue a report that synthesizes the recommendations from all the evaluation studies and provides the Census Bureau review of the dress rehearsal operation. This report will also indicate the Census Bureau's official position on the utilization of these results the Census in 2000 operation. This report will be available July 30th.

Quality of the Data Capture System

July 1999
Kevin D. Haley
Decennial Statistical Studies Division

EXECUTIVE SUMMARY

This evaluation reports on the accuracy of the Data Capture System as it existed during the Census 2000 Dress Rehearsal. It focuses on the overall system and its ability to capture the answers the respondents wish to make. It is not intended as a report on the individual components of the system.

The Data Capture System used in the Census 2000 Dress Rehearsal was still in development. As was stated in the Census 2000 Dress Rehearsal Mid-Term Status Report:

> "The Data Capture System 2000 contractor is still developing the software and hardware capabilities for the data entry and scanning system. This is being done with the full knowledge and acceptance of the Census Bureau. The scanning production measures planned from Dress Rehearsal operations are not available due to continual system changes. Load tests and other tests that were unable to be completed earlier are currently planned to assess the system. Data from these tests will be gathered, analyzed and reported at a later date."

The Dress Rehearsal was the first opportunity to test the system with questionnaires filled out by actual respondents. No amount of testing in a laboratory can simulate the diversity that exists in the way the general public will fill out the questionnaires. It is also impossible to simulate all of the problems that will be encountered in the Data Capture operation during Census 2000. The purpose of the Dress Rehearsal was to expose the Data Capture System to this type of an environment so that any areas that needed improvement could be identified. As the following report demonstrates, the Dress Rehearsal was successful in providing data to make improvements to the Data Capture System for Census 2000.

The Data Capture operation for the Census 2000 Dress Rehearsal utilized digital imaging technology to capture responses from the census questionnaires. The image system consisted of scanning the census questionnaires to create image files. Optical Character Recognition software was used to interpret the handwritten responses, and Optical Mark Recognition software was used to interpret the mark responses. The system was designed with a Key From Image component to display responses on a computer screen to a keyer when the Optical Character Recognition software was uncertain of the correct answers. If a questionnaire could not be scanned it was sent to be Keyed From Paper. The system also used a Data Capture Audit and Resolution process on short form mail returns to detect possible population count problems. If the number captured in the coverage question field did not agree with the number of person panels that contained at least two items with responses, the image of the questionnaire was shown to a clerk who reviewed the image and labeled persons as either valid or invalid. Data from persons labeled as invalid were not used during headquarters processing, or included in this evaluation.

There are two questions that this evaluation will answer for each Dress Rehearsal site, by form type, and by question type:

- What percentage of the responses in the unedited Dress Rehearsal response file is different from the actual responses on the census questionnaires (also referred to as field error rates)?

- Are these field error rates significantly different from one another?

The various questions from the Dress Rehearsal questionnaire can be broken down into write-in and mark fields. A field is the space provided for the respondent to provide an answer, or response. For write-in responses the field is the set of segmented boxes, while for mark responses the field is the set of check boxes. For this analysis we looked at write-in and mark response groupings. For example, rather than comparing Person 1 Last Name and Person 2 Last Name, they were both treated as Last Name. A response in the Dress Rehearsal response file was considered an error if a single character was incorrect. For mark responses, the system was not expected to correctly capture a response if a respondent circled a check box or if a respondent circled the answer description next to a check box. Several of the write-in differences were caused by the truncation of a response by the Data Capture System 2000. These truncation differences were not considered errors for this evaluation. The reason the differences appeared was the truth file was allowed to capture more characters than the number of segmented boxes. When calculating the field error rates, the number of times a field was captured incorrectly was divided by the total number of times that field had a response.

This evaluation was designed to assess the Data Capture System as a whole. No information is provided for the different methods of processing the write-in and mark responses (Optical Mark Recognition, Optical Character Recognition, Key From Image, Key From Paper). Due to unexpected problems, the scope of this evaluation is limited to data from the mailout/mail return short form questionnaires. The results from this evaluation cannot be generalized beyond the Census 2000 Dress Rehearsal.

The following results were observed during the course of this evaluation:

- The field error rate for the write-in responses was 3.01 percent. The goal for capturing write-in responses during Census 2000 is to have a field error rate of no more than 2.0 percent, which is consistent with the Bureau's historical expectations for data capture.

- The number of write-in response errors on the Dress Rehearsal response file was reduced by 6.64 percent by the Data Capture Audit and Resolution process. Without the Data Capture Audit and Resolution process the field error rate for write-in responses would have been 3.21 percent.

- Approximately 24 percent of the write-in response errors may have been due to the way the respondent filled out the questionnaire.

- Approximately 6.6 percent of the write-in response errors were from questionnaires that were

checked into the Data Capture System 2000 but had no data on the Dress Rehearsal response file. Either the data were lost after the questionnaires were processed, or the questionnaires were never processed.

- The field error rate for all write-in responses was not significantly different between Sacramento and Columbia. Four of the individual write-in response groupings were significantly different between Sacramento and Columbia:

  < Last Name (5.35 percent Sacramento, 4.23 percent Columbia)
  < American Indian Tribe (13.58 percent Sacramento, 6.98 percent Columbia)
  < Other Race (9.68 percent Sacramento, 22.89 percent Columbia)
  < Area Code (1.64 percent Sacramento, 3.83 percent Columbia)

- Of the errors found for the write-in responses: 63.66 percent had the wrong characters or numbers, 13.79 percent were omitted responses that should have been on the Dress Rehearsal response file, 10.91 percent had characters or numbers omitted, 5.5 percent had characters or numbers added, 1.68 percent were added responses that should not have been on the Dress Rehearsal response file, and 4.46 percent were characters in numeric fields (or vice versa).

- The field error rate for mark responses was 0.81 percent. The results from the Image Data Capture Evaluation from the 1995 Census Test estimated the mark response error rate at 4.2 percent.

- The number of mark response errors on the Dress Rehearsal response file was reduced by 43.80 percent by the Data Capture Audit and Resolution process. Without the Data Capture Audit and Resolution process the field error rate for mark responses would have been 1.42 percent.

- The field error rate for all mark responses was significantly different between Sacramento (0.92 percent) and Columbia (0.74 percent). Only one of the individual mark response groupings, Race, was significantly different between Sacramento (1.51 percent) and Columbia (0.82 percent).

- Approximately 41 percent of the mark response errors may have been due to the way the respondent filled out the questionnaire.

- Approximately 25 percent of the mark response errors were from questionnaires that were checked into the Data Capture System 2000 but had no data on the Dress Rehearsal response file. Either the data was lost after the questionnaires were processed, or the questionnaires were never processed.

- For the cases where a respondent marked more than one Race box, 15.25 percent (se[1] of 1.95) of the responses had at least one mark omitted. For the cases where a respondent marked more than one Hispanic Origin box, 23.19 percent (se of 5.08) of the responses had at least one mark omitted.

- Of the errors found for the mark responses: 21.85 percent were added responses that should not have been on the Dress Rehearsal response file, 52.76 percent were omitted responses that should have been on the Dress Rehearsal response file, and 25.39 percent had the wrong response captured.

The following conclusions with accompanying recommendations, and current status (shown in bold) were drawn from the results of this evaluation:

- There did not appear to be an overall difference in the data capture quality between the Dress Rehearsal sites.

- Many of the errors that were observed were due to the way the respondent filled out the questionnaire. It should not be expected that all responses will be written as intended within the segmented boxes.

  While it may not be reasonable to expect the Data Capture System 2000 to capture every response on a questionnaire, regardless of where it is written, there are several situations that appear to be reasonable to capture:

  < When a respondent makes a mistake and crosses out what is written, and then writes the response immediately above or below the field.
  < When a respondent writes two characters in the same segmented box.
  < When part of the response extends outside of the segmented boxes, or field.

- The use of the exact match criteria for capturing write-in alpha responses may not be necessary to satisfy the current processing needs.

  The definition of an error for write-in alpha responses should be modified to include only significant deviation from what is present on the questionnaire, as long as it does not impact the usage of the data. This would be consistent with historic methods for monitoring the capture of write-in alpha responses. This could have an impact on the ability of the Data Capture System 2000 to meet the Census 2000 write-in field error rate goal.

  This recommendation has been accepted, and will be used by the Data Capture System 2000.

- The write-in response that had the lowest error rate was the coverage question. This response

[1] Standard error.

was not treated the same as all other write-in responses. The coverage question had a content edit applied to try to ensure that the captured responses were complete.

It appears that the use of a content edit on this response had an impact on the quality of the captured responses. Consideration should be taken in sending more response groupings through content edits as a way of improving the data capture quality.

This recommendation has been accepted. More content edits are being applied to fields such as the age and year of birth. Subject matter dictionaries are also being used to help determine if a response captured by the Optical Character Recognition Software is correct.

- Multiple mark responses for the Race and Hispanic Origin questions were not adequately captured. The multiple mark responses that were not captured represented approximately 29 percent of all mark response omission errors.

  Since the methodology used by the Data Capture System 2000 during Dress Rehearsal did not adequately capture the multiple responses for the Race and Hispanic Origin questions, other alternatives to capturing the responses should be investigated.

  This recommendation has been accepted. A method for handling multiple responses to the Race and Hispanic Origin questions is currently under development.

- The quality of the Data Capture System 2000 captured responses varied by the response grouping.

  If the Data Capture System 2000 cannot capture all response groupings at an equal level of quality, the quality requirements for each response grouping should be established based upon the capabilities of the Data Capture System 2000, with agreement from the appropriate customers. The quality assurance system should be refined to monitor the quality for each response grouping. The Data Capture System 2000 should have the ability to identify and correct errors for all methods of data capture (Optical Mark Recognition, Optical Character Recognition, Key From Image, Key From Paper).

- Omission of a response accounted for approximately 14 percent of the write-in response errors and approximately 53 percent of the mark response errors. There are two ways that a response could be omitted: either the system interprets the response as a blank, or the system does not process and store the data on the questionnaires.

  Precautions should be taken when a person exists and a response is interpreted as being blank. If after the first pass through the system a response is thought to be blank, that decision should be validated.

The Data Capture System 2000 must have better controls in place to ensure that questionnaires that are checked into the system have all data captured and stored. Check points should be

established at each step in the system to ensure that all questionnaires that enter a process, leave the process with all of the expected output. Correction of this problem could have an impact of up to a 40 percent reduction in mark response omissions, and a reduction of around 45 percent in write-in response omissions.

Currently a check-out function has been added to the Data Capture System 2000 to ensure that data are captured for all questionnaires that are scanned.

- Several errors occurred because of responses that were written either partially, or completely outside of a response field.

  The boundaries set by Data Capture System 2000 to look outside of a response field may need to be widened so that these types of responses are captured correctly.

- Several of the responses on the Dress Rehearsal response file were truncated by the Data Capture System 2000. This was due to a system requirement on the number of characters that the system could capture for each field.

  The number of characters that the Data Capture System 2000 can capture for each field should be increased. Currently, the number of segmented boxes in a field is the number of characters that the system can capture.

  This recommendation can not be accepted at this time. It is too late to make this alteration to the Data Capture System 2000.

- The Data Capture Audit and Resolution process was effective at identifying and excluding extraneous data from being used during Census processing.

  The Data Capture Audit and Resolution process should be used during Census 2000.

  The Data Capture Audit and Resolution process will be used during Census 2000.

- Several changes have been made to the Data Capture System 2000. In order to determine the impact of these changes, a new evaluation should be performed prior to Census 2000. This evaluation should provide information for each field from the four main questionnaires: short form mailout/mailback, long form mailout/mailback, short form enumerator, and long form enumerator. The Operational Test and Dry Run planned for the Baltimore Data Capture Center would be the next opportunity to evaluate the Data Capture System 2000.

  This recommendation has been accepted. A study plan for the evaluation is under development