Census 2000 Topic Report No.3
Census 2000 Testing, Experimentation, and Evaluation Program

Issued January 2004

TR-3

Census 2000 Data Capture

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

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

United States Census 2000

Acknowledgments

The Census 2000 Evaluations Executive Steering Committee provided oversight for the Census 2000 Testing, Experimentation, and Evaluations (TXE) Program. Members included **Cynthia Z. F. Clark**, Associate Director for Methodology and Standards; **Preston J. Waite**, Associate Director for Decennial Census; **Carol M. Van Horn**, Chief of Staff; **Teresa Angueira**, Chief of the Decennial Management Division; **Robert E. Fay III**, Senior Mathematical Statistician; **Howard R. Hogan**, (former) Chief of the Decennial Statistical Studies Division; **Ruth Ann Killion**, Chief of the Planning, Research and Evaluation Division; **Susan M. Miskura**, (former) Chief of the Decennial Management Division; **Rajendra P. Singh**, Chief of the Decennial Statistical Studies Division; **Elizabeth Ann Martin**, Senior Survey Methodologist; **Alan R. Tupek**, Chief of the Demographic Statistical Methods Division; **Deborah E. Bolton**, Assistant Division Chief for Program Coordination of the Planning, Research and Evaluation Division; **Jon R. Clark**, Assistant Division Chief for Census Design of the Decennial Statistical Studies Division; **David L. Hubble**, (former) Assistant Division Chief for Evaluations of the Planning, Research and Evaluation Division; **Fay F. Nash**, (former) Assistant Division Chief for Statistical Design/Special Census Programs of the Decennial Management Division; **James B. Treat**, Assistant Division Chief for Evaluations of the Planning, Research and Evaluation Division; and **Violeta Vazquez** of the Decennial Management Division.

As an integral part of the Census 2000 TXE Program, the Evaluations Executive Steering Committee chartered a team to develop and administer the Census 2000 Quality Assurance Process for reports. Past and present members of this team include: **Deborah E. Bolton**, Assistant Division Chief for Program Coordination of the Planning, Research and Evaluation Division; **Jon R. Clark**, Assistant Division Chief for Census Design of the Decennial Statistical Studies Division; **David L. Hubble**, (former) Assistant Division Chief for Evaluations and **James B. Treat**, Assistant Division Chief for Evaluations of the Planning, Research and Evaluation Division; **Florence H. Abramson**, **Linda S. Brudvig**, **Jason D. Machowski**, and **Randall J. Neugebauer** of the Planning, Research and Evaluation Division; **Violeta Vazquez** of the Decennial Management Division; and **Frank A. Vitrano** (formerly) of the Planning, Research and Evaluation Division.

The Census 2000 TXE Program was coordinated by the Planning, Research and Evaluation Division: **Ruth Ann Killion**, Division Chief; **Deborah E. Bolton**, Assistant Division Chief; and **Randall J. Neugebauer** and **George Francis Train III**, Staff Group Leaders. **Keith A. Bennett**, **Linda S. Brudvig**, **Kathleen Hays Guevara**, **Christine Louise Hough**, **Jason D. Machowski**, **Monica Parrott Jones**, **Joyce A. Price**, **Tammie M. Shanks**, **Kevin A. Shaw**, **George A. Sledge**, **Mary Ann Sykes**, and **Cassandra H. Thomas** provided coordination support. **Florence H. Abramson** provided editorial review.

This report was prepared under contract by **Donald Kline** of the Titan Systems Corporation. The project manager was **Kevin A. Shaw** of the Planning, Research and Evaluation Division. The following authors and project managers prepared Census 2000 experiments and evaluations that contributed to this report:

Decennial Statistical Studies Division:
Joseph D. Conklin

Planning, Research and Evaluation Division:
Kevin A. Shaw

Independent contractor:
Donald Kline, Titan Systems Corporation

Greg Carroll and **Everett L. Dove** of the Administrative and Customer Services Division, **Walter C. Odom**, Chief, provided publications and printing management, graphics design and composition, and editorial review for print and electronic media. General direction and production management were provided by **James R. Clark**, Assistant Division Chief, and **Susan L. Rappa**, Chief, Publications Services Branch.

Census 2000 Topic Report No. 3

Census 2000 Testing, Experimentation, and Evaluation Program

Issued January 2004

TR-3

CENSUS 2000 DATA CAPTURE

U.S. Department of Commerce
Donald L. Evans,
Secretary

Samuel W. Bodman,
Deputy Secretary

Economics and Statistics Administration
Kathleen B. Cooper,
Under Secretary for Economic Affairs

U.S. CENSUS BUREAU
Charles Louis Kincannon,
Director

Suggested Citation

Donald Kline
Census 2000 Testing, Experimentation, and Evaluation Program Topic Report No. 3, TR-3, Census 2000 Data Capture,
U. S. Census Bureau,
Washington, DC. 20233

For sale by the Superintendent of Documents, U.S. Government Printing Office
Internet: bookstore.gpo.gov Phone: toll-free 866-512-1800; DC area 202-512-1800
Fax: 202-512-2250 Mail: Stop SSOP, Washington, DC 20402-0001

Contents

This page intentionally left blank.

Foreword

The Census 2000 Testing, Experimentation, and Evaluation Program provides measures of effectiveness for the Census 2000 design, operations, systems, and processes and provides information on the value of new or different methodologies. By providing measures of how well Census 2000 was conducted, this program fully supports the Census Bureau's strategy to integrate the 2010 planning process with ongoing Master Address File/TIGER enhancements and the American Community Survey. The purpose of the report that follows is to integrate findings and provide context and background for interpretation of related Census 2000 evaluations, experiments, and other assessments to make recommendations for planning the 2010 Census. Census 2000 Testing, Experimentation, and Evaluation reports are available on the Census Bureau's Internet site at: *http://www.census.gov/pred/www/.*

This page intentionally left blank.

1. Background

This report provides an overall synthesis of issues that were identified in several studies addressing the technical and operational elements of the complex and large-scale Census 2000 data capture system. The U.S. Census Bureau outsourced the two major components of the Census 2000 data capture program. Those components were the Data Capture System 2000 (DCS 2000) which was awarded to Lockheed Martin and the Data Capture Services Contract (DCSC) awarded to TRW. Lockheed Martin provided equipment for imaging, recognition, and data keying as well as the processing systems for four Data Capture Centers (DCCs). TRW provided staff and services for data capture, facilities management, office equipment, supplies, and office automation for three of the DCCs. (A fourth DCC was managed by the National Processing Center (NPC), a permanent Census Bureau facility in Jeffersonville, Indiana.) Within the report, a distinction is made between the two components, as appropriate.

The underlying system technology was developed through a contract awarded to Lockheed Martin. The contractor characterized this program as one of the largest image processing projects in history. The data capture system processed and captured data from 152 million census forms with an extremely high accuracy rate, which exceeded established goals (see Section 4.1). In actuality, the total number of census forms exceeded this figure. Based on a cost/benefit analysis, low volume forms were deliberately excluded from DCS 2000 as a risk mitigation strategy. The automated system was, in fact, designed to process 80 per cent of the forms volume while the remaining 20 percent of low volume forms were processed in a different manner.

Advanced technologies were employed to capture forms by creating a digital image of each page and then interpreting respondents' entries using Optical Mark Recognition (OMR) and Optical Character Recognition (OCR) processes.[1] This was the first time that the Census Bureau had used high speed OCR technology to capture hand written entries by respondents. Although OMR had been used in 1990, the automation in 2000 was more sophisticated because it included key from image (KFI) and OCR technologies as well as OMR. The system was highly automated but still relied on extensive operational support from contractors. Despite the reliance on technology, manual data entry methods were still needed to capture data in cases where the data were not machine readable, or if the form was damaged and could not be scanned or if the forms were low volume.

[1] OMR technology uses an optical scanner and computer software to scan a page, recognize the presence of marks in predesignated areas, and assign a value to the mark depending on its specific location and intensity on a page; OCR technology uses an optical scanner and computer software to "read" human handwriting and convert it into electronic form.

One aspect of the data capture system that has perhaps been overshadowed by the highly visible use of technology is the control processes used to manage the flow of forms through the data capture system and to monitor image quality. This workflow management system was a very effective mechanism that ensured all data were captured. According to Lockheed Martin, 1.2 million forms were rerun through the system. Although this was a small percentage of the overall number of forms that were processed, it nonetheless provided an indication of the stringent controls applied to monitor the process. There was a significant amount of census data associated with those 1.2 million forms. Different types and sizes of forms were processed through DCS 2000 and, in addition to capturing respondent answers, DCS 2000 electronically interpreted identification and control information on the forms. DCS 2000 had an automated monitoring feature that examined image quality by detecting over a dozen types of errors. A form recovery procedure was developed and implemented to handle questionnaires with those types of errors.

The data capture system employed a two-pass approach to capture data. The first pass commenced on March 6, 2000 and was completed on September 15, 2000. It captured the 100 percent census data (from both the long and short forms) needed for apportionment. The second pass captured the social and economic data (i.e., the

sample data). This was a shorter phase that started on August 28, 2000 and completed on November 15, 2000. The two-pass approach was used because the original keying rate estimates were too optimistic and the two-pass approach would ensure that data capture deadlines were met. The accuracy rate for OCR and OMR during both passes exceeded program goals. The manual keying accuracy rate also exceeded expectations.

Lockheed Martin, the prime contractor for DCS 2000, cited the systemic nature of DCS 2000 when explaining how it achieved high accuracy rates (Lockheed Martin, 2001b):

> *Automated data capture and the quality of the information produced lies at the heart of the DCS 2000 system. Many times in the image processing industry, products or systems claim automated character recognition rates of 99% or higher. But these rates are frequently calculated on preprocessed character test decks that rarely give an indication of how a system will work in an operations environment. DCS 2000 can make the same accuracy claim, but at a question level and on live Census production data. Moreover, this rate is obtained with nearly 80% of the data captured automatically. This level of automated capture did not come from simply a careful selection of commercial products or even by fine tuning the individual OCR and OMR components. These production statistics are the result of in depth tuning and complex integration of every component of the system.*

Indeed, there were 15 commercial-off-the-shelf (COTS) products integrated into the system. This approach was necessary given the limited time available to develop, test, and deploy the system. The COTS components provided the following functions: mail check-in and sorting; paper to digital image conversion; data base management; workflow management; digital image processing; optical character and mark recognition; data review and correction; digital tape backup and recovery; and system administration. The integration and tuning of these components were major accomplishments given the complexity of the DCS 2000 architecture.

According to the Data Capture Program Master Plan (PMP) (Brinson and Fowler, 2001), of the approximately 152.3 million census forms entered into data capture, approximately 83.9 million were mailback forms, 59.7 million were enumerator forms, 600,000 were Be Counted forms, and 8.1 million were Group Quarters (GQ) forms.[2]

[2] Actual numbers were reported after the completion of Census 2000. The final PMP issue date was March 30, 2001.

The Data Capture PMP reported that a cost model projected that the total *number of forms* to be processed would be 149.7 million. It further stated that approximately 1.5 billion form pages were processed during the data capture period. DCS 2000 output files were transmitted to the Decennial Systems and Contracts Management Office (DSCMO) on a daily basis. In order to manage this enormous workflow, DCS 2000 continually generated progress reports for management.

The overall management of the data capture system was a critical element contributing to the system's success. In addition to the NPC and the three DCCs, an Operations Control Center (OCC) was established in Lanham, Maryland to oversee all data capture operations. To assist the OCC with the management of the DCCs and their associated operations, the DCSC Management Information System (DMIS) was developed to provide a variety of integrated office automation tools. Raw data were transmitted to the DSCMO on a daily basis.

The data capture system succeeded in providing the population data needed for purposes of determining congressional apportionment, redistricting, and the distribution of over $100 billion of federal funds to state and local governments.

2. Scope and Limitations

The main focus of this report is to address the following four topics: performance of the data capture system; the system's ability to capture questionnaire data; the impact of data capture requirements on the questionnaire design and other factors; and the appropriateness of requirements identified for the data capture system. Other salient observations are included as well in view of their potential importance to future data capture systems and processes. The following documents were reviewed for this report:

1. Data Capture Program Master Plan (PMP) - Data Capture Systems and Operations
2. R.3.d. - Census 2000 Data Capture System Requirements Study by Titan Systems Corporation
3. K.1.b. - Evaluation of the Quality of the Data Capture System and the Impact of Questionnaire Capture and Processing on Data Quality
4. Lockheed Martin - Phase II Lessons Learned (including Appendix A, Technical Lessons Learned White Paper
5. TRW - Lessons Learned from DCSC Final Report
6. Rochester Institute of Technology Research Corporation - DCS 2000 Data Quality
7. Census 2000 Questionnaire Design Study by Titan Systems Corporation
8. Assessment Report for Data Capture of Paper Questionnaires, prepared by Andrea F. Brinson and Charles F. Fowler, Decennial Management Division
9. Lessons Learned for Census 2000, the Forms Design and Printing Office
10. Memorandum from Howard Hogan, January 24, 2000. Subject: Proposal for Quality Assurance of Census 2000 Data Capture.
11. Memorandum from Daniel H. Weinberg, December 7, 2000. Subject: Actions to Correct Pass 2 Keying Errors in Census Sample Monetary Fields.

Only two of the reference sources (#3 and #6 above) are based on empirical research. All other sources provide qualitative data.

In addressing the topics identified above, this report summarizes the key findings and major recommendations of the documents reviewed and seeks to identify any common themes or conflicting information between them. Therefore, this report is a high level, integrated assessment rather than being a critique of every facet of each study reviewed. It is not the intent of this report to re-visit the detailed statistical data contained in the documents that were reviewed.

Limitations stated in other reference sources also indirectly applied to this study. The two Titan studies and the K.1.b evaluation cited the limits identified below. Specific details on each limit are defined within the respective documents and are not fully described here due to space limitations.

Census 2000 Data Capture System Requirements Study

- The perception of those persons participating in the interview process can significantly influence the quality of information gathered
- In some cases, interviews were conducted several months, even years, after the participant had been involved in system development activities
- Each interview was completed within a one to two hour period, with some telephone followup to solicit clarification on interview results
- Every effort was made to identify key personnel and operational customers who actively participated in development efforts

Census 2000 Questionnaire Design Study

- The perception of those persons participating in the interview process can significantly influence the quality of information gathered
- Nearly two years have passed since participants were last involved in supporting Census 2000 activities

- Due to availability problems, Titan analysts were unable to interview the full range of personnel with knowledge about processing issues

K.1.b Evaluation

- Raw data are not a random representative sample of the U.S. population
- Failure to obtain all data originally planned
- Resolution of 666,711 records not matched to the twelve regional census center files
- Subjectivity in interpreting the most likely intent of the respondent
- Data reflect multiple sources of error beyond those attributable to system design

The collection of documents reviewed for this report identified important issues related to data capture topics. There were additional evaluations of data capture operations planned, which may have identified more issues. However, these evaluations were either not available by the time this report was completed or were cancelled altogether. Initially, this report intended to reflect the content of up to 11 documents, but due to the smaller number of references, the consolidated findings and recommendations will not be as extensive as originally planned. Despite this limitation, the report still covers a broad range of data capture issues, reflecting both quantitative and qualitative assessments.

3. Data Capture Topics Addressed in this Report

An expansion of each topic is provided below to give an appreciation for the scope of issues that were examined across all of the documents reviewed.

3.1 Performance of the data capture system

In order to address performance issues, a clear definition for the system's objective must be articulated. The data capture system was comprised of both automated and manual processes. Data capture equipment and related systems were acquired through a contract awarded in 1997 to Lockheed Martin Mission Systems. The automated system scanned a variety of forms and created digital images that were read by OMR and OCR software. OMR accomplished more than merely identifying marks in boxes. It was capable of recognizing cases when multiple marks appeared and used an Optical Answer Recognition feature that applied an algorithm and logic process to determine the most likely intended response. The OCR component was even more sophisticated. The Lockheed Martin study noted that OCR accuracy was a function of both its inherent ability to recognize characters and its contextual recognition capabilities. The excerpt below (Lockheed Martin, 2001b) explains how the OCR engine achieved the high accuracy rate:

> *First, not only does it recognize characters with a high degree of accuracy, it also provides multiple choices for each character and corresponding bounding character coordinates. This allows subsequent custom developed contextual processing to validate segmentation results as well as use an analysis of multiple recognition hypotheses in context and their probabilities of occurrence in order to further improve the results. Also, by providing a dictionary lookup capability as well as the description of the processing used to match or reject a word as a dictionary entry, the product allows even more opportunity for downstream analysis of the data during contextual analysis. Finally, because the product provides a vast array of definition parameters, it is also customized to treat each individual field with a high degree of detail and specificity, which will also maximize the accuracy and acceptance rates of the output.*

An Automated Image Quality Assessment (AIQA) application analyzed each imaged document. It corrected problems and enhanced images where possible. Once the forms were converted into an electronic format, the DCS 2000 software interpreted the data on the forms to the greatest extent possible. In those cases where OMR/OCR could not interpret the data within a certain range of confidence limits, the form image was automatically sent to KFI (key from image), an operation that required an operator to interpret the "low confidence" response data and then manually key the data into the system. Thus, as the Rochester Institute of Technology Research Corporation (RITRC) put it, "KFI got the bulk of the messy or ambiguous responses." The KFI process was described in the Data Capture PMP as follows:

> *The operators were presented with an image, called a "snippet," of the portion of the form they were to key. If a field required an action, the cursor was positioned on that field. Using their best judgement, the operators then keyed all the characters as they understood them from the image. For several situations, keying rules were provided to assist the operators in interpreting the information and entering standard responses.*

The Data Capture PMP notes that fields read by OCR and designated as "low confidence" images, and therefore automatically sent to KFI, were often correct. KFI had its own quality assurance process involving comparisons with OCR and/or a second keyer. Forms that could not be imaged were run through KFP (key from paper) to capture all data manually. KFP involved two keyers, with the second providing verification of the data entered by the first operator.

RITRC's sampling of production data looked at the acceptance rate[3] for both OCR and OMR for the

[3] RITRC defines acceptance rate as the fraction of fields in which the OCR has high confidence, usually expressed as a percent. Accepted fields are the ones RITRC scored for OCR accuracy; they are not sent to keyers except for QA purposes.

mailback and enumerator forms (100 percent data). For the D-1 short form and the D-2 long form, the acceptance rate was 83.08 percent for OCR and 99.89 percent for OMR. For the D-1E and D-2E enumerator versions of these forms the acceptance rates for OCR and OMR were slightly lower at 79.17 percent and 99.78 percent, respectively. Based on these findings, and other considerations, RITRC concluded the data quality from both sets of forms was about the same, with both exceeding the program goals.

The quality of OMR, OCR, KFI, and KFP was constantly monitored. The accuracy rates for OMR and OCR data capture were contractually specified as 99 percent and 98 percent, respectively. (OCR accuracy was actually sub-divided into two separate accuracy rates of 98 percent for alphabetic data and 98.5 percent for numeric data.) Keyer accuracy for KFI and KFP was also measured. The accuracy standard for KFI was 96.5 percent and KFP was to have no more than a 2 percent error rate.

Given the complexity of the data capture environment, the volume of forms processed, and the use of state-of-the-art technologies, it is instructive to examine the performance of the overall system. The examination of DCS 2000 performance issues is not intended to be a fault finding exercise. Rather, it provides a view into issues that can lead to a better understanding of the effectiveness of the system used during Census 2000. This information can, in turn, benefit future data capture operations by providing insights into the benefits and limitations the technology and manual systems employed.

3.2 The system's ability to capture questionnaire data

A significant drop in the nationwide mail response rate during the 1990 Census led to dramatic changes in questionnaire design strategies for Census 2000. The major impetus for change in the questionnaire design came as a result of Congressional direction, which brought about efforts to make the mailback forms more respondent friendly. The assumption was that respondent friendly forms would lead to an increase in response rates. During the decade leading up to Census 2000, the Census Bureau conducted research into forms design issues in an effort to increase mail response rates and improve data quality. There were a number of methodological tests targeted at improving the Census Bureau's understanding of the array of cognitive, technical, and overall design factors. These factors influence the efficiency and accuracy of data collection and capture processes. The testing included a range of studies that examined the layout of the questions, the testing of matrix formats against individualized formats, and evaluation of different envelope colors.

Reflecting the combination of new design initiatives and the availability of sophisticated scanning technologies, the short form underwent significant changes for the Census 2000. The resulting form exhibited an entirely new and more respondent friendly individual space layout (separate panel for each person) and provided features such as a background color, motivational icons, a Census 2000 logo, check boxes, and segmented write-in spaces.

Lockheed Martin came to appreciate the criticality of forms design and its contribution to capturing respondent data (Lockheed Martin, 2001b):

> *Of all the aspects of an automated data capture system, the absolutely most critical component of the system is the design and printing of the forms. A good form design can increase the stability, flexibility, error detection and recovery, and performance of the system. A poor form design can adversely affect all of these factors and subsequently increase the system cost exponentially. The experiences of DCS 2000 helped to emphasize these points.*

The Assessment Report on Data Capture of Paper Questionnaires (Brinson and Fowler, 2003) pointed out some particular forms design and subsequent effects on printing that affected the system's ability to capture data. It observed that "Forms Design and Printing was not coordinated with the data capture technology...until later in the process making it more difficult to design and test the data capture technology." The report suggested that the automation technology available may not have been fully utilized. The report further states the following with regard to how certain forms design and printing issues impacted the OMR and OCR subsystems:

> *The multiple form types, booklet style formats, question design, and specific colors used made the implementation of OMR and OCR technology more challenging. Also, the lateness in finalizing questionnaires and printing of prototypes made the development of OMR and OCR software more complicated, of higher risk, and more costly.*

The need for forms design and printing to be tightly integrated

within the overall data capture system development environment is apparent and was echoed in several of the documents reviewed for this report.

3.3 The impact of data capture requirements on the questionnaire design and other factors

There were several image capture specifications that created constraints in the forms design environment. In the Census 2000 Data Capture System Requirements Study, Titan identified the following areas where DCS 2000 had clearly defined specifications (Titan, 2002):

- *Size of Scannable Documents.* A set of four specific paper size dimensions for single sheets was approved for DCS 2000. The booklet questionnaires also had defined limits for the size of separated sheets. According to DSCMO, DCS 2000 processed six different sized forms, but was not limited to this number.
- *Optical Character Recognition (OCR) Write-in Fields.* As noted by RITRC, the OCR subsystem was designed to read all of the write-in fields for which there was a high level of confidence. Consequently, there were a variety of very precise criteria defining dimensions and spacing requirements for these fields. The basic purpose of these criteria was to facilitate character recognition by the data capture system.
- *Optical Mark Recognition (OMR) Check Boxes.* In addition to size and spacing requirements for these boxes, there were also specifications indicating that boxes on one side of the page should not coincide with check boxes or image areas on the reverse side of the page. However, according to DSCMO, paper specifications required a high opacity level to minimize "show through."
- *Color and Screen.* A major data capture requirement was that the background color must "drop out" and should not contain any black content. A drop out effect can usually be achieved through a combination of the color and screen dot size.
- *Margins.* Required white space was defined for side, top, and bottom margins.
- *Form Type/Page Identifier.* There was a set of requirements for the use of the Interleaved 2 of 5 bar code, which served to identify form type and page numbers.
- *Document Integrity.* Since booklet forms could become separated during the scanning process, a unique identifier had to be included on all sheets of a long form to link the sheets. Another bar code was printed in the margin area to provide the sheet linkage function necessary to ensure Document Integrity. Document Integrity was also included on both sides of the short form to mitigate the risk of non-detected double feeds.

The constraints imposed on the forms design by the data capture requirements must be viewed in terms of their contribution to highly efficient data capture processes. In all, 26 different forms were scanned using OMR and OCR technologies, resulting in a substantial labor savings achieved from DCS 2000.

3.4 The appropriateness of requirements identified for the data capture system

Like all systems, the Census 2000 data capture system was designed to satisfy a set of requirements. A system cannot provide the right functionality if the requirements defined for it were incomplete. Thus, the efficacy of the requirements definition process determines to a great extent how well the system will work.

Research into new technologies that could make the data capture process more efficient began in the early 1990s. The Rochester Institute of Technology (RIT) tested a variety of commercial off-the-shelf (COTS) products in 1993 and 1995. A Request for Proposals (RFP) was developed in 1996 to procure proven technologies and to outsource the development and operation of the data capture system.

In the pre-award phase, multiple vendors were asked to conduct an operational capabilities demonstration. According to the Census 2000 Data Capture System Requirements Study, this demonstration allowed the Census Bureau to identify the contractor most suited to the task of developing DCS 2000 and served to identify and fine-tune requirements for the data capture system. The award to Lockheed Martin was issued on March 21, 1997 and development activities ensued.

The original statement of work (SOW) was used for development up to the Census 2000 Dress Rehearsal. At that point it was determined that the SOW lacked sufficient detail and required more specifics. Consequently, a Functional Baseline (FBL) document was developed to focus on what

functionality was needed. The FBL was given to Lockheed Martin, but the requirements for DCS 2000 continued to evolve throughout the development of the system. Similarly, TRW had to refine the Operations and Facilities Plan as development proceeded. The basic requirements for this plan were generated jointly by the Census Bureau and Lockheed Martin. The requirements were also included in the RFP for DCSC, but TRW was not awarded the contract until nearly a year after Lockheed Martin had commenced development of the system. By this time, refinements to the plan needed to be made by TRW to keep pace with DCS 2000 developments.

The question posed in relation to requirements can be put in the context of defining what the entire data capture system needed to do — not just the automated scanning technology. Therefore, this report looks at requirements issues from both the automated and operational aspects of the data capture system.

4. Findings

This section provides major findings and key issues that were echoed in the documents reviewed for this report. They are discussed within the context of the four main topic areas of this report; assessing performance, factors affecting the system's ability to capture data, issues relating to the impact of data capture requirements, and examining the appropriateness of those requirements.

4.1 Assessing the performance of the data capture system

4.1.1 DCS 2000 exceeded performance goals. As noted in the DCS 2000 Data Quality report prepared by RITRC, Census 2000 was the first time that the Census Bureau had used commercially available scanners and high-speed digital data processing to capture census data. Although not a completely definitive assessment of quality assurance (QA) for DCS 2000, the RITRC analysis concluded that "The results we obtained from Production Data Quality Sampling indicate that DCS 2000 production data quality significantly exceeded program goals." Putting this success into context, the data capture evaluation (Conklin, 2003) noted that "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."

Performance goals for accuracy were exceeded for both mailback forms and the enumerator forms. RITRC stated that, based on a sampling of images and data, the in-production "OCR accuracy was 99.6 percent, KFI was in excess of 97.8 percent, and the check-box question accuracy was a little over 99.8 percent." RITRC also reported another significant finding: the overall write-in field accuracy for both passes of data capture (merged data combining OCR and KFI) was 99.3 percent. For Title 13 data (merged from OCR and KFI) on the mailback forms, numeric fields were found to have lower error rates than alpha fields. The data capture evaluation found that for all fields, OCR had the lowest error rate, followed by OMR and then KFI.

(Note: The K.1.b findings concerning error rates for the three different data capture modes suggest a discrepancy with respect to RITRC's findings. The OMR accuracy rate reported by RITRC indicates that this mode had the lowest error rate, however the K.1.b evaluation placed OMR after OCR in terms of their respective error rates (i.e., OCR was lower than OMR). This apparent discrepancy is complicated by the fact that these studies used different performance classification methods. It is important to note that error rates differ from accuracy rates. The error rate is calculated by subtracting the accuracy rate from 100, typically yielding a very small number (e.g., .4 percent), whereas accuracy rates are the fraction of accepted fields that are correct, and are therefore associated with much higher numbers (e.g., 99.6 percent). Also, the studies were based on different data sets. Attempting to compare OCR to OMR to see which is the best is not a valid comparison since the two methods are designed to accomplish different types of recognition tasks. Given the exceptionally high performance of the OCR and OMR subsystems, both of which exceeded program goals, there is no utility in attempting to determine a "winner".)

The data capture evaluation examined the issue of whether or not some fields were sent to KFI more often than others. It concluded that name related fields were more likely to go to KFI unnecessarily. It also posed the question of whether certain fields sent automatically to KFI should be processed instead by the automated technology. The discussion of this issue suggests further research is needed:

> *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.*

One important lesson learned cited by Lockheed Martin during the Census 2000 Dress Rehearsal was that the "full-field[4] method of keying provided the most cost-effective combination of accuracy and throughput given the requirements of the DCS 2000 system." Other types of keying methods may have increased throughput, though accuracy would have suffered.

4.1.2 Reasons for errors.
According to RITRC, some percentage of KFI "errors" in their production data quality sampling was due to ambiguous handwriting, misinterpretation of keying rules, or varying interpretations of handwritten entries (where respondent intent was not clear). Therefore, these were not considered as actual keying errors. The data capture evaluation noted that the KFI error is "not necessarily a poor reflection on the automated technology" and observed that the automated technology and the evaluation and production KFI are prone to the following errors:

- failure to read a field on the form
- picking up content that is not really there
- incorrectly capturing the content on the paper
- correctly capturing what the respondent wrote, but not necessarily what the respondent intended

The error rate can be attributed to factors such as the hardware of the automated technology or the software. Lockheed Martin identified the two main contributors to system errors: noise (including noise generated by respondents) that interferes with character recognition and segmentation errors (multiple characters within one box or single characters spanning multiple boxes). This was common to other image processing systems as well as DCS 2000.

Another major cause of keying errors was the use of certain keying rules. For example, the rules called for a write-in field to be filled with "eights" if the number could not be determined and "nines" were to be used if a value exceeded certain limits. RITRC expressed concern about keyer confusion over the use of "eights" and "nines", particularly with respect to the income field with its imbedded dollar sign, comma, and pre-printed cents. They also noted that "the essential KFI problem...was the conflict between "key what you see" and interpretations required by the keying rules." The Decennial Statistical Studies Division (DSSD) has also reported problems with the income field, citing an "error rate of nearly 45 percent for income fields filled with 9s" and "mistakes...in fields containing all 8s." (Weinberg, 2000).

The data capture evaluation reported, across various modes of data capture, that the most frequent reasons for failing to capture the intended responses were:

- Extra-check-box - the output from the automated technology output shows more check-boxes marked than are in the scanned image.
- Missing characters - the output from the automated technology has fewer characters than the scanned image.
- 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 same report listed the most common reasons for these problems as:

- Poor handwriting - the respondent's handwriting makes one letter look like another, but a person 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.

The data capture evaluation reached some significant conclusions. First, it found that "if there is intelligent content in a field, the automated technology will detect it with nearly perfect certainty." Second, despite the fact that the system is not perfect, "a sizeable portion of responses will be captured and interpreted correctly at speeds that are orders of magnitude above KFI." And third, "the largest impediment to automation is not the quality of the hardware or software, but the quality of the responses supplied by human beings." The study suggests that attempting to build a system that could capture nearly any type of response would not be a practical endeavor.

4.1.3 OCR acceptance rate.
According to the data capture evaluation, 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 problems. The percent accepted by OCR for write-in fields (short and long forms) was 78.9 percent, which is quite close to the 81.2

[4] Full field keying involves re-entering the entire field as opposed to character keying where only the low confidence characters in a field are entered.

percent reported by RITRC for Pass 1. However, RITRC reported a much lower rate for Pass 2 of 64.8 percent. The lower rate reflects the problems inherent in interpreting the more difficult long form responses.

The OCR acceptance rate for forms completed by enumerators was lower when compared to the rate for mailback forms. RITRC believes that this may have been due to light pencil marks or incomplete erasures. Thus, this lower acceptance rate was probably more of a function associated with the writing instrument (i.e., pencil), and not a reflection on the effectiveness of the OCR subsystem.

4.1.4 Cluster concept. DCS 2000 was based on the concept of cluster processing. Clusters were autonomous units of image processing, constructed around the capacity of three scanners. Each DCC was equipped with as many clusters as necessary to process their workload. In Lockheed Martin's opinion, the cluster design was a key factor that contributed to the successful development of the overall system. In this concept, each cluster functioned independently, using a set of modules, with a sequential processing order that verified the previous actions taken by other modules. Lockheed Martin described the operation of the cluster concept as follows:

> *This efficiency [of verifying each step of the process], which frequently exceeds automated acceptance rates of 80% of the data at greater than 99% accuracy, also incorporates the self-validation themes of cluster processing. While the obvious validation steps are the keying operator functions, there is also significant use of cross character, cross field, and cross image contextual checks that are performed in order to validate and re-validate the data that is produced. This processing is also spread across the various steps of the cluster workflow. At the context step, OMR box, character, field, and image validations are performed to maximize the efficiency of the automated OCR and OMR. Then, another module assures that control data and form integrity have not been compromised. Next, key from image functions (KFI) are then used to complete the capture of data that could not be done automatically. While these keying steps also include real-time edits, functions for operators to reject entire images for quality, and automated quality assurance checks, there are also subsequent full form validation functions that assure that the captured data is consistent with expected responses. All these processes work together to continuously validate and improve the captured data as it progresses through the system.*

In short, the cluster concept ensured images were subject to stringent quality checks. The prime DCSC contractor, TRW, noted that quality was designed into the process due to the short-term, high volume nature of the work. They concluded that "this provided a high-quality product with a minimum of QA staff and overhead." In this regard, the prime DCS 2000 contractor, Lockheed Martin, stated only one percent of the forms needed to be pulled and reprocessed. Another benefit of the cluster architecture was that it allowed for continued scanning of forms, even when there were component failures within a cluster.

4.1.5 Inefficiencies in the KFP process. The DMD Assessment Report on Data Capture of Paper Questionnaires characterized KFP as being an "inefficient way to capture forms that could not be captured by scanning." The report noted that the software used for KFP was designed for image keying and was therefore "cumbersome". It also took issue with the KFP policy of requiring 100 percent verification stating that this may have "caused more keying than was required since a sample verification may have been sufficient." In general, the DMD report favored maximizing the use of automation and relying less on keying.

4.2 Factors affecting the system's ability to capture questionnaire data

Noting that many forms processing systems can define a form to run through the system in minutes, Lockheed Martin observed that the complete definition of DCS 2000 was dependent upon utilizing all of the optimizations designed into the form itself. This means the form ultimately reflected an in-depth analysis of the total environment including Census rules, form characteristics, and respondent tendencies.

4.2.1 Keying rules and methods. While allowing for respondent tendencies was factored into the forms design process, there was still a need to apply keying rules and methods for capturing data. According to RITRC, one of the major causes of keying errors was the complexity of the rules, which required keyers to make interpretations–while maintaining a high pace of production. According to TRW, the application of keying rules varied between Passes 1 and 2. They noted that the application of rules was limited on Pass 1 because of the limited number of fields; the rules were more critical

in Pass 2 because of the broader range of field types. TRW reported that daily audits showed a high degree of accuracy during Pass 2 keying. Nonetheless, in view of the need for interpretations by keyers and variations of rules between the passes, the more accurately the forms are filled out, the less need there is for keying. Basically, the system's ability to capture questionnaire data can be improved by better forms design practices.

4.2.2 Improvements to forms design. Given the success of DCS 2000 as a high speed, high volume, high quality, data capture mechanism, some aspects of the form could be improved. The data capture evaluation cited and endorsed possible improvements to the form that had been identified in Titan's Census 2000 Questionnaire Design Study.

The data capture evaluation recommended the following be considered as possible improvements:

- Have the person information for household members be filled out from left to right across the page instead of up and down.
- Allow the use of pencils so respondents can easily correct mistakes.
- Change the sizes, fonts, appearance, etc. of the instruction icons so they are easier to spot (or simply eliminate them).
- Allow more spaces for the last name fields.
- Include instructions for filling out or correcting write-in fields.
- Include more detailed instructions for the race and ethnicity questions. While additional instructions may improve recognition, DSCMO and others (e.g., Dillman) expressed concerns that an overcrowded form with too many instructions may hinder response and data capture.
- Try to make the instructions to the head of household for filling out the form more concise.
- Employ the use of 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.
- Select a background color with better visual contrast.

Enhancements to these areas have the potential to further improve the quality of data captured and perhaps make the form even friendlier to respondents.

(Note: The Census 2000 Questionnaire Design Study was a qualitative study that reflected the insights and experience of subject matter experts who had extensive knowledge of forms design. While the findings from this study suggested that certain aspects of the questionnaire could be improved, we caution that further research and testing is needed to determine which, if any, of the recommendations should be implemented. It is important to note that one of the main purposes of this study was to identify and highlight forms design issues that could be candidates for future research efforts.)

4.2.3 Few instructions were provided to respondents. Several of the above bullets touch on this subject. The system's ability to capture respondent data could have been impacted, to a certain extent, by a lack of instructions to respondents. As noted in the Census 2000 Questionnaire Design Study, the short form does not provide guidance to respondents on what to do if an answer exceeds the length of a write-in box or how to correct any mistakes on the form. The study suggested that some questions could benefit from expanded instructions, although there are risk/benefit trade-offs that would need to be assessed. It is conceivable that additional instructions for respondents could have reduced problems and enabled the system to capture more respondent data, rather than rejecting it as unreadable.

4.2.4 Use of banking should be minimized. Most interviewees who participated in the Census 2000 Questionnaire Design Study understood the need for banking but felt that it was not a desirable design feature and its use should be minimized. This technique, especially triple banking, can lead to tightly grouped check boxes. Besides being confusing, it increased the likelihood of a processing problem when a large mark extends beyond the box boundary and into adjacent boxes. Greater spacing between boxes may minimize stray marks that create interpretation problems for OMR.

4.2.5 The race question. Due to the multiple answers allowed by this question, Lockheed Martin found this question difficult to arbitrate with any degree of high accuracy. Owing to the high importance placed on accurately capturing race data, when multiple marks were detected by the OMR subsystem, they were passed to keyers for resolution. According to Lockheed Martin, the accuracy rates for this "particularly sensitive area of interest" increased significantly as a result of the manual interpretation.

4.3 Issues relating to the impact of data capture requirements on the questionnaire design and other factors

The complete redesign of the questionnaire for Census 2000 produced a more respondent friendly form based on an individual-space format. To a large extent, the new design of the form was made possible due to technological advancements made in OCR and OMR technologies.

4.3.1 Space separation between boxes and text. There had to be a space separation between boxes and text of at least .08 inches, and space between boxes (from the bottom of one box to the top of the next box) of at least .05 inches. The DCS 2000 image capture specifications noted that more space between boxes was preferable in order to prevent a large mark from running into another box. Although conforming to these specifications, many boxes on the form were tightly grouped. For example, the Hispanic origin and race questions, and question #2 for Persons 2 - 6 contained numerous boxes that were tightly grouped with minimal vertical space separation. This increased the likelihood of a data capture error when a large mark extended beyond the box boundary into an adjacent check box or segmented boxes. Future use of check boxes should allow greater spacing as requested in the DCS 2000 image capture specifications.

4.3.2 Darker frame around segmented boxes and darker segments. The DCS 2000 image capture specifications prohibited the use of dark outlines surrounding OCR fields. A darker outline could have provided more definition to the boxes and therefore potentially reduce one of the main sources of data capture problems–segmentation errors. The segmentation lines had low contrast and did not show up well, which may have accounted for some of those types of errors where characters spanned more than one box or more than one character was written into a single box. The use of a different background color with a higher contrast to white should alleviate the problem.

4.3.3 Background color was problematic. The choice of the background color (Pantone 129) by the graphics arts firm, Two Twelve Associates, met data processing specifications for "dropping out", or becoming invisible to the scanner. However, according to the Titan's Census 2000 Questionnaire Design Study, this particular color did not provide the best contrast. Another study conducted by RITRC, after the color had been selected for the Census Bureau, stated that this particular color "was on the fringe of acceptable drop-out colors for the Kodak scanner used for DCS 2000."

As stated in the Census 2000 Questionnaire Design Study, DSCMO felt the choice of Pantone 129 compromised the ability of the scanners to use more aggressive settings to read lighter shades of answers. The choice of this color did not generally present any problems for dropping out during the forms processing function. This can be attributed to tight QA monitoring during the forms production process. Without an effective QA process, there could have been additional problems with the form background color failing to drop out, causing the scanning equipment to reject a questionnaire. While the color on the form was controlled within the specification parameters by using instruments, according to the interviewees and the reports provided by DSCMO, color generation was not always consistent during the printing process and there were noticeable variations.

Although technically meeting data capture requirements, the selection of Pantone 129 was, in retrospect, far from being optimal. Personnel with expertise in data capture operations need to have input into the color selection process to assess the implications on data capture operations.

4.4 Examining the appropriateness of requirements identified for the data capture system

Through extensive interviews with Census personnel, the Census 2000 System Requirements Study (Titan, 2002) found that many people felt DCS 2000 was the right system for the job and provided an efficient and effective means to capture census data. However, there were several requirements related areas that could be improved.

4.4.1 Fluid requirements. Leading edge technologies were being employed on a very large-scale operation, so it is understandable that requirements would be altered and evolve as the plan changed. (Ideally, requirements should be sufficiently flexible to accommodate new technology.) However, the fact that a pre-test version of DCS 2000 was used late in the life cycle, in the Census 2000 Dress Rehearsal, suggests that system planning and the requirements definition process needs improvement and schedules/timelines should have been adjusted to ensure the system was fully prepared for the Dress Rehearsal.

According to the Assessment Report on Data Capture of Paper

Questionnaires, numerous requirements (six were identified) had to be added after the Dress Rehearsal based on the lessons learned. The report concluded that "the addition of requirements after the data capture system had been designed or was in the final testing phase provided a significant increase to the contract cost, and risk to the quality of the data, and to the data capture schedule."

Implementing DCS 2000 and the DCSC operations posed significant challenges integrating new technologies and complex operations. A robust testing program seemed to compensate for any lack of requirements or understanding of exactly how the DCS 2000 components and operational procedures were to function in the production environment. Consequently, there was reliance on extensive testing to simulate the Census environment.

4.4.2 Keying rules. According to the Census 2000 Data Capture System Requirements Study, the keying rules changed after production began and continued to remain an issue throughout the contract, creating risk to data quality and to timely completion of data capture. It is unknown to what extent changing the keying rules impacted the quality of the data capture, but the changing of rules between the first and second passes were reported to have occurred. Requirements for keying rules were certainly fluid. For example, the DCSC contractor made staffing decisions based on the expected "key what you see" method, which was subsequently changed. Recognizing that this type of major change presents a significant data quality issue that can greatly increase risks to the program, the Census Bureau should place more emphasis on fully defining firm requirements for keying rules.

4.4.3 Quality Assurance. The requirements for QA could have been better defined. Although the framework for the overall quality assurance plan was decided by the Census Management Integration Team (CMIT), QA specialists in the Census Bureau differed with the CMIT on the application of QA standards for DCS 2000. Complicating this situation was the fact that Lockheed Martin and TRW had their own internal QA programs, and the Census Bureau had implemented its own independent QA of the Census 2000 data capture process at the National Processing Center. Both the 100 percent and long form sample data were monitored by the Census Bureau at this facility. Since it was believed that Lockheed Martin would easily find gross errors, the Census Bureau's QA monitoring scheme concentrated on reviewing rarer events (e.g., multiple races or American Indian tribe designations) that were not captured.

Specific recommendations from the Census Bureau for improving the quality assurance aspect of DCS 2000 were provided very late in the development process and, if adopted, would have necessitated a major redesign of the DCS 2000 software and the post-processing operations at Headquarters. The parameters, processes, and responsibilities for QA measurement should be included as part of the requirements definition process.

4.4.4 Archiving requirements. Another problematic requirement was the need for archiving census data. The Census Bureau was originally advised that ASCII files, not images, would be required by the National Archives and Record Administration (NARA). These requirements were later changed to include microfilmed images and an index. According to the Assessment Report on Data Capture of Paper Questionnaires the cost incurred by the Census Bureau to meet the new archiving requirements was approximately $44 million.

4.4.5 Operational perspective. With the DCSC contract award occurring nearly one year after the award to Lockheed Martin, TRW never had the opportunity to have a major influence on the requirements for DCS 2000. TRW noted that they had "little opportunity to influence system requirements with respect to the user interface and the need for management data on the production floor." Some of the system requirements suggested by TRW during the testing phase could not be implemented due to resource and schedule constraints.

In the Lessons Learned for Census 2000 document, the Forms Design and Printing Office observed that it "did not have a comprehensive set of data capture and processing technical requirements to include in the forms design and print contracts for the DR." The same report noted that after the Dress Rehearsal "late and changing requirements led to major revisions on all forms being electronically data captured" and there was no time to test the revised changes. In conclusion, the report stated that "all these changes put the Census Bureau at risk of failing to meet scheduled deliveries."

5. Results of Analysis

This section assesses the findings and key issues for each of the four areas reviewed and provides an overall, high-level view that reflects the information synthesized from the documents reviewed. It also presents other salient observations about the Census 2000 data capture system that were deemed to be relevant to this report and planning for the 2010 Census.

5.1 System performance

Given the massive volume of forms processed with a very high degree of accuracy and acceptance rates, the data capture system was an unqualified success. As mentioned earlier in this document, the system exceeded all of its performance goals. While some errors could be attributable to limitations of the automated system (noise and segmentation errors), many were attributed to ambiguous or poor handwriting by respondents. In its classification of various Pass 2 write-in errors for overall Title 13 data, RITRC cited ambiguous responses, write-overs, crossouts, and poor handwriting as accounting for a substantial number of errors. RITRC's research also found that "No reason" accounted for over half of the errors in the same data. Similarly, the K.1.b evaluation found a high number of unexplained errors in the data it reviewed. Interestingly, RITRC was comfortable with the large percentage of "No reason" errors. In their assessment, this was a "very good sign...because a well-designed system should have the bulk of its errors in the "noise." The data capture evaluation provided an excellent perspective on system performance: "The largest impediment to automation is not the quality of the hardware or software, but the quality of the responses provided by human beings." The data capture evaluation suggested that attempting to build a system that could capture nearly any type of response would not be a practical endeavor because of the various permutations of human errors.

5.2 Capturing respondent data

Titan's Census 2000 Questionnaire Design Study highlighted the sophistication of the forms design process and its awareness of the need to efficiently capture respondent data. The study provided the following description of the forms design environment:

> *The design of the Census 2000 short form questionnaire was a complex undertaking that reflected the combined efforts of many people at the Census Bureau. Every facet of the form was carefully analyzed for its potential effect on response rates and data quality. Because the forms were to be processed by sophisticated automation employing both optical character and mark recognition technologies, the designers faced the extra challenge of also having to meet image capture specifications that placed constraints on the form.*

While the success of Census 2000 (as discussed in section 5.1) reflected well on the overall forms design effort, the data capture evaluation noted that "considerable human resources were...required to handle the many millions of write-in fields that posed a problem for it." This suggests the need for continued research into ways of improving forms design to minimize the need for manual keying operations.

5.3 The impact of data capture requirements

Although meeting data capture specifications, the background color was widely recognized as being problematic with respect to being on the fringe of acceptable drop-out colors and from a visual contrast standpoint. Selection of this particular color was not the result of a collaborative effort involving subject matter experts or any quantitative analysis. The lesson learned from this, and other related experiences, is that there should be close coordination between data capture personnel and those involved in questionnaire development throughout the forms design life cycle.
Technology may impose some limitations on designers, but that same technology can also enable more sophisticated design techniques. For both of these reasons, tight collaboration and communication between the two groups are essential when requirements are being developed or changed.

5.4 Fluid requirements posed substantial risks

Requirements proved to be subject to revision in several major areas. To a large extent, the deficiencies in requirements were compensated for by extensive system testing that helped to refine the Census 2000 data capture system. This was a very risky approach for such a major, high profile system. The importance of having a well-defined and disciplined structure for developing requirements cannot be overemphasized as they define what the system needs to do and what the performance metrics are, along with establishing operational processes and QA parameters. Additionally, the selection of qualified contractors depends on a thorough understanding of requirements.

It is worth noting that the effort to define requirements was handicapped in several respects. First, there was no well-established process in place to guide and facilitate the development of requirements. Second, Census Bureau personnel were not accustomed to preparing requirements within a contracting environment. This issue was highlighted in the Assessment Report for Data Capture of Paper Questionnaires, which stated that "this was the first time that the Decennial program areas had to do their work for data capture within contracting guidelines and contracting time constraints." And third, adequate funding was not made available early enough to allow system requirements to commence in a proactive fashion. The combined effects of these three factors created a challenging environment for requirements development.

5.5 Other salient observations about Census 2000 data capture system

5.5.1 Agency-contractor partnership was a key success factor. To a large extent, the success of the Census 2000 data capture program was due to a healthy partnership atmosphere that existed between the Census Bureau and the contractors. Considering that the system was not completed in time for Dress Rehearsal, requirements were fluid, and that there were differences over QA processes, it is evident that agency-contractor cooperation was a major factor in ensuring the success of the Census 2000 data capture system.

Within the agency-contractor partnership there was a subtle relationship that also existed between the two prime contractors. Even though there were two separate contracts, one for DCS 2000 and another for DCSC, there was a mutual dependence between them because the award fees were tied to shared performance evaluations. This contractual arrangement fostered a cooperative relationship which ultimately benefitted the Census Bureau.

5.5.2 Change control processes were effective. Given the dynamic requirements, and the constant incorporation of new technologies being applied to make DCS 2000 a robust platform, discipline in the change control process was a major plus that mitigated risks to the program. Both prime contractors adhered to strict change control processes for system components and operational procedures.

The Assessment Report on Data Capture of Paper Questionnaires generally agreed that change control was a contributing factor to the success of the data capture system. The report specifically noted that the DSCMO contract office had established a "highly effective change control process to track, evaluate, and control changes to the Data Capture program... throughout the development of the program." It added that the process received favorable review from oversight bodies because of their focus on cost control and schedule deadlines. While acknowledging the success of DSCMO, DMD noted that it was responsible for gathering requirements and, unlike DSCMO, DMD did not have a dedicated staff to manage change. It would prefer to see a more centralized requirements change control process implemented.

5.5.3 Comprehensive testing. As noted in the Census 2000 Data Capture System Requirements Study, the thoroughness of the testing compensated for the lack of a solid set of requirements. In fact, requirements were very dynamic, changing throughout the entire development period for DCS 2000 and into production. DCS 2000 underwent a series of tests: Site Acceptance Test (SiteAT), Operational Test and Dry Run (OTDR), integrated Four Site preproduction test, and Beta Site testing.[5] According to the Lockheed Martin Phase II Lessons Learned document, these tests were extremely successful. TRW found the OTDRs to be "key contributors to the success of Census 2000 data capture" as they closely resembled live operations and exercised every facet of operations and included the OCC.

[5] According to the Data Capture PMP, the DCS 2000 contractor operated the Bowie Test Site which housed prototypes of the software and hardware environments used in the DCCs. System, integration, and security testing and troubleshooting for DCS 2000 software was performed at the Bowie Test Site by the contractor.

5.5.4 Operational discipline. Stringent control over the consistency of operational procedures helped ensure operational consistency across each of the sites. TRW reported that each of the three contractor-run DCCs was organized in the same way and all had essentially the same functions such as: operation of the data capture process; human resources management; workforce training; QA activities; and facilities management.

5.5.5 Workflow management. As mentioned in the Background section of this report, one aspect of the data capture system that may have been overshadowed by technology is the control process used to manage the flow of forms through the data capture process and monitor image quality. This workflow management system was a very effective and structured mechanism that ensured that all data were captured. Basically, it was responsible for ensuring the complete and orderly flow of forms, identifying problems, and rerouting forms to handle a range of exceptions. Lockheed Martin noted that: "The workflow is superficially straightforward, consisting of a series of sequential steps with most processes passing their results to the next, with little forking of workflow cases. However, underneath it is complicated because of the rerouting that is required."

Another inconspicuous aspect of the workflow management system was the underlying software that integrated the DCS 2000 COTS products. The unique Generic Application Interface allowed new or updated workflow COTS products to be easily integrated into the workflow system.

One critical step in the workflow process was checkout. After batches of forms were processed, they arrived at the checkout station and a verification process ensured that each form that was received was processed. Any forms that needed to be reprocessed were sent to the Exception Checkout handler. This illustrates that rigid controls were built into the workflow process through the last step of the chain.

Most importantly, the workflow process complied with Title 13 protection requirements. The Assessment Report for Data Capture of Paper Questionnaires discussed the importance of questionnaire accountability:

> *The successful protection and security of the questionnaires was of primary concern during the data capture period and subsequent forms destruction. The accountability for the data capture of paper questionnaires once they were received at the Data Capture Centers used a check-in system, batch processing through scanning, or Key From Paper, and a positive checkout system which verified that all forms were processed and that their data was received by Headquarters Data Processing.*

Understanding the criticality of an orderly workflow management process and designing efficiency into that process were key success factors for the data capture system.

5.5.6 Modular system design. DCS 2000 was a flexible system architecture that could adapt to changing requirements. For example, a significant process change occurred very late in the program when a decision was made to use a two-pass data capture process. According to Lockheed Martin, because of the system's modular design, workflow, image replay capability, inherent robustness, and a high level of configurability, conversion to the two-pass method was a relatively simple matter. A major lesson learned that was cited by Lockheed Martin addressed the overall system design and the adaptability of the system:

> *The fact that these changes [switch over to the two-pass method] could be implemented so close to the start of production reemphasizes the positive lessons learned that accompany the benefits of incorporating each of the DCS 2000 design themes at the earliest stages of development.*

5.5.7 Forms Printing. The Census Bureau recognized that monitoring print quality on-site was integral to success. Lockheed Martin concurred with this direction. They cautioned that "problems can quickly affect enormous quantities of forms" and therefore the quality of Census 2000 can depend on maintaining printing standards and consistency. For this reason, the Census Bureau had an extensive and automated print QA process.

This page intentionally left blank.

6. Recommendations

This section provides recommendations stemming from the various lessons learned that were cited in the documentation reviewed.

6.1 Unify the design strategy needed for the data capture system

RITRC recommended future data capture systems be developed within the context of a unified framework. Their view of the system would includes the data capture components and, in addition, the printing, forms design, recognition, edits, and coding components. Lockheed Martin provided a comment that touched on this recommendation:

> *A technical interface between form design, printing, and data capture was also extremely beneficial to the program's success and should be established very early in the program lifecycle. This worked well on DCS 2000, but could have been established even earlier. All three of these aspects of forms processing must work in concert with each other in order to maximize the productivity of the process as a whole.*

In keeping with the theme of a unified data capture system, TRW recommended "starting the system and services contracts at the same time so strong working relationships can be developed from the beginning." They further added that "Having both contractors work requirements together will result in a better system and better operational procedures" if operational perspectives are reflected in the requirements. TRW specifically noted that development of DCS 2000 was initiated without input from the users (i.e., the operations staff). They added that this was a source of frustration at the DCCs and required development of a management information system that was far more extensive than they had originally planned to implement.

The Assessment Report on Data Capture of Paper Questionnaires was in general agreement with the need for a unified system development environment. It recommended "integrated development" involving internal stakeholders early in the planning phase and the need for "better integration of forms design and printing specifications with data capture system development." It found that forms design was largely independent of data capture and processing system designs and therefore this was a factor that led to questionnaire designers being over confident in the capabilities of OMR and OCR technologies.

6.2 Define requirements early

As noted in Titan's Census 2000 Data Capture System Requirements Study, requirements establish the very foundation for a system. Their importance cannot be overstated, especially in an environment where a substantial R&D investment is necessary. Delays in defining requirements, or not fully defining them, increases the likelihood that the system will not meet data capture expectations or perform at the level required. Or, in the case of quality assurance requirements, waiting until late in the development cycle may not allow for sufficient time for implementation of the mechanisms needed to generate appropriate metrics or ensure adequate quality. Starting the planning and development earlier would provide a greater chance that all identified requirements will be implemented and that sufficient time will exist for testing and system refinement.

6.3 Develop quality assurance standards early

Given that requirements for QA were not well established, the recommendation is that more emphasis be placed on defining QA measurements, processes, and reports for the 2010 Census. TRW recommended that both the DCS 2000 and DCSC participants develop an integrated QA program. Additionally, as stated in the RITRC report, the Census Bureau should investigate new QA technologies to bring QA evaluation time closer to production time. In this regard, the Assessment Report on Data Capture of Paper Questionnaires pointed out the need for the data capture system to provide for real-time access to data for quality assurance review.

6.4 Focus on redesigning key data capture problem areas

As discussed in this report, segmentation errors account for many of the processing problems. While

problematic, this facet of forms design provides a potentially high payback area for future research. DCS 2000 image capture specifications prohibited the use of dark outlines surrounding OCR fields, but the use of a different background color with a higher contrast to white should alleviate the definition problem.

Another example of a significant problem area on the forms is the Hispanic and race question. Many of the participants in the Census 2000 Questionnaire Design Study agreed that this was an especially troublesome area. Although there were improvements in terms of the presentation and sequencing of these questions in 2000, there is still room for further improvements that could make them easier to understand and less prone to interpretation problems during the data capture process. More empirical research is needed on this particular design topic.

6.5 Limit the number of forms

RITRC observed that the number of forms grew to 75 for Census 2000 (only 26 were data captured by DCS 2000). They expressed concerns that, with separate form definition templates being created for each form and that each had to be tested, this proliferation of forms became "unwieldy." RITRC has suggested that the "80-20 rule"[6] might apply in regards to the number of forms that are used for the 2010 Census. That is, the use of fewer, more generic forms might prove to be cost effective from a data capture perspective.

There are two major issues to consider. First, the level of effort required to develop the templates has major time and cost implications for the Census Bureau. Second, since identical information is being asked across multiple forms, there is a risk that wording or phraseology may not be consistent across all forms. The recommendation is for the Census Bureau to consider combining forms, when possible, so a single form may serve more than one purpose. There is a risk inherent in over reliance on too many generic forms.

TRW noted that it encountered disruptions to operations due to the need to adapt to different types of forms. They also recommended the Census Bureau should minimize the number of form types and design them as early as possible so the processing implications can be understood prior to the DCSC developing its procedures.

6.6 Assess future role of automated data capture technology

The data capture evaluation cited two possibilities with regard to the future use of automated data capture and imaging technology in the decennial census. If it is seen as having a *supporting role*, it would be used primarily for rapidly capturing the clear and easy responses. In this scenario, traditional methods, although resource intensive, would still be used to capture especially difficult responses. On the other hand, automated technology could have a *dominant role* assuming that census forms were dramatically streamlined and the long form with its numerous write-in responses was no longer captured in the decennial census. If the long form is retained, the data capture evaluation asked if it is then worthwhile to put a high priority on improving the quality performance of the automated technology. The data capture evaluation suggested that this issue could be a research question for testing leading up to the 2006 Census Test.

With poor handwriting accounting for many errors, the data capture evaluation suggested giving consideration to reducing some write-in fields to check boxes, reducing the set of questions, or using more enumerators to get long form data.

6.7 Implement a unified Help Desk

In TRW's opinion, the existence of two different types of help desks (i.e., one for DCS 2000 and one for DCSC) "introduced unnecessary complexity, overlap and confusion." They recommend using only one integrated help desk.

6.8 Train on the actual working system

Given the size of the workforce for data capture operations (over 10,000 temporary employees during the data capture operations phase), TRW was concerned whether all employees had the necessary skills. TRW recognized the need for consistency in training across all sites. To provide quality training, TRW cited a key lesson learned: "From the start of the project, instructional design teams need as much access to the actual working system and equipment as possible to ensure accurate training documentation. This includes participation in dress rehearsals and OTDRs."

6.9 Expand the time for testing

As previously mentioned, the various tests were extremely beneficial and exercising the system provided invaluable insights. However, TRW was concerned that the OTDRs started too late in the

[6] Also referred to as Pareto's Law, this is the principle that 20 percent of something accounts for 80 percent of the results.

Census schedule (August 1999) and would like to see more time allotted between tests. Since managers from all four DCCs participated in each sequentially scheduled OTDR, their time was divided between participation in an OTDR at another site, preparing their own OTDR, and preparing their site for data capture operations. TRW noted that "preparing for an OTDR is a full-time project that could have been performed better if the participants were not busy at previous tests at the same time that they needed to be preparing for their own test."

6.10 Better define Decennial Management Information Systems for future contracts

TRW expressed concerns about the level of effort required to implement the DCSC Management Information System (DMIS) and the fact that it required more resources than anticipated. Their assessment was stated as follows:

> *The scope of DMIS was severely underestimated. Instead of a few COTS tools sitting on desktops, it grew into a medium sized networked information system. This was a difficult task to complete in less than two years. We should have scaled back the scope of DMIS given the time constraint. We were understaffed early in the process. It was difficult to attract developers to the project due to the lack of programming work and the short-term nature of this project.*

These concerns should be noted for future planning purposes and the scope of requirements for systems like DMIS must be fully understood. A management information system typically interfaces with numerous other systems and this increases the complexity of development and testing. The magnitude of effort and the time required to develop robust information systems essential to the management of large-scale programs needs to be factored into planning.

6.11 Provide more information on the scope of documentation requirements

As with DMIS, TRW had similar concerns about the amount of effort it took to produce management documentation. They provided the following perspective on this issue:

> *In retrospect, when the entire documentation effort is considered, it needs to be recognized that the importance of the documentation was not emphasized in the original RFP and therefore assumed to be less significant by TRW in our proposal. We are sure in hindsight that no one meant to minimize the effort, but it should be recognized that comprehensive technical and administrative procedures were essential to running consistent operations at the data capture centers. The operational procedures and assembly-line-like process necessary to handle the forms required all parties to know and understand what procedures to follow and understand what changes were made to the procedures during operations to ensure optimum productivity. It was also critical that the DCC administrative and facilities staffs have guidelines and procedures to follow to reduce employee-relations issues. And, certain documents were added to the effort and were critical to the program, but were not in Section F [of the contract]. These documents included DCSC Management Information System (DMIS) documentation and operations documentation such as the OCC CONOPS [Concept of Operations] and other OCC related documents.*

TRW recommended that documentation requirements be given more emphasis in future RFPs.

6.12 Minimize the use of keying from paper

The Assessment Report on Data Capture of Paper Questionnaires cites inefficiencies associated with KFP due to its cumbersome software and 100 percent validation requirement. With much of the need for KFP being generated by damaged forms, light pencil marks, and incomplete erasures, it may be possible to devise methods to significantly reduce the number of forms submitted for KFP

6.13 Produce real-time cost data

The Assessment Report on Data Capture of Paper Questionnaires addressed the need for current cost information. The combined costs for DCS 2000 and DCSC were approximately $520 million, and contract modifications accounted for significant increases over the original contract baselines. In spite of the magnitude of the data capture system, real-time labor expenses were not available, nor were cost data by operation for each DCC. To facilitate management of the 2010 Census, the report recommends using more detailed and current cost data.

This page intentionally left blank.

7. Author's Recommendations

7.1 Implement a more structured approach to defining requirements for the data capture system

More attention needs to be paid to defining a solid set of requirements, especially because the data capture system in 2010 may well incorporate additional functionality and have a more complex architecture. For instance, the Lockheed Martin Phase II Lessons Learned document speculated that:

> *Additions such as data entry from wireless and internet-based applications or complex address recognition can have significant positive effects on various aspects of the system and its operation. It may also be beneficial to look outside the current bounds of the system for opportunities to improve the overall Census capture operation. Such areas of interest include the integration of post processing coding systems or uses for two-dimensional bar codes. So, in order to take full advantage of all of the experiences of the 2000 Census, enhancements of these types should be investigated and possibly applied to future systems with similar requirements.*

This will require a pro-active, disciplined, and systematic approach to requirements definition. A major management commitment will be necessary to make this a reality. Without it, the Census Bureau will face a high degree of risk in implementing data capture systems and operations for the 2010 Census.

7.2 Research the potential for expanding the all digital data capture environment

In citing lessons learned, one of the recommendations put forward by RITRC suggested that "Due to the unique nature and scope of the census, it requires an early investment in research and development." Titan endorses RITRC's advice, but suggests that research and development efforts focus beyond paper-based systems. To be sure, the capabilities and sophistication of such systems have steadily improved, as evidenced by the DCS 2000 performance and quality issues discussed in this report. However, these systems still incur substantial labor costs for printing and processing forms and for nonresponse followup. The system also requires major capital investment in scanning technologies in order to process millions of forms. In short, paper based systems are still cumbersome because of the conversion from paper into images, and subsequent storage in electronic format. This is a very complex process.

Although the data capture system was unquestionably successful during Census 2000, we have transitioned into an era when information is predominately collected, stored, transformed, and transmitted in a digital format. In theory, future census activities could be performed in an all-digital environment. In fact, the Census Bureau is already moving toward expanding the use of digital media having made a commitment to use mobile computing devices in data collection operations involving enumerators. And, the feasibility of a Web-based census data collection system was proven during Census 2000 and will provide yet another means of collecting data digitally in the next decennial census. This is especially true if the Census proactively markets an online option for the 2010 Census. Numerous electronic data files being maintained by state and federal agencies may provide more sources of personal data that can be conveniently extracted for nonrespondents.

An ambitious goal for the Census Bureau would be to eliminate the need for processing massive quantities of paper forms by planning for an all digital data capture environment. Advanced technologies on the horizon are introducing new forms of digital media, some of which are in the early stages of development. These efforts may identify enabling technologies that could further streamline census data capture operations. This may not be achievable in time for the 2010 Census, but research can help pave the way for the transition into the digital world.

This page intentionally left blank.

8. Conclusion

The Census 2000 data capture system was a very complex, well managed system that employed advanced technologies to automate data capture processes. It captured massive amounts of data within a relatively short period of time with high accuracy rates. However effective the automation, it still relied on some manual processing to handle millions of write-in fields that posed a problem for the automated character and mark recognition subsystems. The data capture evaluation raises an interesting issue: is it worth attempting to make further refinements to the automated capture processes in order to capture the problematic responses? This is something that the Census Bureau may want to study in order to determine whether the extra cost and effort could be justified. Alternatively, improved forms design may allow more data to be captured with fewer errors. Many of the problematic types of responses were identified in the documentation reviewed for this report and are well understood by Census Bureau staff. Perhaps some of the design improvements listed in this report could help to minimize these types of errors from occurring in the future.

A goal of this report was to look for common themes in the documentation as well as any conflicting perceptions about the data capture system. No significant differences of opinion were detected during the review. To the contrary, there were many similar, complimentary views and findings that reinforced our confidence in stating the issues presented in this report.

This page intentionally left blank.

References

Brinson, A. and Fowler, C. (2003) "Assessment Report for Data Capture of Paper Questionnaires," February 19, 2003.

Conklin, J. (2003) "Evaluation of the Quality of the Data Capture System and the Impact of the Data Capture Mode on the Data Quality, (K.1.b)," Final Report, March 12, 2003.

Brinson, A. and Fowler, C. (2001) "Program Master Plan, Data Capture Systems and Operations," March 30, 2001.

Forms Design and Printing Office. (2002) "Forms Design and Printing Lessons Learned," May 30, 2002.

Lockheed Martin. (2001a) "Phase II Lessons Learned," February 21, 2001. (Includes Appendix A listed below as a separate white paper.)

Lockheed Martin. (2001b) "Appendix A, Technical Lessons Learned," White Paper, Release 1 Version 1, February 23, 2001.

Rochester Institute of Technology. (2002) "DCS 2000 Data Quality, v.2.1 Final," September 20, 2002.

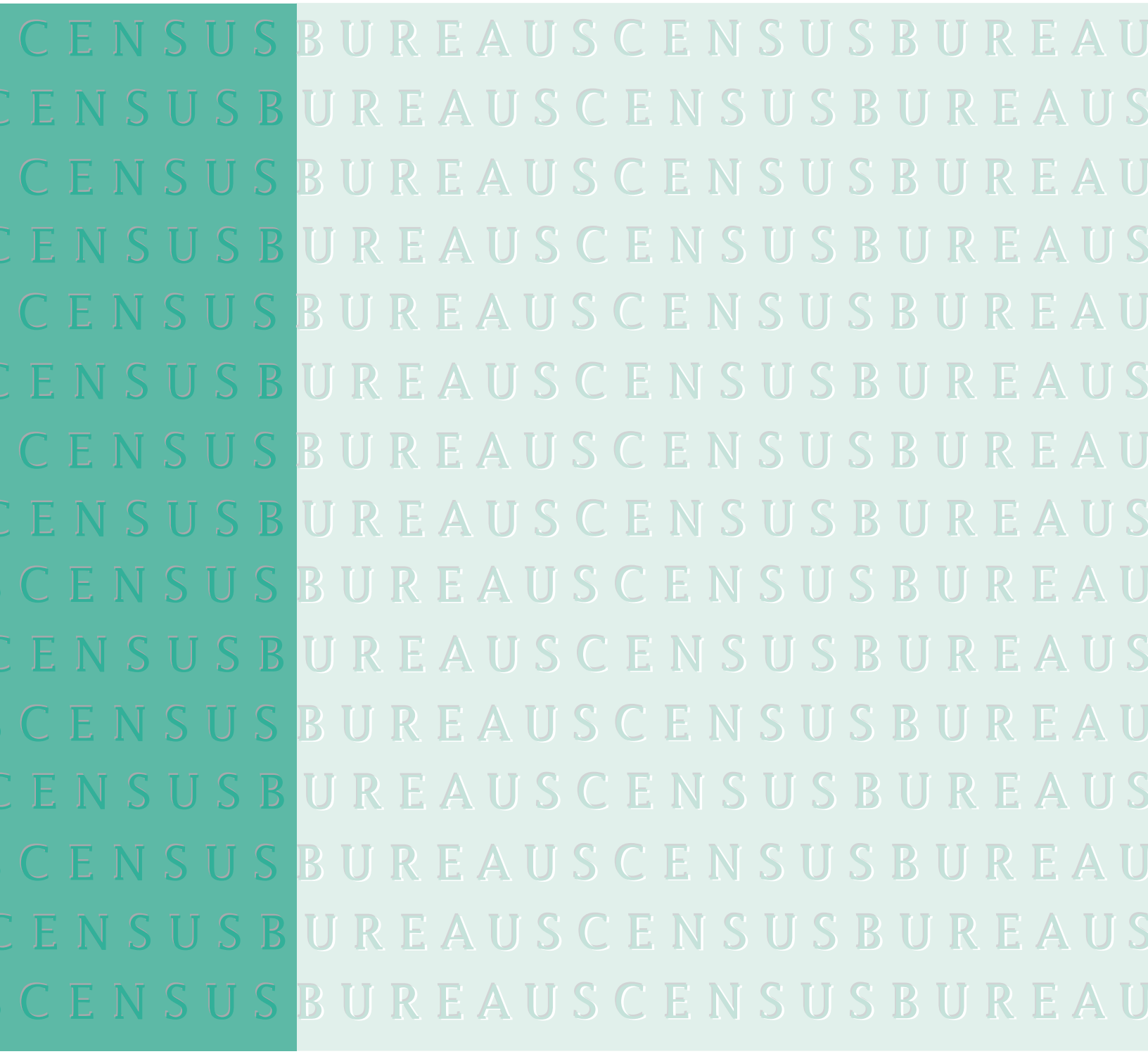
Titan Systems Corporation/System Resource Division, (2002) "Census 2000 Data Capture System Requirements Study (R.3.d)," Final Report, August 23, 2002.

Titan Systems Corporation/Civil Government Services Group. (2003) "Census 2000 Questionnaire Design Study," Final Report, March 19, 2003.

TRW. (2001) "Data Capture Services Contract Final Report," February 9, 2001.

Weinberg, D. (2000) Memorandum from Daniel H. Weinberg, Subject: Actions to Correct Pass 2 Keying Errors in Census Sample Monetary Fields, December 7, 2000.

This page intentionally left blank.


CENSUS BUREAUSCENSUSBUREAU
CENSUSBUREAUSCENSUSBUREAUS
CENSUS BUREAUSCENSUSBUREAU
CENSUSBUREAUSCENSUSBUREAUS
CENSUS BUREAUSCENSUSBUREAU
CENSUSBUREAUSCENSUSBUREAUS
CENSUS BUREAUSCENSUSBUREAU
CENSUSBUREAUSCENSUSBUREAUS
CENSUS BUREAUSCENSUSBUREAU
CENSUSBUREAUSCENSUSBUREAUS
CENSUS BUREAUSCENSUSBUREAU
CENSUSBUREAUSCENSUSBUREAUS
CENSUS BUREAUSCENSUSBUREAU