STUDY SERIES
(Survey Methodology #2010-04)

Data Reliability Indicator Based on the Coefficient of Variation: Results from the Second Round of Testing

Kathleen T. Ashenfelter

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: April 1, 2010

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

USCENSUSBUREAU

Statistical Research Division
Usability Laboratory
Washington, D.C. 20233

Date: March 9, 2010

To: DSSD Data Reliability Indicator Team: Anthony Tersine, Michael Springer, and Jennifer Tancreto

From: Kathleen T. Ashenfelter, SRD Usability Laboratory

Subject: Data Reliability Indicator Based on the Coefficient of Variation: Results from the Second Round of Testing

1 Abstract

This study was the second round of usability testing for the Data Reliability Indicator for American Community Survey (ACS) data tables proposed by the sponsor team. Three prototype tables with a color-coded indicator based on an estimate's coefficient of variation were compared to a baseline table that represented a past production version of an ACS data profile. One prototype had three levels of reliability and the other two had four levels of reliability. One version of the four-level table was labeled with the terms "excellent," "good," "fair," and "poor." The other was labeled with the terms "reliable," "mostly reliable," "somewhat reliable," and "unreliable."

No differences were found in users' accuracy and efficiency (time on task) across the four tested table designs. Overall, 19 of the 21 participants indicated that they preferred the prototype tables over the baseline tables, although their preference for either the three–(n=10) or four–level (n=9) tables or either of the two alternative wording options was fairly evenly split (nine preferred the "good" wording, ten preferred the "reliable" wording, and two had no preference). However, there were some significant differences in satisfaction scores between the tables. There is evidence that satisfaction as measured by the QUIS instrument was significantly higher for both the baseline and the four–level tables than for the three–level table.

An analysis of how likely participants were to mention the margin of error (MOE) or report it along with the estimate showed no significant differences by table design. Eye-tracking heat maps showed that participants looked at the MOE column on the table fewer times for the three–level condition.

Participants did use the reliability indicator and frequently said that they would report the message contained in the reliability column along with the estimate (12% indicator reporting for the three–level condition and 43% reporting for each of the four–level conditions). Participants in both the four–level "good" table and the four–level "reliable" table conditions were significantly more likely to report the message from the color–coded reliability indicator than the three-level indicator in a post–hoc test. Also, participants in both the four–level "good" table and the four–level "reliable" table conditions were significantly more likely to explicitly report the message from the color–coded reliability indicator along with the estimate than participants in the three-level indicator condition. There was no significant difference between the two four–level tables themselves on this variable.

More detailed results and potential usability issues are discussed.

Key Words: **data reliability indicator, coefficient of variation, color-coded data tables, usability**

2 Introduction

This second round of testing took a more empirical approach to the evaluation of these prototypes and builds on the more exploratory method used in the first round of testing (Ashenfelter, Beck, & Murphy, 2009). While internal Census Bureau employees were the participants for the first round of testing, a group of Census–external ACS data users (i.e., Washington, D.C. area researchers, federal employees, and journalists) were recruited as participants for this round of testing. Findings from this second round of testing will inform the design-and-development team on areas of user satisfaction and success as well as areas where the participants struggled while accessing and using the data.

2.1 Background

This project aimed to address an issue that arises with the ACS data tables because the estimates have varying levels of reliability. Some of the data, especially some single-year estimates, have high coefficients of variation (CVs). Some users may use these estimates without taking into account their reliability. The goal of this project is to provide some guidance to help data users more easily detect when there are potential reliability issues as measured by the CV (although the decision of whether or not to use the estimate is ultimately that of the data user).

The proposed method for addressing the issue of the reliability of the estimate was to color-code each estimate with the appropriate level of reliability along with an associated word (e.g., "reliable" or "unreliable"), as measured by the coefficient of variation (Whitford & Weinberg, 2008). The choice of CV as the estimate of sampling error to be tested was based on the goal to produce a standardized measure of reliability that might be easier for users to interpret. There has also been the observation that, although the margin of error (MOE) is currently provided with each estimate, ACS data users routinely ignore the MOE. Another reason for using the CV as a metric for reliability is that there are published

Census Bureau standards for data reliability based on the CV. The existing data reliability standard, Quality Requirements for Releasing Data Products (Cahoon et al., 2007), states (Page 7 section 1.B.): If the estimated coefficients of variation (CV) for key statistics are larger than 30 percent, the data product will be released under the requirements for category 2 or category 3 data [1].

As a starting point, a four-level categorization based on this documented Census 0.30 standard was proposed by the sponsoring team in the Decennial Statistical Studies Division (DSSD). The idea was to color-code the estimate according to its reliability, as evaluated by its associated CV. For both the first and second round of usability testing, a red color indicates a low-reliability estimate and green indicates a reliable estimate. Mid-range reliability is indicated by yellow or orange coding. The prototypes and baseline table that were tested in this second-round evaluation of the ACS data reliability indicators are included in Appendix A.

The tasks that participants completed for the second round of testing are provided in this test plan as Appendix B. These tasks were kept as similar as possible to those used in the first round of usability testing, but they were updated to incorporate findings from the first round of testing as well as feedback from team members and the Census Methodology and Standards council. A more objective method of assigning a task-difficulty rating was also used for this round of testing. Information about the task difficulty metric used can be found in Appendix B.

The following are key differences between the first and second round of testing:

- **Revised tasks -** The tasks were worded to avoid potentially biasing participants toward using color-coding as opposed to MOE or CV; additional tasks were added to the protocol, and existing tasks were re-written to be more realistic and challenging for participants.
- **Revised prototypes -** Only three- and four-level indicator prototypes were tested this round due to the poor testing performance of the two-level indicator in the first round of testing (although a baseline table was still used for comparison). The indicator legend box explaining the reliability column was also moved to a spot directly above the table. The three–level indicator prototype used the wording “good” (green; $CV < .30$), “fair” (yellow; $.30 \leq CV < 0.61$), and “poor” (red; $CV \geq .61$). This indicator was similar to the three–level table evaluated in the first round of testing. Two different versions of the four–level table were tested with alternative wording of the text message included in the reliability indicator column. One version, referenced in this report as the four–level “good” indicator (since this is how participants frequently referred to it), included the messages “excellent” (green; $CV < .10$), “good” (yellow; $.10 \leq CV < .30$), “fair”

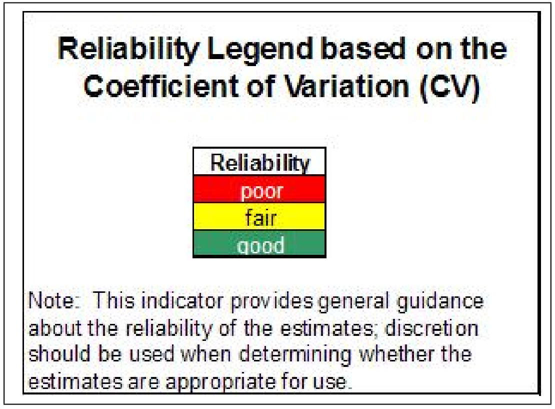
[1] Category 2 and 3 data are defined in the Quality Requirements for Releasing Data Products document: “Data in the first category satisfy the Public Data Release Criteria. Category 2 consists of data that do not satisfy these criteria but have published release dates. Data in the second category will always be released. Category 3 consists of any other data that do not satisfy the criteria.”

(orange; $.30 \leq \text{CV} < 0.61$), and "poor" (red; $.61 \leq \text{CV} < 1.0$). This indicator prototype was similar to the four–level prototype that was evaluated in the first round of usability testing. The second four–level indicator tested included the same cutoff levels for CV, but with different labels referring to the reliability of the data [2]: "reliable" (green), "mostly reliable" (yellow), "less reliable" (orange), and "unreliable" (red). Figures 1, 2, and 3 show the data reliability legend from the three–level, four–level "good", and four–level "reliable" indicators, respectively. These prototype tables were compared to each other and to a "baseline" table, which represented what the 2006 ACS data profile tables looked like, which are similar to current versions of (2008) ACS detailed data tables. Beginning in 2007, the ACS data profiles were redesigned to add percentages next to the estimates in addition to the MOE, so the baseline tables tested here must now be called "previous" versions of the ACS tables instead of "current."

- **Different participant pool-** While participants for the first round of testing were internal Census employees, participants for the second round consisted of real ACS data users from outside the Census Bureau. Some participants were recruited from a list of D.C.–area journalists who use ACS data products, which was provided by the Census Bureau Public Information Office (PIO). Real ACS data users were also recruited through a local nonprofit research institution and the Council of Professional Associations on Federal Statistics (COPAFS), as well as from an ACS Interagency list. Also, since one of the participants from the first round of testing was colorblind and some participants randomly selected from the sampling frame were likely to be colorblind, all participants were asked to report the status of their color vision. However, none of the participants in this round of testing were colorblind (determined by self–report and color–blindness test).

[2] The term "reliable" only acknowledges the reliability from a sampling standpoint

Figure 1. Data Reliability Indicator Legend from Three–Level Prototype

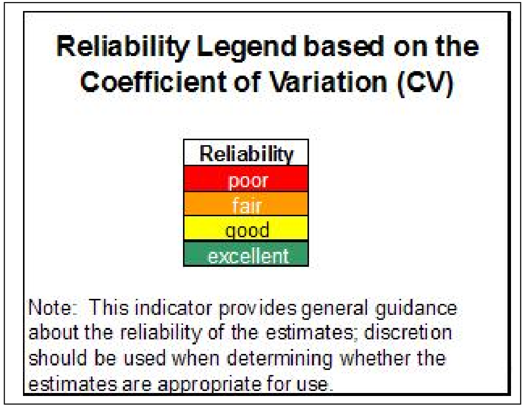


Reliability Legend based on the Coefficient of Variation (CV)

Reliability
poor
fair
good

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

Figure 2. Data Reliability Indicator Legend from Four–Level "Good" Prototype

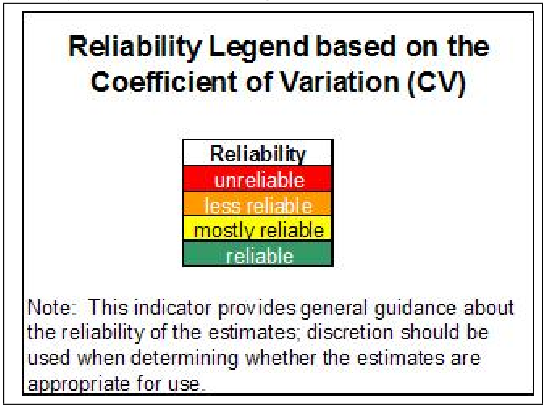


Reliability Legend based on the Coefficient of Variation (CV)

Reliability
poor
fair
good
excellent

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

Figure 3. Data Reliability Indicator Legend from Four–Level "Reliable" Prototype



Reliability Legend based on the Coefficient of Variation (CV)

Reliability
unreliable
less reliable
mostly reliable
reliable

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

2.2 Research Goals

The usability goals for this study were defined in three categories: user accuracy, efficiency, and satisfaction.

Goal 1: To achieve a high level of accuracy in completing the given tasks using the data tables. It was decided that the user should be able to complete 75% of the tasks successfully. The goal for the first round of testing was set at 70% accuracy, but this goal was exceeded by all of the participants. Since the average accuracy rate was 83% in the first round of testing, the goal was raised for the second round (especially considering that the participating journalists and other ACS data users are likely to be familiar with ACS data). The results showed that all three tables had over 85% accuracy scores, so all four tables passed this usability goal. A related sub–goal was to evaluate whether the color–coded data reliability indicator would prompt users to pay attention to and report an estimate's reliability. The results showed that 12% of participants would report the estimate's reliability for the three–level table, 43% would report it for the four–level "good" table, and 43% would report it for the four–level "reliable" table. This indicates that participants were often attending to the reliability column and reporting it along with their answer to a task.

Goal 2: To achieve a high level of efficiency in using the data tables. It was decided that the test participants should be able to complete the tasks in an efficient manner taking no longer than 3 minutes for a harder task, 2 minutes for a medium task, and 1 minute for an easier task. While the designation of tasks as "medium" or "hard" was determined in team meetings for the first round of testing, a more objective metric based on empirical research on cognitive workload was used for this round of testing (see Appendix B). The difficulty rating assigned to each task can also be found in Appendix B. All tasks were rated either medium or hard (there were no easy tasks). The average time for completion of medium tasks was 1.61 minutes, so this passes the usability efficiency goal for medium tasks. The average completion time for hard tasks was 3.53 minutes, so this did not pass the usability efficiency goal set before testing.

Goal 3: For the users to experience a moderate to high level of satisfaction from their experience with the data tables. A tailored version of the University of Maryland's Questionnaire for User Interaction Satisfaction (QUIS) (Chin, Diehl, & Norman, 1988) was implemented. The overall mean of the QUIS ratings for the data tables should be above the mean (above 5 on a nine-point scale, where 1 is the lowest rating and 9 is the highest rating). The same should hold true for the individual QUIS items. Each table (including the baseline table) had an overall QUIS score of over 6.11, so this passes the usability QUIS goal. Additionally, each individual item had an overall rating of at least 5.75, so this is also consistent with this usability goal.

2.3 Scope

A specific set of user interactions with the tables (as portrayed in the prototypes provided by the sponsor) was within the scope of the usability evaluation. The user interface was not tested for compliance with the Section 508 regulations, although members of of the Systems Support Division (SSD) did consult with the usability and sponsor team about potential accessibility issues associated with color–coding data tables before the first round of usability testing took place. Since these may become data tables that could potentially be accessed through a government Web site, they must comply with Section 508 regulations before the Web site becomes available, unless a waiver is granted.

2.4 Assumptions

- Participants had at least one year of prior Internet and computer experience.
- Participants had prior knowledge of how to navigate a Web site.
- Participants had some prior experience in using ACS data products.
- Participants had no known disabilities, but were screened for color blindness.

3 Method

3.1 Participants

The original goal for this study was to recruit sixty participants from the metro Washington, D.C. area from a list of local data users and local ACS data users to come to the SRD Usability Laboratory in Suitland, MD for testing. However, the usability staff had some difficulty with recruiting participants and only 21 people participated in the study. The SRD Usability staff recommends that recruitment for the next round of testing coincide with an on–site conference of ACS data users, since one main problem was a general unwillingness for participants to travel to Suitland, MD for the testing (e.g., the State Data Centers (SDC) annual conference or the Census Information Centers (CIC) and SDC annual joint conference). The usability staff will also investigate the possibility of off–site testing for future research.

Initially, sixty people were randomly sampled from a sampling frame of 342 local Washington, D.C.-area journalists provided by the Census Bureau's Public Information Office (PIO). From this sample of sixty journalists, the expected response rate was forty participants. However, due to a low response rate and a large number of journalists being listed with out–of–date contact information, an invitation to participate was extended to the remainder of the list. However, only six journalists participated. Additional real ACS data users were recruited through the a local nonprofit research institution and the Council of Professional Associations on Federal Statistics (COPAFS). Six people participated in the testing from these institutions. Participants were also recruited from an ACS Interagency list obtained from the American Community Survey Office (ACSO). Nine people contacted through this list participated. Because of the difficulty recruiting participants, it was not possible to stratify their random assignment to condition. However, each participant was assigned to a condition using the SAS Proc Plan function. They were assigned to one of four conditions corresponding to one of the three prototypical ACS data reliability indicators or the baseline table. Each version of the table had at least 5 participants who used it to complete the testing. The order in which the participants performed the tasks was also randomized for each person. Table 1 lists the participants' occupations, the first task they performed, and the table condition to which they were randomly assigned. Six participants were journalists, six were researchers (e.g., nonprofit institution, COPAFS, or other), and 9 were federal employees recruited through the ACS Interagency list. Twelve of the participants were male and nine were female.

Each test participant had at least one year of prior experience in navigating different Web sites. Participants varied in their levels of familiarity with the ACS and ACS data tables, but all used ACS data products at least occasionally for their jobs. The amount of time that participants reported that they have been using ACS data products or tabulations based on them ranged from two years to the very beginning of the ACS. The average age of the participants was 45, with a minimum of 27 and a maximum of 71.

Observers from the Decennial Statistical Studies Division (DSSD) Data Reliability Indicator team, the Math Stat Council, and ACSO were invited to watch the usability tests on television screens in a separate room from the test participant and test administrator. At the end of each test session, the test administrator and observer(s) discussed the findings from that session.

Table 1. Participant Condition, First Task Performed, Gender, and Occupation

Participant	Table Seen	First Task	Gender	Occupation
P1	3-Level	4	Male	Nonprofit
P2	4-Level Good	5	Female	Nonprofit
P3	3-Level	7	Male	Journalist
P4	4-Level Reliable	9	Male	Federal
P5	4-Level Reliable	1	Female	Journalist
P6	Baseline	3	Male	Journalist
P7	4-Level Good	5	Female	Nonprofit
P8	Baseline	1	Female	Federal
P9	Baseline	3	Female	Journalist
P10	3-Level	5	Female	Journalist
P11	4-Level Reliable	6	Male	COPAFS
P12	4-Level Good	8	Male	Federal
P13	4-Level Good	1	Male	Journalist
P14	Baseline	2	Female	Federal
P15	4-Level Reliable	4	Male	Nonprofit
P16	3-Level	6	Male	Federal
P17	3-Level	7	Female	Federal
P18	Baseline	9	Male	Federal
P19	4-Level Reliable	2	Female	Federal
P20	4-Level Good	3	Male	COPAFS
P21	3-Level	5	Male	Federal

The assignment of participants to condition did result in some conditions having different types of participants (e.g., more federal employees vs. journalists, etc). This limits the generalizability of the results of the analyses.

3.2 Facilities and Equipment

Testing Facilities

The test participant sat in a small room (5K512), facing a one-way glass and a wall camera, in front of an LCD monitor equipped with an eye-tracking machine that is placed on a table at standard desktop height. The test participant and test administrator were in the same room for the reading of the general protocol, the think–aloud practice, and eye–tracking calibration. The test administrator then went into the control room for the usability testing segment of the session and returned to sit in the same room as the participant for the debriefing segment.

Computing Environment

The participant's workstation consisted of a Dell personal computer, a 21-inch Tobii LCD monitor (Tobii model 2150) equipped with cameras for eye tracking, a standard keyboard, and a standard mouse with a wheel. The operating system was Windows XP for all participants.

Audio and Video Recording

Video of the application on the test participant's monitor was fed through a PC Video Hyperconverter Gold Scan Converter, mixed in a picture-in-picture format with the camera video, and recorded via a Sony DSR-20 digital Videocassette Recorder on 124-minute, Sony PDV metal-evaporated digital videocassette tape. Audio for the videotape was picked up from one desk and one ceiling microphone near the test participant. The audio sources are mixed in a Shure audio system, eliminating feedback, and fed to the videocassette recorder.

Eye–Tracking

The participant's eye movements were recorded during the usability test using a trial version of Tobii Studio Enterprise Edition (Tobii Technology, 2008). Some of the participants' data were collected using the older Tobii Clearview software after the trial license for Tobii Studio expired on July 4, 2009. The Tobii eye-tracking device monitors the participant's eye movements and records eye-gaze data. The data recorded represent the physical position of the eye as measured by the the reflection of a near–infrared beam off of the pupil. The horizontal and physical position of the pupil are recorded for both eyes at a rate of 50 Hz (e.g., 50 samples per second) on this model of eye tracker. This type of eye-tracking requires the calibration of each eye. Data collected from the eye-tracking device includes eye-gaze position, timing for each data point, eye position, and areas of interest. The Tobii eye tracker records data at a rate of 50 Hz. When a participant looks away or blinks, or if the eye tracker loses track of the participant's pupil, this data is recorded as missing data and this does not stop the data recording. Often, the eye tracker will regain tracking status of the participant's pupil and data recording will begin again within a few seconds following a glance away from the computer screen.

3.3 Materials

Usability testing requires the use of various testing materials. Testing materials included the following items provided in the appendices. There were three different prototypes corresponding to different possible ways of displaying the data reliability indicator (two 4-level and one 3-level indicator). There was one baseline table used to represent ACS data tables without data reliability indicators. Versions of these prototypes and the baseline are available in Appendix A. Following the initial probe item (i.e., "What is the first thing that that you noticed about this table?"), the tasks for each prototype were the same and were presented in a randomized order for each participant.

Only the small geography of Hays City, Kansas, was used for this round of testing, which differs from the three geographical pairs used in the first round of testing. Pairs of geographical locations were used in the first round of testing to contrast the difference in data reliability based on the CV (i.e., California is an area with a large population, so all of the estimates will be highly reliable, but Wilmington, Delaware, is much smaller and will have less reliable estimates). The smaller geographies have more variability in the reliability of their estimates and will require some judgment on the part of the data user as to whether to use the estimates for the task at hand. Hays City, Kansas was selected because its estimates had a wide range of reliability and each level of the data reliability indicator was represented for both the three– and four–level indicators.

Prototypes

The two-level indicator was eliminated based on the first round of usability testing. Two versions of the four-level indicator and one version of the three-level indicator were tested in this second-round investigation. These three prototypes and the baseline (previous ACS data table) can be found in Appendix A. The images have been truncated for legibility.

Tasks

Members of the ACS data-reliability indicators team and members of the Census Bureau's Usability Lab created the tasks with input from the Math Stat council. The tasks are designed to capture the participant's interaction with, and reactions to, the design and functionality of the ACS data reliability indicators. The first question asked of the participants is not a task in the traditional sense because it simply asks them to report the first thing that they notice about the tables, so it is called the "initial probe" question and is not considered an official task. The rest of the tasks were designed so that the participant would look for estimates that were located in different areas of the table. The tasks themselves were randomized for each participant. Table 1 lists the first task that each participant performed based on the randomized order of tasks they each received. Appendix B provides the version of the tasks that were used in this second round of testing [3]

[3] The wording in task 9 was changed from "find out how many *people* are 18 or older in your hometown of Hays, KS" to "find out how many *civilians* are 18 or older" after many participants struggled with this

General Protocol

Each participant was read a general protocol, which can be found in Appendix C. The test administrator read some background material and explained several key points about the session. The general protocol emphasizes that the participant's skills and abilities are not being tested, but that the participant is helping in an evaluation of the data table's overall usability.

Consent Form

Prior to beginning the usability test, the test participants completed a general consent form supplied in Appendix D. The consent form documents the participant's agreement to permit videotaping of the testing session and states that the study is authorized under Title 13 of the U.S. Code.

Questionnaire on Statistical Experience, Computer Use and Internet Experience

Prior to the usability test, the test participant completed this questionnaire, which gathered information on the participant's demographics, experience using statistics, computer use, and Internet experience (Appendix E). This information helped us determine whether there is a relationship between these three experience factors and performance and preference scores found during testing.

Future research will examine whether there is a relationship between experience, expertise, and the difficulty rating that participants assign to the tasks in the task-difficulty rating questionnaire (see below).

Questionnaire for User Interaction Satisfaction (QUIS)

The original version of the QUIS includes dozens of items related to user satisfaction with a user interface (Chin et al., 1988). In a usability test at the Census Bureau, SRD typically uses 10 to 12 items that the usability team has tailored to the particular user interface being evaluated. This study used a modified version that includes items worded for the ACS data-reliability indicators context (Appendix F). The experimenter handed the QUIS to the participant at the same time as the task-difficulty rating questionnaire (below).

Task-Difficulty Rating Questionnaire

Participants were asked to provide a difficulty rating for each task, which was used for validation of the "medium" versus "hard" designation during analysis. This short survey can be found in Appendix G.

task.

Debriefing Questions

After completing the tasks, the experimenter read aloud debriefing questions to the participants about their overall experience using the prototype ACS Data Reliability Indicator (Appendix H). The debriefing questions included an inquiry about each participant's color vision, followed by a brief Ishihara test of colorblindness. These questions are included in the debriefing segment of the protocol following testing and not included in the survey administered to the participants before testing so as not to prime them to focus intentionally on color during testing.

Procedure

Each test participant was escorted to the usability lab at the U.S. Census Bureau headquarters building in Suitland, Maryland. Upon arriving, the test participant was seated with the test administrator in the testing room (5K512). The test administrator greeted the participant, thanked him or her for his or her time, and read the general introduction. Next, the participant read and signed the consent form. After signing the consent form, the test participant completed the questionnaire on demographics, experience with statistics, computer use and Internet experience.

Since this test used the eye-tracking device, the participant's eyes were calibrated after the general protocol was read and the consent form was signed. Calibration was usually completed in about fifteen to twenty seconds by having the participant look at a dot moving across the computer screen. Once calibration was completed, the test administrator exited the room and continued the testing process from the control room (5K509).

Following calibration, the participant began to complete the tasks on the ACS data reliability indicators prototype. At the start of each task, the participant read the task aloud. While completing the task, the participants were encouraged to think aloud and share what they were thinking about the task. This interaction was not intended to be a conversation. If at any time the participant became quiet, the test administrator probed the participant about what they were looking for in the table. The content of the so-called "think-aloud" protocol allows us to gain a greater understanding on how the participant is completing the task and to identify issues with the tables. In order to make sure that the participants understood what was expected by the instruction to think aloud, they engaged in a practice think-aloud task where they walk through their thought process while performing a task using a commonly accessed Web page (the end of Appendix C).

At the conclusion of each task, the participant stated a "final answer" to the task. During the task or while watching the tapes of the sessions at a later time, the test administrator noted any observable struggles or other noteworthy behaviors, including comments and body language. After the participant completes all tasks, the eye-tracking device was stopped, the test administrator returned to the testing room, and the video recording continued. The test participant then completed the modified QUIS and task-difficulty rating questionnaire silently. When the participant completed the two paper forms, the test administrated asked the participant a series of debriefing questions (Appendix H). At the conclusion of the usability evaluation, the video recording was stopped. Overall, the usability session ran between 45 and 60 minutes.

4 Results

4.1 Accuracy

The number of participants for this study was small (5-6 per condition for a total of 21 participants), which means that the statistical anaylses had low statistical power. The small number of participants should be taken into account before generalizing to any population. The accuracy score was calculated by scoring whether the participant found the correct estimate in the table (1) or not (0). The issue of whether the participant would report the estimate, margin of error, or color–coded indicator message where applicable was scored separately. The initial screening probe question (What is the first thing you noticed about this table?) was not scored for accuracy. For task 3, which asked participants to find the number of people of German ancestry, the number of people of Slovak ancestry, and then decide which estimate had better data quality, the average was taken for these three sub–questions to compute an accuracy score for each participant. Similarly, for task 4, which had asked for three different estimate in parts a, b, and c, the average of these three parts was calculated. Part d was a subjective decision about whether to hold the concert in Hays, KS or not and was not scored for accuracy. The sample size is the number of task scores available for each table (e.g., 9 tasks per participant for 5 participants would be a sample size of 45). Table 2 shows the overall percentage of correct responses for each table. A one–way ANOVA showed that there was no significant difference among the tables in terms of the accuracy scores ($\alpha = 0.05, F(3, 185) = 1.1, p > 0.05$. They are all equally accurate.

Table 2. Accuracy Results by Table

Participant	Sample Size	Percent Correct
Baseline	45	92.6
Three–Level	54	94.4
Four–Level “Good”	45	93.3
Four–Level “Reliable”	45	85.2

Table 3 lists the accuracy score results by task number. Task 9 had the lowest accuracy score by far, and several participants commented on its difficulty. A sample size of 21 reflects the number of participants who completed each task.

One-way ANOVAs (across all four table types) were conducted to check for the possible influence of differing participant experience and educational levels. No significant differences in accuracy were found with education, how long the participant has been using ACS products, how often the participant uses ACS products, number of statistics courses taken, or self-rated level of expertise with statistics as independent variables ($p > 0.05$).

Table 3. Accuracy Results by Task

Task	Sample Size	Percent Correct
1	21	81.0
2	21	95.2
3	21	100
4	21	95.2
5	21	100
6	21	95.2
7	21	100
8	21	100
9	21	57.1

Mentioning and Reporting the Estimate, MOE, and Indicator Message: Participant's Judgment

The following tasks asked participants whether they would report or use the estimate they found in the table or whether they would include a particular group of people in a category based on the information in the table. Different wording was used when writing the tasks so that these questions about the consideration of sampling error did not stand out as being too similar to the participants. Their responses were scored as 1) whether or not they would report the estimate and 2) whether this decision is consistent with the appropriateness of reporting this estimate according to the reliability of the data and the context given in the task scenario (e.g., what is at stake based on the hypothetical decision). The responses were also coded according to whether they mentioned the MOE and the color–coded reliability indicator color or message and, if the participants said they would report the estimate, whether they explicitly stated that they would report the MOE or indicator message along with the estimate.

Reporting the Estimate

Whether or not the participants would report an estimate was a question of interest for several of the tasks. Specifically, the following tasks were designed so that participants' decisions about reporting the estimates could be compared to the reliability information associated with them. In most cases, there were only 5 participants per condition, so 20% would mean "1 out of 5."

Task 1: Your supervisor asks you to find some information about the number of women ages 15 to 50 who gave birth in the past 12 months for your hometown of Hays, KS. What information would you report to your supervisor?

The first part of the question asks for the estimate itself, which was scored in the accuracy portion of this study. However, whether or not the participants would report the estimate to the supervisor in this vignette is a separate issue. For this question, the correct estimate is 307, and the MOE is ±127. For the three–level table, the reliability column cell was coded green with the message “good” and for the four–level tables, the reliability column cell was coded yellow with the message “good” or “mostly reliable.”

Of the participants that found the correct estimate, for the baseline table, 60% of participants would report this estimate, 75% would report it for the three–level table, 100% would report it for the four–level “good” table, and 67% would report it for the four–level “reliable” table. The majority of participants would report these estimates for all of the tables.

Task 2: You are researching background information for a paper and need to find the number of people of West Indian descent in Hays, KS. What do you report in the paper based on your findings in the tables?

For this question, the correct estimate is 13, and the MOE is ±25. For both the three–level tables, the reliability column cell was coded red with the message “poor” for the three–level table and “poor” or “unreliable” for the four–level tables.

The results show that for the baseline table, 75% of participants who found the correct estimate would report this estimate, 50% would report it for the three–level table, 40% would report it for the four–level “good” table, and 20% would report it for the four–level “reliable” table. Although this estimate is unreliable due to a very large CV, three quarters of the participants in the baseline condition would use it. This may indicate that the data reliability indicator may be dissuading participants from using the unreliable estimate for this task. This issue will be further examined in the third round of testing.

Task 5: You are asked to report to state leaders the number of people of Italian descent living in Hays. What answer would you give them?

For this question, the correct estimate is 155, and the MOE is ±145. For the three–level table, the reliability column is coded yellow with the message “fair.” For the four–level tables, the reliability column is coded orange with the message “fair” or “less reliable.”

The results show that for the baseline table, 60% of participants who found the correct estimate would report this estimate, 83% would report it for the three–level table, 60% would report it for the four–level “good” table, and 100% would report it for the four–level “reliable” table.

Task 6: The mayor of Hays said that if there are more than 300 people ages 5 to 15 with disabilities in Hays, the city might be eligible to receive some government funding to develop programs for the disabled. He asks you if the there are at least 300 people in this age group with disabilities in Hays. What would you tell him using ACS data?

For this question, the correct estimate is 229, and the MOE is ±184. For the three–level table, the reliability column is coded yellow with the message “fair.” For the four–level table, the reliability column is coded orange with the message “fair” or “less reliable.”

The results show that for the baseline table, 60% of participants would report this estimate, 60% would report it for the three–level table, 40% would report it for the four–level “good” table, and 20% would report it for the four–level “reliable” table.

Task 7: The Danish embassy wants a listing of all cities with more than 200 people of Danish descent. Would you include the city of Hays based on the ACS data?

For this question, the correct estimate is 69, and the MOE is ±111. For the three–level table, the reliability column is coded red with the message “poor.” For the four–level table, the reliability column is coded red with the message “poor” or “unreliable.”

The results show that for the baseline table, 0% of participants who found the correct estimate would report this estimate, 0% would report it for the three–level table, 40% would report it for the four–level “good” table, and 0% would report it for the four–level “reliable” table. Less than half of the participants in the four–level table would report the estimate, and no one in the other three tables would report it. Overall, the participants decided not to use the estimate. Some commented that since the embassy was probably going to use the information to make an important decision that it was not a good idea to base it on that estimate.

Task 8: Cities with less than 200 people of French Canadian descent will engage in an outreach program designed to attract more people of French Canadian descent. Does Hays qualify based on ACS data?

For this question, the correct estimate is 180, and the MOE is ±114. For the three–level table, the reliability column is coded yellow with the message “fair.” For the four–level table, the reliability column is coded orange with the message “fair” or “less reliable.”

The results show that for the baseline table, 80% of participants who found the correct estimate would report this estimate, 33% would report it for the three–level table, 80% would report it for the four–level “good” table, and 80% would report it for the four–level “reliable” table. The lowest number of participants said that they would use the estimate in the three–level table condition. One possible difference in the results between the three– and four–level tables is that the data reliability message is coded yellow for the three–level table and orange for the four–level tables. There may have been a difference in the way that participants viewed the message in the context of two different colors.

Task 9: You are writing a news article about voter turnout in the 2008 presidential election and want to find out how many civilians are 18 or older in your home town of Hays, KS. What results do you find in the table?

For this question, the correct estimate is 16,098, and the MOE is ±378. For the three–level table, the reliability column is coded green with the message "good." For the four–level table, the reliability column is coded green with the message "excellent" or "reliable."

Several participants did not find the correct estimate for this task. In order to determine whether they would or would not report the estimate as described in the vignette, participants must first find that estimate. Therefore, this determination was not applicable for several participants. A few participants who did find the correct answer expressed a lack of confidence that the answer was correct, which may have impacted their decision of whether or not to report that estimate. Of the participants who found the correct answer, 67% of participants in the baseline condition would report this estimate, 80% would report it for the three–level table, 67% would report it for the four–level "good" table, and 0% would report it for the four–level "reliable" table. This task posed some difficulty for many participants and many of them never found the correct answer. The percentages listed are for those participants who found the correct answer. The wording of this question will be changed for the third round of testing.

Mentioning and Reporting the Margin of Error

Whether or not participants would report the margin of error along with the estimate and whether the color–coded reliability indicator would influence this decision was a question of interest for the sponsor. These totals do not include scores for tasks 3 and 4, which had several sub–questions. The sample size reflects the number of tasks completed (out of a possible total of 7) multiplied by the number of participants who worked with each table. Table 4 shows the percentage of responses by table type where the participant mentioned the margin of error in their response. A one–way ANOVA showed that there were no significant differences in mentioning the MOE among the groups.

Table 4. Mention MOE in Response

Participant	Sample Size	Mentioned MOE (%)
Baseline	35	51.4
Three–Level	42	40.5
Four–Level "Good"	35	57.1
Four–Level "Reliable"	35	51.4

Table 5 shows the percentage of responses by table type where the participants explicitly stated that they would report the MOE along with the estimate. A one–way ANOVA showed that there were no significant differences in reporting the MOE among the groups. These results indicate that there were no significant differences between the tables.

Reporting the Color–Coded Reliability Message

As in the first round of testing, many participants tended to stop using the MOE once they started using the color–coded reliability indicator. One participant even commented that

Table 5. Report MOE with Estimate

Participant	Sample Size	Would Report MOE (%)
Baseline	35	57.1
Three–Level	42	40.5
Four–Level "Good"	35	54.3
Four–Level "Reliable"	35	48.6

the indicator was "addictive" and that he was aware that he had stopped using the MOE and was using this color–coding instead.

No participants said that they would report the color itself (e.g., red, green, yellow), but rather the message contained in the column. Although a few participants recommended getting rid of the "Reliability" column and just highlighting the estimate itself, there is an inherent benefit to including the message within that column in addition to addressing 508 issues. In particular, it conveys a message with suggested wording for reporting caution or reliability along with the estimate. This is especially important when the participant does not also report the MOE.

Table 6 shows the percentage of responses by table type where the participant mentioned the margin of error in their response. These totals do not include scores for tasks 3 and 4, which had several sub–questions. The sample size reflects the number of number of tasks completed (out of a possible total of 7) multiplied by the number of participants who completed each table. A one–way ANOVA comparing the three prototype tables on this variable (since the baseline table did not have a color–coded indicator, it was excluded from the analysis) showed that there was at least one significant difference among the groups ($\alpha = 0.05, F(2, 109) = 13.07, p < 0.001$). Post–hoc Tukey t–tests indicated that participants in both the four–level "good" table ($\alpha = 0.05$,mean difference=0.485, $p < 0.001$) and the four–level "reliable" table conditions ($\alpha = 0.05$,mean difference=0.400, $p < 0.001$) were significantly more likely to mention the message from the color–coded reliability indicator than participants in the three-level indicator condition. There was no significant difference between the two four–level tables themselves on this variable ($\alpha = 0.05$,mean difference = 0.86, $p > 0.05$).

Table 6. Mention Color–Coded Indicator or Message in Response

Participant	Sample Size	Mentioned Indicator (%)
Baseline	35	NA
Three–Level	42	14.3
Four–Level "Good"	35	62.9
Four–Level "Reliable"	35	54.3

Table 7 shows the percentage of responses by table type where the participants explicitly stated that they would report the label message from the color–coded reliability indicator along with the estimate. That is, whether the participate would use the data in the context of the task vignette and said the label message as part of their final answer. A one–way ANOVA comparing the three prototype tables on this variable (since the baseline table did not have a color–coded indicator, it was excluded from the analysis) showed that there was at least one significant difference among the groups ($\alpha = 0.05, F(2, 109) = 6.36, p = 0.002$). Post–hoc Tukey tests indicated that participants in both the four–level "good" table ($\alpha = 0.05$, mean difference=0.31, $p = 0.008$) and the four–level "reliable" table conditions ($\alpha = 0.05$,mean difference=0.31, $p = 0.008$) were significantly more likely to explicitly report the message from the color–coded reliability indicator along with the estimate than participants in the three-level indicator condition. There was no significant difference between the two four–level tables themselves on this variable ($\alpha = 0.05$,mean difference = 0.00, $p > 0.05$).

Table 7. Report Color–Coded Indicator or Message with Estimate

Participant	Sample Size	Would Report (%)
Baseline	35	NA
Three–Level	42	11.9
Four–Level "Good"	35	42.9
Four–Level "Reliable"	35	42.9

4.2 Efficiency

The start and stop times for the different tasks were obtained from the time stamps on the eye–tracking data in order to calculate average times to complete the tasks. Efficiency scores for tasks 3 and 4, which had sub–questions, consist of a total time–on–task for all sub–parts to the question. The initial screening probe question (What is the first thing you noticed about this table?) was not scored for efficiency. Table 8 shows the efficiency scores in seconds by table. The sample size is the number of participants in each condition multiplied by the number of tasks they completed. A sample size of 54 means that 6 people completed 9 tasks each. A one–way ANOVA showed that there were no significant differences in efficiency among the tables ($\alpha = 0.05, F(3, 185) = 0.6, p > 0.05$. Some participants had trouble finding the geographic area associated with the table for the first task that they performed, which may have added extra time to their efficiency scores for the first task they performed. However, the tasks were presented in a random order to each participant, which should have ameliorated the effect of this issue.

Table 9 lists the efficiency score results by task number. A sample size of 21 reflects the number of participants who completed each task.

In summary, there were no significant differences across treatments in the amount of time required for a participant to complete the assigned tasks. One-way ANOVAs (across all four table types) were conducted to check for the possible influence of differing participant experience and educational levels. No significant differences in efficiency were found with education, how long the participant has been using ACS products, how often the participant

Table 8. Efficiency Results by Table

Participant	Sample Size	Average Time (sec)
Baseline	45	136
Three–Level	54	157
Four–Level “Good”	45	163
Four–Level “Reliable”	45	133

Table 9. Efficiency Results by Task

Task	Sample Size	Average Time (sec)
1	21	128
2	21	100
3	21	136
4	21	359
5	21	84
6	21	129
7	21	69
8	21	102
9	21	224

uses ACS products, number of statistics courses taken, or self-rated level of expertise with statistics as independent variables ($p > 0.05$).

4.3 Satisfaction

The modified QUIS instrument (Chin et al., 1988) asks participants to score items on a scale of 1 to 9. For reference and for the specific scale labels for each item, a copy of the entire QUIS survey can be found in Appendix F. Scores in the tables below were calculated by taking the average satisfaction score across table type (Table 10) and QUIS item (Table 11). This satisfaction questionnaire measures how satisfied participants were with using the data tables during the session.

In Table 10, sample size corresponds to the number of completed items for each table. Some participants chose “not applicable” or skipped some items, and these skips are reflected in the differing sample sizes. A one–way ANOVA showed that there was at least one significant difference among the tables on this variable ($\alpha = 0.05, F(3, 151) = 5.431, p = 0.001$). Planned comparisons ($\alpha = 0.05$) between the tables were performed. The baseline table was coded as condition 1, the three–level table was condition 2, the four–level “good” table was condition 3, and the four–level “reliable” table was condition 4.

1. The baseline table was compared to all of the prototypes and the results were not significant ($t(153) = 1.45, p > 0.05$).

2. The baseline table was compared to the three–level table and the results indicated that the baseline table scored significantly higher than the three–level table ($t(153) = -3.128, p = 0.002$).

3. The baseline table was compared to both four–level tables and the results were not significant ($t(153) = -0.33, p > 0.05$).

4. The three–level table was compared against both four–level tables and the results indicated that the four–level tables scored significantly higher than the three–level table ($t(153) = 3.43, p = 0.001$).

5. The two four–level tables were compared to each other and the results were not significant ($t(153) = 1.45, p > 0.05$).

For this round of testing, both the four–level table and the baseline table had significantly higher satisfaction scores than the three–level table, and the baseline table had significantly higher scores than the three–level table.

Table 10. QUIS Scores by Table

Participant	Sample Size	Average QUIS
Baseline	34	7.2
Three–Level	47	6.1
Four–Level “Good”	39	6.8
Four–Level “Reliable”	37	7.3

Table 11 lists the average QUIS scores by item. The sample size refers to the number of participants who completed that particular item. Sample sizes differ among tasks due to skipped and not applicable responses. Item 7 asked about the color–coded reliability indicator, so the baseline–condition participants marked “not applicable” for that item.

Table 11. QUIS by Item

Question	Sample Size	Average Score
1	21	6.4
2	20	5.8
3	19	7.7
4	20	6.6
5	21	6.5
6	20	7.6
7	15	6.7
8	21	7.2

One-way ANOVAs (across all four table types) were conducted on the participant background variables to check for the possible influence of differing participant experience and educational levels. The questions from the background survey can be found in Appendix E. A significant effect was found for Education ($F(3, 145) = 3.5, p = 0.0.17$). Post-hoc Tukey tests revealed that people with a 4-year college degree had significantly higher satisfaction scores than participants with some college education (mean difference=1.45, $p = 0.024$); participants with four-year degrees had higher satisfaction scores than people with some post-graduate education (mean difference=1.55,$p = 0.013$); and people with post-graduate degrees had higher satisfaction scores than people with some post-graduate education (mean difference=1.18, $p = 0.03$).

Comments (Item 9)

Participants had the opportunity to write in comments at the end of the QUIS instrument for Item 9.

Here is a list of comments given on these forms without the participant number for privacy purposes:

Baseline Table

- Disability item not defined - not sure what number means since 64 different definitions of disability in fed programs - need to publish each response category and age covered. Item presented is not very useful for analysis.

3–Level Table

- Like to see general totals at top.
  P10: I didn't read how reliability was defined.

- Basis of color-coding needs to be upfront if to be used (Is it based on Variance, SD, what?)

4–Level "Good" Table

- This would be helpful to have for the other tables (i.e., demographic, economic).
- Would like the ability to click for more info about methodology and reliability estimates. Would like percentages to be calculated for subgroups- especially useful for areas larger than a town like this.
- The tables can be arranged in a different format for easier navigation to extract data.

- Margin of error column very confusing. I was looking for: MOE relationship to CV, definition of MOE, desired precision.

- I like color codes. Good for average user, whom I often talk to. For small area sample size would be nice. Keep margin of error. Perhaps more link (footnotes) that can be click if I want more info on say a "household" v. a "family."

4–Level "Reliable" Table

- Perhaps instead of having words "reliable" or "unreliable", have signs: + + + means very reliable and + means not reliable

- The info display question seems not too relevant to me, since this is one small selection of data, and there is a lot more out there

- Geographic area "Hays City, KS" should be highlighted; Did not think first screen would be target city - thought it would be USA or 50 states; Helpful if alternative rows were shaded due to distance between variable and associated data; Assumed reliability codewords were valid, compared est. with +/− values also to arrive at decision to accept or reject data.

4.4 Task Difficulty

In order to examine the validity of the easy–medium–hard designation assigned to the tasks before testing based on a review of existing cognitive and other literature, participants completed a task difficulty rating survey after the usability session. Each task was listed and participants were asked to rate the task on scale of 1 to 10, with 1 being very easy and 10 being very difficult. Table 12 lists the original rating and the average difficulty score given by the participants for each task.

Table 12. Task Difficulty Ratings

Task	Original Rating	Average Participant Rating
1	Medium	2.8
2	Medium	2.5
3	Hard	2.3
4	Hard	4.4
5	Medium	2.2
6	Hard	3.3
7	Medium	2.3
8	Medium	2.5
9	Hard	5

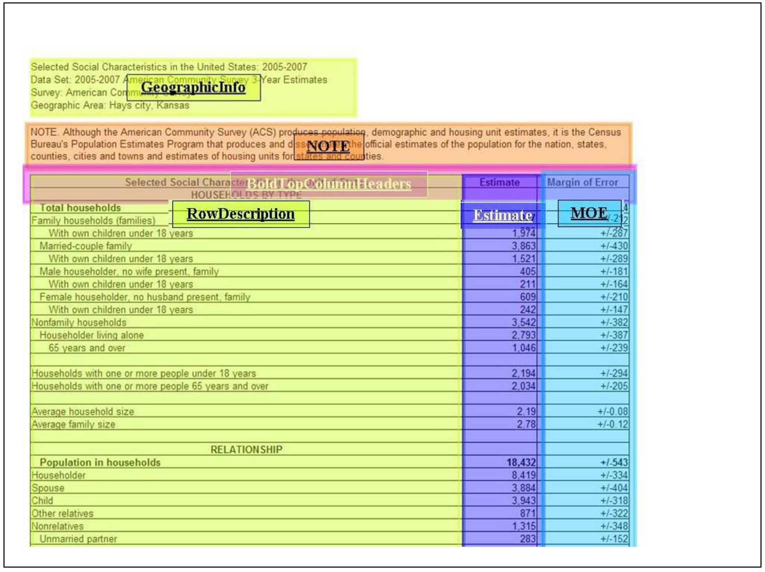
The results indicate that the three tasks originally rated as "hard" did score the three highest difficulty rating scales as rated by the participants: Tasks 4, 6, and 9. The fact that the highest average difficulty rating was a 5, and this rating was for Task 9 where the gaze plot data indicates further evidence of confusions, might be evidence that the difficulty rating scales should be changed from a 10–point scale to a 5–point scale. Additionally, these participants were experienced ACS data users and likely did not experience the same amount of difficulty with the tasks as a novice user would. Future research may compare the difficulty ratings of experts versus novice participants with respect to these difficulty ratings.

4.5 Eye–Tracking Results

The eye–tracking analysis captures evidence of the participants' cognitive process while they are completing the tasks. The horizontal and vertical position of the participant's eye is captured in real time and we can tell where a person looked and for how long. When fixations are repetitive and indicate repeated searching for information, for example, it might mean that the person is confused. If fixation durations are long for an area of the table, it might mean that it is the most interesting or relevant part of the table. If there are no fixations on an area, it means that the partic

Areas of interest (AOIs) for the tables were defined prior to the usability evaluation and can be found illustrated spatially in Figures 4, 5, 6, and 7, respectively. AOIs are typically used in eye–tracking analysis to evaluate how many times and how long participants looked at a certain area of the screen. The unit of measurement for a digital display on the Tobii system software and hardware is one pixel, and AOIs are defined by their X (vertical) and Y (horizontal) pixel coordinates. The entire screen has a resolution of 1024 by 768 pixels.

Figure 4. Areas of Interest for the Baseline Table

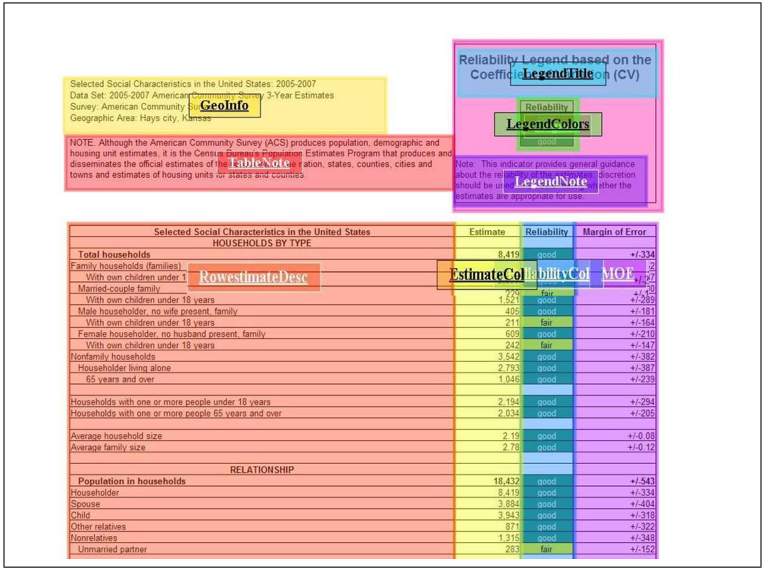


Selected Social Characteristics in the United States: 2005-2007
Data Set: 2005-2007 American Community Survey 3-Year Estimates
Survey: American Community Survey
Geographic Area: Hays city, Kansas

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Selected Social Characteristics in the United States	Estimate	Margin of Error
HOUSEHOLDS BY TYPE		
Total households	[illegible]	[illegible]
Family households (families)	[illegible]	+/-212
With own children under 18 years	1,974	+/-287
Married-couple family	3,863	+/-430
With own children under 18 years	1,521	+/-289
Male householder, no wife present, family	405	+/-181
With own children under 18 years	211	+/-164
Female householder, no husband present, family	609	+/-210
With own children under 18 years	242	+/-147
Nonfamily households	3,542	+/-382
Householder living alone	2,793	+/-387
65 years and over	1,046	+/-239
Households with one or more people under 18 years	2,194	+/-294
Households with one or more people 65 years and over	2,034	+/-205
Average household size	2.19	+/-0.08
Average family size	2.78	+/-0.12
RELATIONSHIP		
Population in households	**18,432**	**+/-543**
Householder	8,419	+/-334
Spouse	3,884	+/-404
Child	3,943	+/-318
Other relatives	871	+/-322
Nonrelatives	1,315	+/-348
Unmarried partner	283	+/-152

Figure 5. Areas of Interest for the Three–Level Table



Reliability Legend based on the Coefficient of Variation (CV)

Reliability
good

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used in determining whether the estimates are appropriate for use.

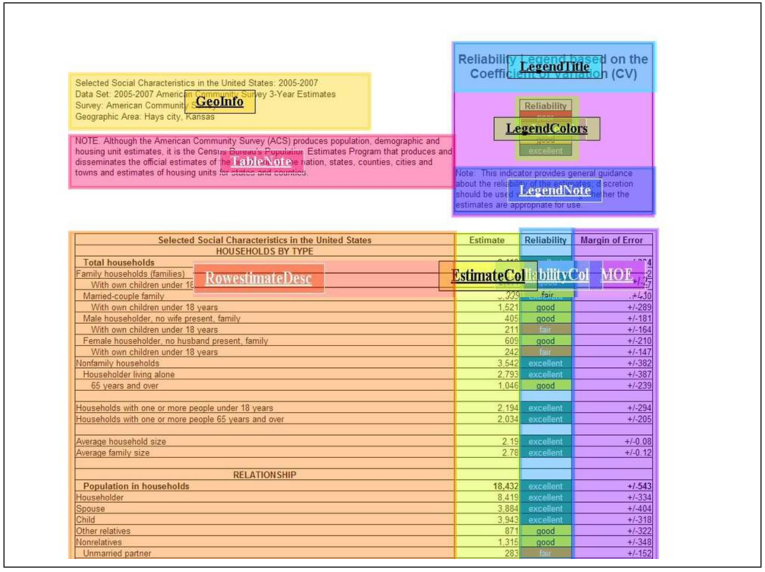
Selected Social Characteristics in the United States: 2005-2007
Data Set: 2005-2007 American Community Survey 3-Year Estimates
Survey: American Community Survey
Geographic Area: Hays city, Kansas

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Selected Social Characteristics in the United States	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	8,419	good	+/-334
Family households (families)	[illegible]	[illegible]	[illegible]
With own children under 1	[illegible]	[illegible]	[illegible]
Married-couple family	[illegible]	fair	[illegible]
With own children under 18 years	1,521	good	+/-289
Male householder, no wife present, family	405	good	+/-181
With own children under 18 years	211	fair	+/-164
Female householder, no husband present, family	609	good	+/-210
With own children under 18 years	242	fair	+/-147
Nonfamily households	3,542	good	+/-382
Householder living alone	2,793	good	+/-387
65 years and over	1,046	good	+/-239
Households with one or more people under 18 years	2,194	good	+/-294
Households with one or more people 65 years and over	2,034	good	+/-205
Average household size	2.19	good	+/-0.08
Average family size	2.78	good	+/-0.12
RELATIONSHIP			
Population in households	**18,432**	good	**+/-543**
Householder	8,419	good	+/-334
Spouse	3,884	good	+/-404
Child	3,943	good	+/-318
Other relatives	871	good	+/-322
Nonrelatives	1,315	good	+/-348
Unmarried partner	283	fair	+/-152

Tables 13, 14, 15, and 16 show the average fixation durations for all of the participants on each area of interest by table condition. Each fixation on an area of interest that lasted at least 100 milliseconds is recorded along with the duration of that fixation by the Tobii software.

Figure 6. Areas of Interest for the 4–Level "Good" Table



Reliability Legend based on the Coefficient of Variation (CV)

Selected Social Characteristics in the United States: 2005-2007
Data Set: 2005-2007 American Community Survey 3-Year Estimates
Survey: American Community Survey
Geographic Area: Hays city, Kansas

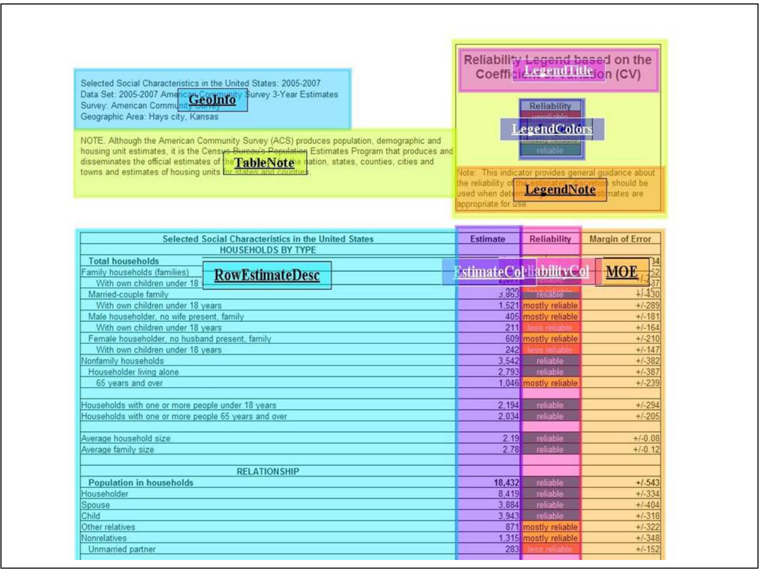
Reliability
[illegible]
[illegible]
good
excellent

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

Selected Social Characteristics in the United States	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	[illegible]	[illegible]	[illegible]
Family households (families)	[illegible]	[illegible]	[illegible]
With own children under 18 years	[illegible]	[illegible]	[illegible]
Married-couple family	[illegible]	fair	+/-430
With own children under 18 years	1,521	good	+/-289
Male householder, no wife present, family	405	good	+/-181
With own children under 18 years	211	fair	+/-164
Female householder, no husband present, family	609	good	+/-210
With own children under 18 years	242	fair	+/-147
Nonfamily households	3,542	excellent	+/-382
Householder living alone	2,793	excellent	+/-387
65 years and over	1,046	good	+/-239
Households with one or more people under 18 years	2,194	excellent	+/-294
Households with one or more people 65 years and over	2,034	excellent	+/-205
Average household size	2.19	excellent	+/-0.08
Average family size	2.78	excellent	+/-0.12
RELATIONSHIP			
Population in households	**18,432**	excellent	**+/-543**
Householder	8,419	excellent	+/-334
Spouse	3,884	excellent	+/-404
Child	3,943	excellent	+/-318
Other relatives	871	good	+/-322
Nonrelatives	1,315	good	+/-348
Unmarried partner	283	fair	+/-152

Figure 7. Areas of Interest for the 4–Level "Reliable" Table



Reliability Legend based on the Coefficient of Variation (CV)

Selected Social Characteristics in the United States: 2005-2007
Data Set: 2005-2007 American Community Survey 3-Year Estimates
Survey: American Community Survey
Geographic Area: Hays city, Kansas

Reliability
unreliable
[illegible]
mostly reliable
reliable

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

Selected Social Characteristics in the United States	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	[illegible]	[illegible]	[illegible]
Family households (families)	[illegible]	[illegible]	[illegible]
With own children under 18 years	[illegible]	[illegible]	[illegible]
Married-couple family	3,863	reliable	+/-430
With own children under 18 years	1,521	mostly reliable	+/-289
Male householder, no wife present, family	405	mostly reliable	+/-181
With own children under 18 years	211	less reliable	+/-164
Female householder, no husband present, family	609	mostly reliable	+/-210
With own children under 18 years	242	less reliable	+/-147
Nonfamily households	3,542	reliable	+/-382
Householder living alone	2,793	reliable	+/-387
65 years and over	1,046	mostly reliable	+/-239
Households with one or more people under 18 years	2,194	reliable	+/-294
Households with one or more people 65 years and over	2,034	reliable	+/-205
Average household size	2.19	reliable	+/-0.08
Average family size	2.78	reliable	+/-0.12
RELATIONSHIP			
Population in households	**18,432**	reliable	**+/-543**
Householder	8,419	reliable	+/-334
Spouse	3,884	reliable	+/-404
Child	3,943	reliable	+/-318
Other relatives	871	mostly reliable	+/-322
Nonrelatives	1,315	mostly reliable	+/-348
Unmarried partner	283	less reliable	+/-152

Table 13. Fixation Durations on Areas of Interest in Seconds for Baseline Table - 5 Participants

AOI	Average	SD	Min	Max
Bold Col. Head.	14.5	15.5	0.6	40.0
Geog. Info	13.6	15.6	1.0	40.1
Table Note	11.5	8.5	0.7	23.9
MOE	52.7	80.0	0.5	189.3
Estimate Col.	12.9	112.4	1.1	257.1
Estimate Desc.	155.9	183.1	4.0	378.7

Table 14. Fixation Durations on Areas of Interest in Seconds for Three–Level Table - 6 Participants

AOI	Average	SD	Min	Max
Legend (Box)	4.1	2.2	2.6	50.7
Legend Note	0.7	0.7	0.2	2.0
Legend Title	1.2	0.8	0	2.3
Legend Colors	1.6	1.2	0.1	2.9
Geog. Info	4.3	2.4	0.2	7.0
Table Note	2.9	2.7	0	7.8
MOE	20.0	17.2	0.2	42.5
Reliability Col.	52.0	52.9	2.3	152.2
Estimate Col.	43.8	35.4	3.1	108.0
Estimate Desc.	122.2	68.9	4.3	193.0

Table 15. Fixation Durations on Areas of Interest in Seconds for Four–Level "Good" Table - 5 Participants

AOI	Average	SD	Min	Max
Legend (Box)	11.7	8.9	1.5	18.3
Legend Note	4.8	5.8	0	13.9
Legend Title	2.1	1.9	0.7	5.0
Legend Colors	2.9	3.3	0.2	8.1
Geog. Info	3.7	5.5	0.6	13.5
Table Note	4.3	4.6	0.3	21.4
MOE	24.8	16.8	0.1	47.5
Reliability Col.	47.5	34.3	1.5	92.6
Estimate Col.	48.9	42.6	0.6	110.5
Estimate Desc.	133.2	101.3	6.6	232.7

Table 16. Fixation Durations on Areas of Interest in Seconds for Four–Level "Reliable" Table - 5 Participants

AOI	Average	SD	Min	Max
Legend (Box)	14.4	11.0	0	26.0
Legend Note	1.7	2.6	0	6.1
Legend Title	2.5	2.4	0	5.6
Legend Colors	7.3	6.5	0	15.0
Geog. Info	12.4	10.0	0.1	23.1
Table Note	7.7	4.0	1.1	11.7
MOE	26.6	19.3	0	52.3
Reliability Col.	86.5	103.8	2.1	262.2
Estimate Col.	72.6	51.3	0.8	132.9
Estimate Desc.	222.6	172.3	4.4	444.3

One–way ANOVAs by condition for each of the AOIs revealed that there were no significant differences in fixation durations among the table conditions for any of the AOIs, including the MOE column ($F(3, 17) = 2.8, p = 0.07$). While not significant at the $\alpha = 0.05$ level, these results do suggest that the additional information in the 3– and 4–level tables gains attention and time spent looking at MOEs.

Heat maps were constructed for each of the tables. The Tobii Studio software used to run most of the participants in this study uses a red–yellow–green scale as a default where red indicates areas that had the most fixations, yellow indicates a mid–level amount of fixations, and green indicates relatively few fixations. These fixations are relative, so areas shaded in green may have been fixated upon multiple times, but not as frequently as the areas shaded in yellow or red.

Heat maps of each of the tables examined in this study show the overall distribution of eye fixations that lasted at least 100 milliseconds for each table, averaged across all of the tasks. Perhaps the most striking difference among the concentrations of fixations on these tables is on the MOE column. Specifically, the heat maps show that participants did not tend to look at the MOE column as often for the three–level table (Figure 9) as for the other three tables (Figures 8, 10, and 11. Although there are different numbers of participants represented on these composite figures (due to an incompatibility between the plotting tools for Tobii Clearview and Studio software packages), these results are consistent with the results of the fixation duration analysis, which can be found in Tables 13, 14, 15, and 16. One possible explanation for this trend in eye fixations is that the the three–level indicator provided more implicit meaning than the four–level indicator and that the participants did not feel the need to seek additional information from the MOE column as often, which is a concept that may be examined in future studies.

The statistical analysis combined the data from the Studio and Clearview packages and represents all of the participants. There were no significant differences for this analysis as shown by a one-way ANOVA ($p > 0.05$).

Figure 8. Composite Eye–tracking Heat Map for the Baseline Table- 5 Participants

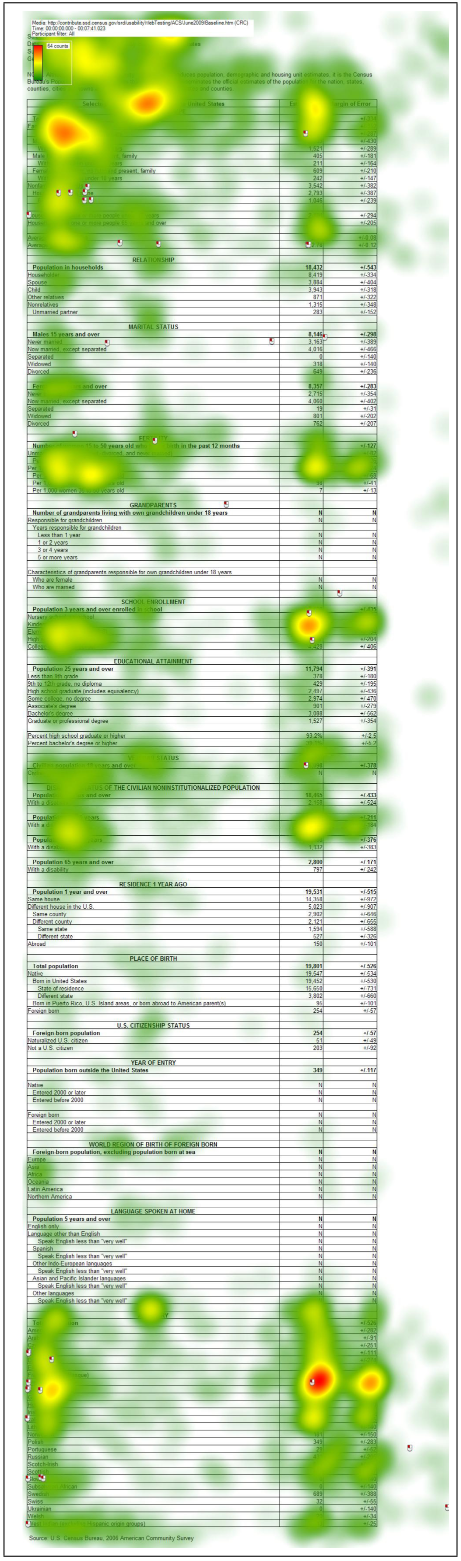
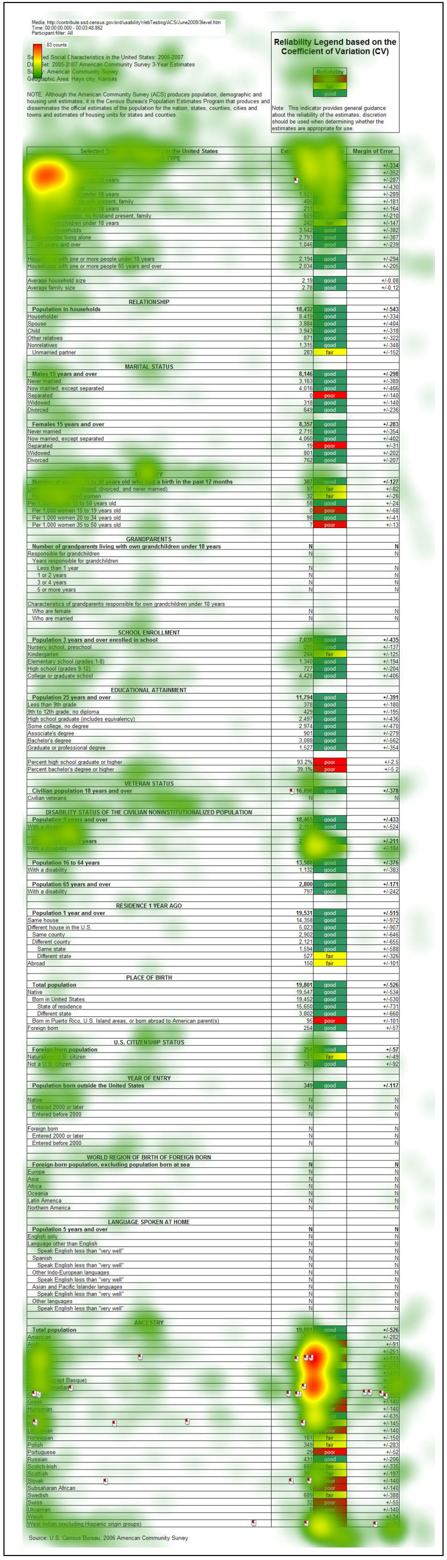

Media: http://contribute.ssd.census.gov/srd/usabilityWebTesting/ACS/June2009/Baseline.htm (CRC)
Time: 00:00:00.000 - 00:07:41.023
Participant filter: All
64 counts
Source: U.S. Census Bureau, 2006 American Community Survey

Figure 9. Composite Eye–tracking Heat Map for the Three–Level Table- 4 Participants



Media: http://contribute.ssd.census.gov/srd/usabilityWebTesting/ACS/June2009/3level.htm
Time: 00:00:00.000 - 00:03:48.862
Participant filter: All

Se[illegible]ted Social Characteristics in the United States: 2005-2007
Da[illegible]Set: 2005-2007 American Community Survey 3-Year Estimates
Su[illegible]y: American Community Survey
Geographic Area: Hays city, Kansas

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Reliability Legend based on the Coefficient of Variation (CV)

Reliability
poor
fair
good

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

Selected So[illegible] in the United States	Esti[illegible]	[illegible]	Margin of Error
HOUS[illegible] BY TYPE			
[illegible]	[illegible]	[illegible]	+/-334
[illegible]	[illegible]	[illegible]	+/-352
[illegible]der 18 years	[illegible]	[illegible]	+/-287
[illegible]y	3,6[illegible]	[illegible]	+/-430
[illegible]n under 18 years	1,521	[illegible]	+/-289
[illegible]no wife present, family	405	[illegible]	+/-181
[illegible]ren under 18 years	211	[illegible]	+/-164
[illegible]seholder, no husband present, family	609	[illegible]	+/-210
[illegible]wn children under 18 years	242	fair	+/-147
No[illegible] households	3,542	good	+/-382
Householder living alone	2,793	good	+/-387
65 years and over	1,046	good	+/-239
Households with one or more people under 18 years	2,194	good	+/-294
Households with one or more people 65 years and over	2,034	good	+/-205
Average household size	2.19	good	+/-0.08
Average family size	2.78	good	+/-0.12
RELATIONSHIP			
Population in households	**18,432**	good	**+/-543**
Householder	8,419	good	+/-334
Spouse	3,884	good	+/-404
Child	3,943	good	+/-318
Other relatives	871	good	+/-322
Nonrelatives	1,315	good	+/-348
Unmarried partner	283	fair	+/-152
MARITAL STATUS			
Males 15 years and over	**8,146**	good	**+/-298**
Never married	3,163	good	+/-389
Now married, except separated	4,016	good	+/-466
Separated	0	poor	+/-140
Widowed	318	good	+/-140
Divorced	649	good	+/-236
Females 15 years and over	**8,357**	good	**+/-283**
Never married	2,715	good	+/-354
Now married, except separated	4,060	good	+/-402
Separated	19	poor	+/-31
Widowed	801	good	+/-202
Divorced	762	good	+/-207
F[illegible]ITY			
Number of wom[illegible] 15 to 50 years old who [illegible] a birth in the past 12 months	307	good	**+/-127**
Unm[illegible]idowed, divorced, and never married)	97	fair	+/-82
Per [illegible]ried women	32	fair	+/-26
Per 1,000 [illegible] 15 to 50 years old	58	good	+/-24
Per 1,000 women 15 to 19 years old	0	poor	+/-68
Per 1,000 women 20 to 34 years old	98	good	+/-41
Per 1,000 women 35 to 50 years old	7	poor	+/-13
GRANDPARENTS			
Number of grandparents living with own grandchildren under 18 years	**N**		**N**
Responsible for grandchildren	N		N
Years responsible for grandchildren			
Less than 1 year	N		N
1 or 2 years	N		N
3 or 4 years	N		N
5 or more years	N		N
Characteristics of grandparents responsible for own grandchildren under 18 years			
Who are female	N		N
Who are married	N		N
SCHOOL ENROLLMENT			
Population 3 years and over enrolled in school	**7,038**	good	**+/-435**
Nursery school, preschool	299	good	+/-137
Kindergarten	244	fair	+/-125
Elementary school (grades 1-8)	1,340	good	+/-194
High school (grades 9-12)	727	good	+/-204
College or graduate school	4,428	good	+/-406
EDUCATIONAL ATTAINMENT			
Population 25 years and over	**11,794**	good	**+/-391**
Less than 9th grade	378	good	+/-180
9th to 12th grade, no diploma	429	good	+/-195
High school graduate (includes equivalency)	2,497	good	+/-436
Some college, no degree	2,974	good	+/-470
Associate's degree	901	good	+/-279
Bachelor's degree	3,088	good	+/-562
Graduate or professional degree	1,527	good	+/-354
Percent high school graduate or higher	93.2%	poor	+/-2.5
Percent bachelor's degree or higher	39.1%	poor	+/-5.2
VETERAN STATUS			
Civilian population 18 years and over	**16,098**	good	**+/-378**
Civilian veterans	N		N
DISABILITY STATUS OF THE CIVILIAN NONINSTITUTIONALIZED POPULATION			
Population 5 years and over	**18,465**	good	**+/-433**
With a disabil[illegible]	2,158	good	+/-524
Po[illegible] 15 years	**2[illegible]**	[illegible]	**+/-211**
With a disability	[illegible]	[illegible]	+/-184
Population 16 to 64 years	**13,588**	good	**+/-376**
With a disability	1,132	good	+/-383
Population 65 years and over	**2,800**	good	**+/-171**
With a disability	797	good	+/-242
RESIDENCE 1 YEAR AGO			
Population 1 year and over	**19,531**	good	**+/-515**
Same house	14,358	good	+/-972
Different house in the U.S.	5,023	good	+/-907
Same county	2,902	good	+/-646
Different county	2,121	good	+/-655
Same state	1,594	good	+/-588
Different state	527	fair	+/-326
Abroad	150	fair	+/-101
PLACE OF BIRTH			
Total population	**19,801**	good	**+/-526**
Native	19,547	good	+/-534
Born in United States	19,452	good	+/-530
State of residence	15,650	good	+/-731
Different state	3,802	good	+/-660
Born in Puerto Rico, U.S. Island areas, or born abroad to American parent(s)	95	poor	+/-101
Foreign born	254	good	+/-57
U.S. CITIZENSHIP STATUS			
Foreign-born population	**254**	good	**+/-57**
Naturalized U.S. citizen	51	fair	+/-49
Not a U.S. citizen	203	good	+/-92
YEAR OF ENTRY			
Population born outside the United States	**349**	good	**+/-117**
Native	N		N
Entered 2000 or later	N		N
Entered before 2000	N		N
Foreign born	N		N
Entered 2000 or later	N		N
Entered before 2000	N		N
WORLD REGION OF BIRTH OF FOREIGN BORN			
Foreign-born population, excluding population born at sea	**N**		**N**
Europe	N		N
Asia	N		N
Africa	N		N
Oceania	N		N
Latin America	N		N
Northern America	N		N
LANGUAGE SPOKEN AT HOME			
Population 5 years and over	**N**		**N**
English only	N		N
Language other than English	N		N
Speak English less than "very well"	N		N
Spanish	N		N
Speak English less than "very well"	N		N
Other Indo-European languages	N		N
Speak English less than "very well"	N		N
Asian and Pacific Islander languages	N		N
Speak English less than "very well"	N		N
Other languages	N		N
Speak English less than "very well"	N		N
ANCESTRY			
Total population	**19,801**	good	**+/-526**
American	[illegible]	[illegible]	+/-282
Arab	[illegible]	[illegible]	+/-91
C[illegible]	[illegible]	[illegible]	+/-251
[illegible]	[illegible]	[illegible]	+/-111
[illegible]	[illegible]	[illegible]	+/-274
E[illegible]	[illegible]	[illegible]	+/-496
F[illegible]xcept Basque)	[illegible]	[illegible]	+/-2[illegible]
F[illegible]anadian	[illegible]	[illegible]	[illegible]
G[illegible]an	[illegible]	[illegible]	+/-962
Greek	[illegible]	[illegible]	+/-140
Hungarian	[illegible]	poor	+/-140
Irish	2[illegible]	good	+/-635
It[illegible]	[illegible]	fair	+/-145
Lithuanian	[illegible]	poor	+/-140
Norwegian	181	fair	+/-150
Polish	349	fair	+/-283
Portuguese	29	poor	+/-52
Russian	431	good	+/-206
Scotch-Irish	666	fair	+/-335
Scottish	36[illegible]	fair	+/-197
Slovak	[illegible]	poor	+/-140
Subsaharan African	0	poor	+/-140
Swedish	689	fair	+/-388
Swiss	32	poor	+/-55
Ukrainian	0	poor	+/-140
Welsh	[illegible]	[illegible]	+/-34
West Indian (excluding Hispanic origin groups)	[illegible]	[illegible]	+/-25

Source: U.S. Census Bureau, 2006 American Community Survey

Figure 10. Composite Eye–tracking Heat Map for the Four–Level "Good" Table- 3 Participants

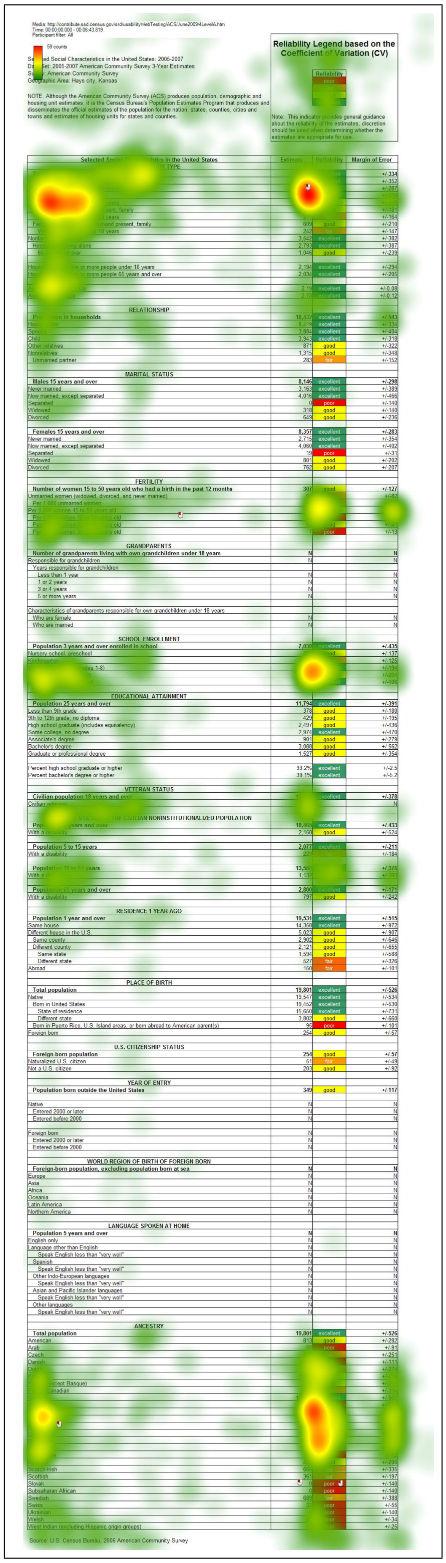
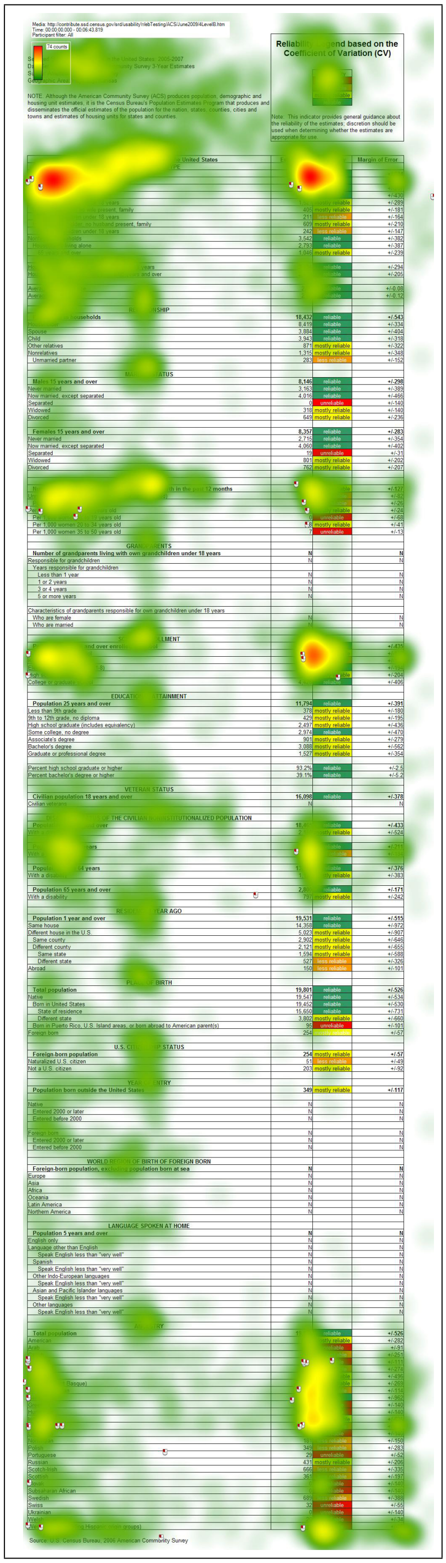


Figure 11. Composite Eye–tracking Heat Map for the Four–Level "Reliable" Table-5 Participants



Media: http://contribute.ssd.census.gov/srd/usabilityWebTesting/ACS/June2009/4LevelB.htm
Time: 00:00:00.000 - 00:06:43.819
Participant filter: All

Se[illegible] n the United States: 2005-2007
Da[illegible] munity Survey 3-Year Estimates
Su[illegible] y
Geographic Area: [illegible] ansas

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Reliabil[illegible] gend based on the Coefficient of Variation (CV)

[illegible] ty
[illegible]
[illegible] e
mostly reliable
reliable

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

[illegible] the United States	E[illegible]	[illegible] y	Margin of Error
[illegible] YPE			
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	+/-[illegible]
[illegible]	[illegible]	[illegible]	+/-430
[illegible] 18 years	1,52[illegible]	mostly reliable	+/-289
[illegible] wife present, family	405	mostly reliable	+/-181
[illegible] ren under 18 years	211	less reliable	+/-164
[illegible] lder, no husband present, family	609	mostly reliable	+/-210
[illegible] dren under 18 years	242	less reliable	+/-147
Nonf[illegible] eholds	3,542	reliable	+/-382
Househ[illegible] living alone	2,793	reliable	+/-387
65 years and over	1,046	mostly reliable	+/-239
Ho[illegible] years	[illegible]	reliable	+/-294
Hou[illegible] years and over	[illegible]	reliable	+/-205
Avera[illegible] e	2[illegible]	reliable	+/-0.08
Averag[illegible]	2[illegible]	reliable	+/-0.12
RE[illegible] NSHIP			
[illegible] households	18,432	reliable	+/-543
Householder	8,419	reliable	+/-334
Spouse	3,884	reliable	+/-404
Child	3,943	reliable	+/-318
Other relatives	871	mostly reliable	+/-322
Nonrelatives	1,315	mostly reliable	+/-348
Unmarried partner	283	less reliable	+/-152
MAR[illegible] TATUS			
Males 15 years and over	8,146	reliable	+/-298
Never married	3,163	reliable	+/-389
Now married, except separated	4,016	reliable	+/-466
Separated	0	unreliable	+/-140
Widowed	318	mostly reliable	+/-140
Divorced	649	mostly reliable	+/-236
Females 15 years and over	8,357	reliable	+/-283
Never married	2,715	reliable	+/-354
Now married, except separated	4,060	reliable	+/-402
Separated	19	unreliable	+/-31
Widowed	801	mostly reliable	+/-202
Divorced	762	mostly reliable	+/-207
Nu[illegible] th in the past 12 months	[illegible]	[illegible] able	+/-127
Un[illegible] d)	[illegible]	[illegible]	+/-82
P[illegible]	[illegible]	[illegible] able	+/-26
Per[illegible] ears old	[illegible]	[illegible] tly reliable	+/-24
Per 1,[illegible] to 19 years old	0	unreliable	+/-68
Per 1,000 women 20 to 34 years old	[illegible]8	mostly reliable	+/-41
Per 1,000 women 35 to 50 years old	7	unreliable	+/-13
GRAND[illegible] ENTS			
Number of grandparents living with own grandchildren under 18 years	N		N
Responsible for grandchildren	N		N
Years responsible for grandchildren			
Less than 1 year	N		N
1 or 2 years	N		N
3 or 4 years	N		N
5 or more years	N		N
Characteristics of grandparents responsible for own grandchildren under 18 years			
Who are female	N		N
Who are married	N		N
S[illegible] LLMENT			
P[illegible] and over enroll[illegible] ol	[illegible]	[illegible]	+/-435
[illegible] ol	[illegible]	[illegible]	+/-[illegible]
K[illegible]	[illegible]	[illegible]	+/-[illegible]
Ele[illegible] 1-8)	[illegible]	[illegible]	+/-194
High sc[illegible]	[illegible]	[illegible] able	+/-204
College or graduate school	4,42[illegible]	reliable	+/-406
EDUCATIO[illegible] TTAINMENT			
Population 25 years and over	11,794	reliable	+/-391
Less than 9th grade	378	mostly reliable	+/-180
9th to 12th grade, no diploma	429	mostly reliable	+/-195
High school graduate (includes equivalency)	2,497	mostly reliable	+/-436
Some college, no degree	2,974	reliable	+/-470
Associate's degree	901	mostly reliable	+/-279
Bachelor's degree	3,088	mostly reliable	+/-562
Graduate or professional degree	1,527	mostly reliable	+/-354
Percent high school graduate or higher	93.2%	reliable	+/-2.5
Percent bachelor's degree or higher	39.1%	reliable	+/-5.2
VETERAN STATUS			
Civilian population 18 years and over	16,098	reliable	+/-378
Civilian veterans	N		N
DIS[illegible] OF THE CIVILIAN N[illegible] NSTITUTIONALIZED POPULATION			
Populati[illegible] and over	18,46[illegible]	reliable	+/-433
With a d[illegible]	2,1[illegible]	mostly reliable	+/-524
Po[illegible] ears	[illegible]	reliable	+/-211
With a[illegible]	[illegible]	[illegible] eliable	+/-184
Popula[illegible] 64 years	1[illegible]	reliable	+/-376
With a disability	1,[illegible]	mostly reliable	+/-383
Population 65 years and over	2,80[illegible]	reliable	+/-171
With a disability	797	mostly reliable	+/-242
RESIDEN[illegible] YEAR AGO			
Population 1 year and over	19,531	reliable	+/-515
Same house	14,358	reliable	+/-972
Different house in the U.S.	5,023	mostly reliable	+/-907
Same county	2,902	mostly reliable	+/-646
Different county	2,121	mostly reliable	+/-655
Same state	1,594	mostly reliable	+/-588
Different state	527	less reliable	+/-326
Abroad	150	less reliable	+/-101
PLAC[illegible] F BIRTH			
Total population	19,801	reliable	+/-526
Native	19,547	reliable	+/-534
Born in United States	19,452	reliable	+/-530
State of residence	15,650	reliable	+/-731
Different state	3,802	mostly reliable	+/-660
Born in Puerto Rico, U.S. Island areas, or born abroad to American parent(s)	95	unreliable	+/-101
Foreign born	254	mostly reliable	+/-57
U.S. CITI[illegible] IP STATUS			
Foreign-born population	254	mostly reliable	+/-57
Naturalized U.S. citizen	51	less reliable	+/-49
Not a U.S. citizen	203	mostly reliable	+/-92
YEAR [illegible] ENTRY			
Population born outside the United States	349	mostly reliable	+/-117
Native	N		N
Entered 2000 or later	N		N
Entered before 2000	N		N
Foreign born	N		N
Entered 2000 or later	N		N
Entered before 2000	N		N
WORLD REGION OF BIRTH OF FOREIGN BORN			
Foreign-born population, excluding population born at sea	N		N
Europe	N		N
Asia	N		N
Africa	N		N
Oceania	N		N
Latin America	N		N
Northern America	N		N
LANGUAGE SPOKEN AT HOME			
Population 5 years and over	N		N
English only	N		N
Language other than English	N		N
Speak English less than "very well"	N		N
Spanish	N		N
Speak English less than "very well"	N		N
Other Indo-European languages	N		N
Speak English less than "very well"	N		N
Asian and Pacific Islander languages	N		N
Speak English less than "very well"	N		N
Other languages	N		N
Speak English less than "very well"	N		N
AN[illegible] TRY			
Total population	19[illegible]	reliable	+/-526
American	[illegible]	mostly reliable	+/-282
Arab	[illegible]	unreliable	+/-91
C[illegible]	[illegible]	[illegible] able	+/-251
[illegible]	[illegible]	[illegible]	+/-111
[illegible]	[illegible]	[illegible]	+/-274
[illegible]	[illegible]	[illegible] able	+/-496
[illegible] t Basque)	[illegible]	[illegible] able	+/-269
[illegible] an	[illegible]	[illegible]	+/-114
[illegible]	[illegible]	[illegible]	+/-962
Gre[illegible]	[illegible]	[illegible]	+/-140
Hu[illegible]	[illegible]	[illegible]	+/-140
Iri[illegible]	[illegible]	[illegible] le	+/-[illegible]
[illegible]	[illegible]	[illegible]	+/-[illegible]
Li[illegible]	[illegible]	[illegible] able	+/-140
Norwegian	181	less reliable	+/-150
Polish	349	less reliable	+/-283
Portuguese	29	unreliable	+/-52
Russian	431	mostly reliable	+/-206
Scotch-Irish	666	less reliable	+/-335
Scottish	361	less reliable	+/-197
Slovak	0	unreliable	+/-140
Subsaharan African	0	unreliable	+/-140
Swedish	689	less reliable	+/-388
Swiss	32	unreliable	+/-55
Ukrainian	0	unreliable	+/-140
Welsh	2[illegible]	[illegible] able	+/-34
Wes[illegible] ing Hispanic origin groups)	[illegible]	[illegible]	[illegible]

Source: U.S. Census Bureau, 2006 American Community Survey

4.6 Participant Preference for Indicator Type

Participants were shown all three versions of the prototypical data reliability indicator tables along with the baseline table, were allowed to explore the tables for a few minutes, and were asked to state their preference (i.e., choose which version they would consider easiest to use) for the number of levels of indicator as well as the wording used in the indicator (e.g., "excellent/good/fair/poor" versus "unreliable/mostly reliable/less reliable/unreliable." The results of asking participants to state their preference can be found in Table 17. These results reflect only the participants' subjective choice of which table they would prefer to use, which is a qualitative evaluation of the instrument. Future testing may incorporate on participant's suggestion to rename the middle category as "fairly reliable."

Table 17. Preferred Version of Table and Indicator Wording by Participant

Participant	Table Seen	Occupation	Preferred Table	Preferred Wording
P1	3–Level	Researcher (Other)	Baseline	NA
P2	4–Level Good	Nonprofit	4-Level	Reliable
P3	3–Level	Journalist	3-Level	No Pref
P4	4–Level Reliable	Federal	4-Level	Reliable
P5	4–Level Reliable	Journalist	3-Level	Reliable
P6	Baseline	Journalist	3-Level	Reliable
P7	4–Level Good	Nonprofit	4-Level	Good
P8	Baseline	Federal	3-Level	Good
P9	Baseline	Journalist	3-Level	Reliable
P10	3–Level	Journalist	4-Level	Reliable
P11	4–Level Reliable	COPAFS	3-Level	Good
P12	4–Level Good	Federal	4-Level	Good
P13	4–Level Good	Journalist	4-Level	Good
P14	Baseline	Federal	No pref	Reliable
P15	4–Level Reliable	Nonprofit	4-Level	Reliable
P16	3–Level	Federal	3-Level	Good
P17	3–Level	Federal	3-Level	Reliable
P18	Baseline	Federal	3-Level	Good
P19	4–Level Reliable	Federal	4-Level	Good
P20	4–Level Good	COPAFS	4-Level	Good
P21	3–Level	Federal	3-Level	Reliable

The results of this question showed no clear preference overall for three versus four levels or for the "good" versus "reliable" wording. Ten participants preferred the three–level indicator, nine participants preferred the four–level indicator, one preferred the baseline table, and one had no preference. As for wording preference, nine preferred "good," ten preferred "reliable," and two had no preference. As in the first round of testing, there was an overwhelming preference for the prototypes over the baseline table overall. However, the real ACS data users were split almost equally on their preference for the three– versus four–level tables and for the two versions of the wording.

4.7 Usability Issues and Observations

Results reported include all identified usability issues and resolutions recommended by the team. Identified issues are prioritized based on the following criteria:

- **High:** This problem brought the test participant to a stand still. He or she was not able to complete the task.
- **Moderate:** This problem caused some difficulty or confusion, but the test participant was able to complete the task.
- **Low:** This problem caused minor annoyance but does not interfere with the flow of the tasks.

High–Priority Issues

1. Usability Issue: No formula for CV cutoffs or explanation of relationship between MOE and CV.

   Many participants expressed confusion over the lack of an explanation for the criteria used in determining the color–coded levels of the indicator. Similarly, several participants remarked that they were not sure why the CV was being used to determine the cutoffs when the MOE was provided in the table itself. There were many suggestions from participants about including more information about the cutoffs and either an explanation or a formula relating the CV and MOE either within the table itself or accessible through a hyperlink on another Web site.

   For example, participant 5 specifically mentioned that she wanted to see the formula, while Participant 12 did not think enough methodology information was given. Participant 13 said it was confusing to her that the legend said it was based on CV, but then the column gives you MOE. She said that she was not sure whether that means that it could be interpreted to mean the MOE was determined to be "Excellent." Participant 18 mentioned she did not know why the Census Bureau "switched" to CV to definite reliability when MOE is in the table. Participant 7 mentioned wanting to see more information in general about the CV and MOE and Participant 16 would want to see definition of the scales because if someone is reporting the estimate to someone like supervisor, they would need to know why it is called fair, etc. Participant 21 said that people would really like the color-coding, but the Census Bureau should show how we got the levels. Future testing may include either a mathematical formula relating the MOE and CV or an explanation of how they are related. Similarly, tables with more information about the cutoffs may be tested once criteria for these cutoffs have been approved by the Methodology and Standards council.

2. Usability Issue: Civilian Population 18 and Over (Task 9).

   Many participants could not find the answer to this question and many that did find it were not confident that it was the correct answer. The wording was changed ("civilian" to “people”) after the first few participants had difficulty finding the answer to this task. However, this did not improve the response rate for the rest of the participants. This task had one of the lowest overall accuracy scores and participants had some of their least efficient performances while completing this task. Many participants rated it as the most difficult question on the task difficulty rating scale (Appendix G). Future rounds of testing may require a revision in the wording of this task so that it is more clear to participants that the answer can actually be found in the table.

Eye–tracking gaze plots of this task for each of the tables indicates that participants looked all over the table before either succeeding or failing to find the correct estimate. The correct estimate is near the middle of the table, yet the gaze plots in Figures 12, 14, 16, and 18 show that participants did not easily find it. The area on the table where the correct answer could be found is circled in red.

Figure 12. Composite Eye–tracking Gaze Plot for Task 9 on the Baseline Table- 5 Participants

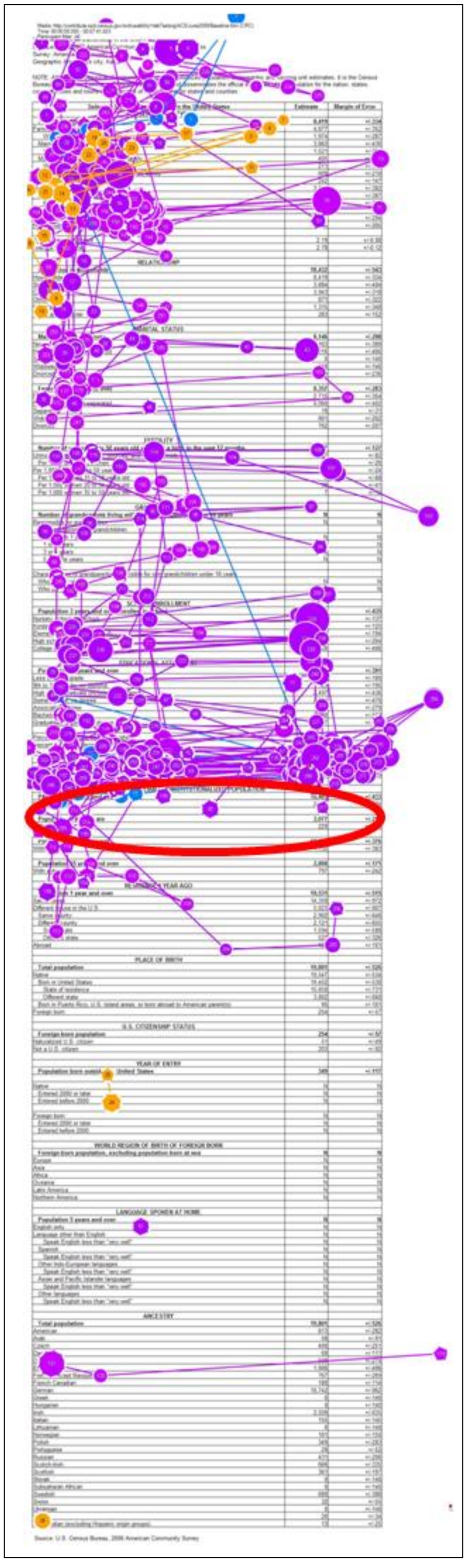


Figure 14. Composite Eye–tracking Gaze Plot for Task 9 on the Three–Level Table- 4 Participants

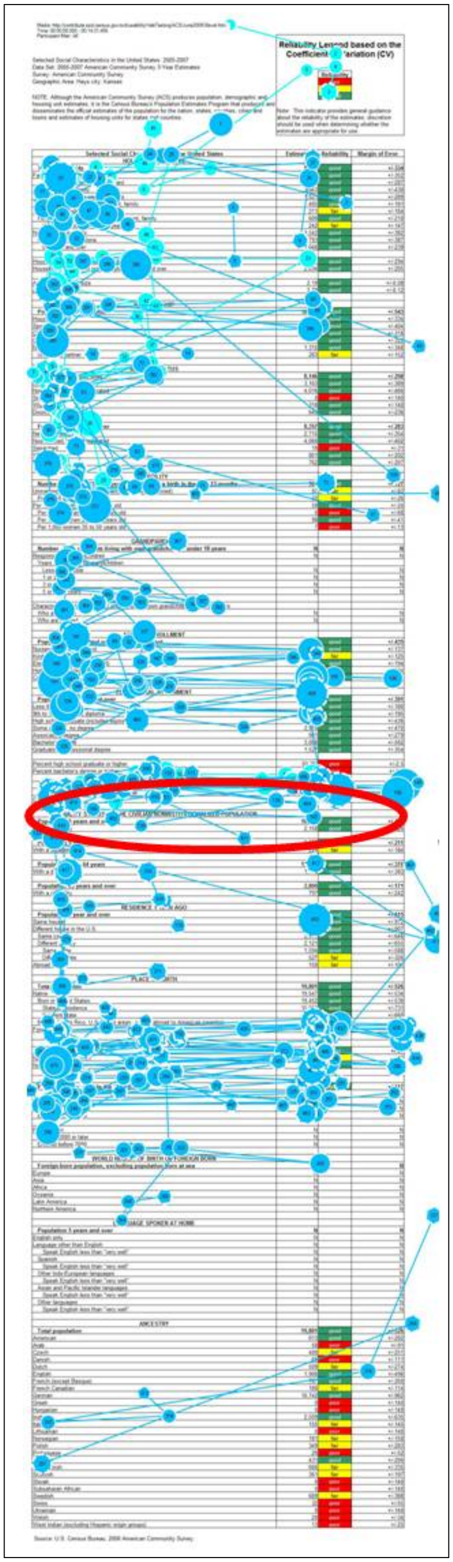

Reliability Legend based on the Coefficient of Variation (CV)

Figure 16. Composite Eye–tracking Gaze Plot for Task 9 on the Four–Level “Good” Table- 3 Participants

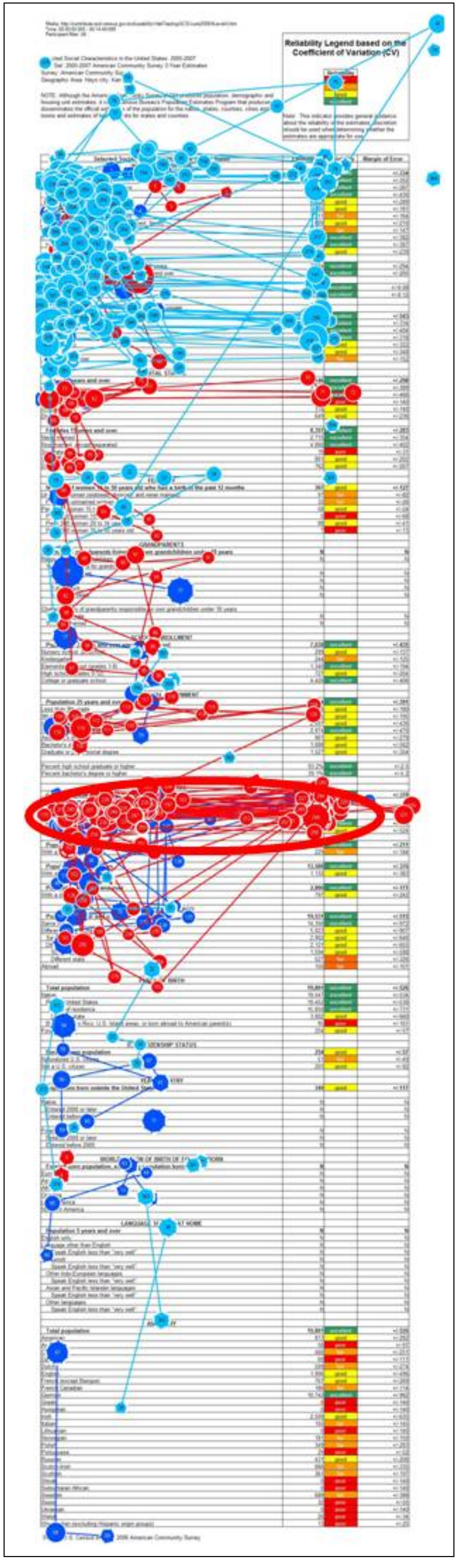
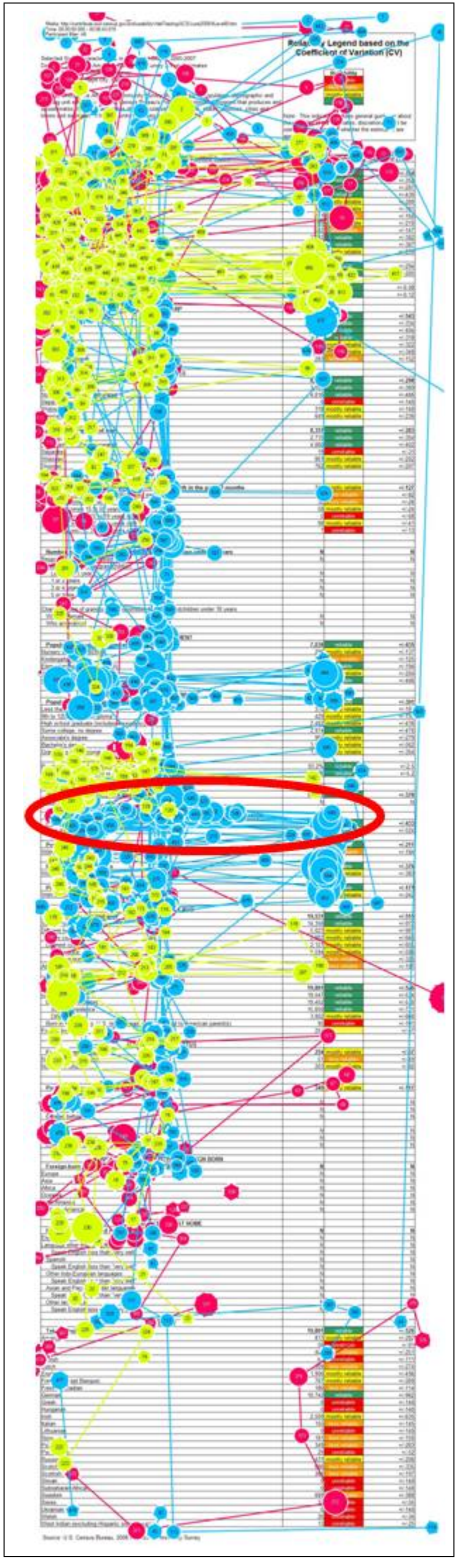

Reliability Legend based on the Coefficient of Variation (CV)

Figure 18. Composite Eye–tracking Gaze Plot for Task 9 on the Four–Level “Reliable” Table-5 Participants


Reliability Legend based on the Coefficient of Variation (CV)
Source: U.S. Census Bureau, 2006

3. Usability Issue: Width of descriptive column

   As in the first round of testing, participants reported that they had trouble tracking the correct estimate across the screen because the description of the estimate (left–most column) was so wide. Participant 6 mentioned that he would like to see the estimate "right next to" the name. Participant 8 said highlighting alternate rows and moving the columns closer together would help her because she had trouble "keeping track of where I am." Participant 11 said he had a "little bit of difficulty reading across" and used the mouse to highlight across the screen while selecting an estimate for almost every answer and made a mistake for Task 4a because he selected the estimate from the incorrect row. Participant 16 mentioned that she could highlight, but it was difficult to follow it across and said, "I wish I could click German and the whole row would highlight. That would really be great." Participant 19 said that it would help to shade alternate lines because it was hard to read the way it was.

   Some ACS new data tables use this type of alternate shading strategy Figure 20 shows a table of three–year ACS estimates for place of birth by sex in the United States.

Figure 20. New Table of ACS Three–Year Estimates from AFF

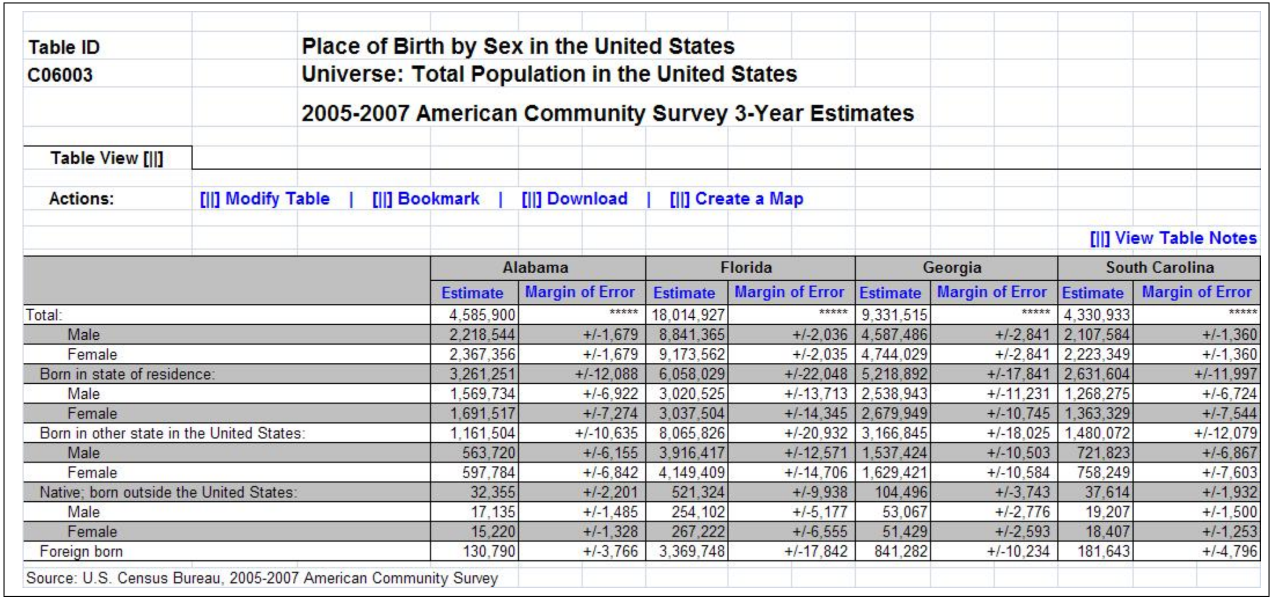


Table ID
C06003

Place of Birth by Sex in the United States
Universe: Total Population in the United States

2005-2007 American Community Survey 3-Year Estimates

Table View [||]
Actions: [||] Modify Table | [||] Bookmark | [||] Download | [||] Create a Map
[||] View Table Notes

	Alabama		Florida		Georgia		South Carolina	
	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error
Total:	4,585,900	*****	18,014,927	*****	9,331,515	*****	4,330,933	*****
Male	2,218,544	+/-1,679	8,841,365	+/-2,036	4,587,486	+/-2,841	2,107,584	+/-1,360
Female	2,367,356	+/-1,679	9,173,562	+/-2,035	4,744,029	+/-2,841	2,223,349	+/-1,360
Born in state of residence:	3,261,251	+/-12,088	6,058,029	+/-22,048	5,218,892	+/-17,841	2,631,604	+/-11,997
Male	1,569,734	+/-6,922	3,020,525	+/-13,713	2,538,943	+/-11,231	1,268,275	+/-6,724
Female	1,691,517	+/-7,274	3,037,504	+/-14,345	2,679,949	+/-10,745	1,363,329	+/-7,544
Born in other state in the United States:	1,161,504	+/-10,635	8,065,826	+/-20,932	3,166,845	+/-18,025	1,480,072	+/-12,079
Male	563,720	+/-6,155	3,916,417	+/-12,571	1,537,424	+/-10,503	721,823	+/-6,867
Female	597,784	+/-6,842	4,149,409	+/-14,706	1,629,421	+/-10,584	758,249	+/-7,603
Native; born outside the United States:	32,355	+/-2,201	521,324	+/-9,938	104,496	+/-3,743	37,614	+/-1,932
Male	17,135	+/-1,485	254,102	+/-5,177	53,067	+/-2,776	19,207	+/-1,500
Female	15,220	+/-1,328	267,222	+/-6,555	51,429	+/-2,593	18,407	+/-1,253
Foreign born	130,790	+/-3,766	3,369,748	+/-17,842	841,282	+/-10,234	181,643	+/-4,796

Source: U.S. Census Bureau, 2005-2007 American Community Survey

4. Usability Issue: The geographic area did not stand out enough and was often overlooked until one or two tasks into the session.

   This issue may be an artifact of the test stimuli and not of the indicator tables themselves. Most participants did not notice that the table was for the small geographic area of Hays City, Kansas right away during the initial probe question or while completing the first task. Many participants also mentioned that they wished the name of the area was bolded or larger at the top of the table. It is also possible that the data reliability indicator's more prominent location for this round of testing helped users to notice it at the top of the table, but may have also distracted them from noticing the geographic area label. This possibility will be investigated in the eye–tracking data for the final version of this report. Having the participants start their tasks at the table is somewhat artificial and unique to the lab setting because real–world users would have accessed the table through American Fact Finder (AFF) or another Census Web site and would have had to choose their geography at that time. These participants were experienced ACS data users, so they were likely to have accessed ACS data through AFF before.

   Participant 11 said that the geographic area was buried a little lower than he would like to see it. Participant 13 answered first task before saying "hold on" and looking for the geographic area to verify that she was looking at the right table. During debriefing, she said about the baseline that this one did not distract her from noticing the geographic area. Participant 19 did not see the geographic area during task 2 for a long time and then said he did not read the text in the top left-hand corner and wished it was bold like the data reliability indicator. Participant 21 said that the geographic area should be bold as well. This participant continued to say that the geographic information was too overwhelmed by the legend.

Medium–Priority Issues

1. Usability Issue: Reference to 12–month periods in tables with multi–year estimates.

   Participant 16 pointed out that it was unclear how she should interpret the 12-month period referenced in Task 1 for this table, which included 3-year estimates. This participant believed that she could not answer the question because it was a 3-year dataset, although this information is valid and interpretable. This issue is out–of–scope for the current project, but may be the focus of future usability testing.

Low–Priority Issues

1. Usability Issue: Total Population and sample size

   Many participants (such as Participant 3) mentioned that they would like to see the total population estimate at the top of the table. Also, many participants expressed the wish that they could click on a link or look somewhere in the table to find the total sample size for the geographic area in order to get a better idea of how large the estimate they found in the table was.

2. Usability Issue:Intensity/saturation of colors may need to be adjusted for the Web

   Participants 1 and 10 mentioned that the colors used in the indicator were too distracting or unappealing and that we should refer to some Web design guidelines to improve them.

Usability Observations

1. Participants tended to notice the indicator as they progressed through the usability session and not necessarily right at the beginning of the first task. Perhaps including a "training" video or pop–up window explaining the reliability indicator would be helpful in getting people to use the indicator sooner. To avoid the frustration of having this training item pop up with each visit to the site, it could be offered before the user gets to the table using a prominently displayed hyperlink.

2. Participants found the answers to the ancestry–based questions more quickly after they had already answered one or two similar tasks. The tasks were randomized for each participant to account for this learning factor (see Table 1). The randomization of tasks should also account for the difficulty that some participants had with finding the geographic area associated with the table for the first task they performed.

3. Users stop looking at MOE as they progress through the tasks (just like the first round) One participant said using the indicator was "addictive" to do so and noticed that he did that.

4. Participants 3, 4, and 6 mentioned that reporting numbers for polls requires more precision, care, and double–checking sampling error than other uses of estimates. There may be a perception among the participant pool that polls are a more appropriate for using MOE, CV, and sampling error in general than for the type of tasks that we asked them to perform. Participant 6 explicitly mentioned that his "polling department" uses CV.

5. There was a recurring comment among participants that the MOE should not be larger than the estimate itself. This seemed to be a commonly held standard among this participant pool. Participant 9 mentioned that she would not take MOE into account, but when the MOE is so close or larger than the actual estimate, there is not much choice. She also mentioned that it would be a lot to explain to her readers.

6. Participants did mention repeatedly that the decision and what is "at stake" (government funding, holding a concert, reporting to state leaders, etc.) does make a difference in whether or not they would report a number and whether or not the MOE was important.

7. The color–coded reliability indicator was used by a few participants for Task 3, which asked the participant to compare the quality of estimates for German and Slovak ancestry in Hays, KS, who did not otherwise use the indicator. There was a trend overall for the participants to use the indicator and report its message for this comparison

task. Participant said, “Now I can use the color-coded column” when answering this task. Participants 4 and 5 relied solely on color coding to answer the third part of this question about reliability. Participants 6 and 10 used the color indicator for task 3, but then stopped looking at error again after that task. Participant 17 used the indicator for the first time on task 7 and only mentioned that an estimate looked good on task 9 later in the session.

5 Limitations

Although the original plan was to recruit 60 ACS data users from the Washington, D.C. area, only 21 could be brought into the usability lab during the course of the study. The small sample size means that the statistical tests presented in this report have low statistical power. These participants were also experienced ACS data users; novice users may have had a difference experience using the tables.

Additionally, the ACS produces many kinds of tables in a variety of different formats, while this test only examined one simple table style. The addition of a color-coded data reliability indicator on some of the other ACS tables would likely create a far "busier" appearance. This issue is currently examined in the third round of testing by testing several different types of ACS tables with the data reliability indicator added.

A related limitation is that only totals or levels were evaluated in this study (e.g., the number of persons of Danish descent - 69 in Hays /111. The ACS also reports the characteristic distribution of the population (e.g., the number of persons of Danish descent - 0.3 percent, /0.6 percentage points). This limitation is also being addressed in the third round of testing, which includes some tables with both estimates and percentages and tasks that ask the participant to interpret both.

Some comments made by participants during testing did express positive opinions about the baseline tables and some negative opinions about the "busy-ness" that the color-coded tables had. This indicates that there are strengths to the baseline table and some weaknesses to the color-coded tables.

6 Discussion and Future Directions

Although there was a difference in terms of preference between the baseline table and the prototype tables, the baseline table did not have significantly lower accuracy and satisfaction scores than the other two tables. In fact, the baseline table had a higher satisfaction score than the three–level table in a post-hoc test. Because the participants were experienced ACS data users as participants for this round of testing, future testing may include novice ACS data users to examine their difference in performance of the baseline and prototype tables.

Overall, the baseline table, three–level prototype, and four–level prototype did not differ much in terms of accuracy or efficiency, although the prototypes were reported as preferred more often. However, the four–level indicator and baseline tables were associated with significantly higher rates of participants mentioning and choosing to include the message contained in the indicator or the MOE as part of their final answer (a.k.a. report it) over the three–level prototype. Participants in both the four–level "good" table and the four–level "reliable" table conditions were significantly more likely to report the message from the color–coded reliability indicator than the three-level indicator. Also, participants in both the four–level "good" table and the four–level "reliable" table conditions were significantly more likely to explicitly report the message from the color–coded reliability indicator along with the estimate than participants in the three-level indicator condition. There was no significant difference between the two four–level tables themselves on this variable.

The four–level "reliable" table was also associated with significantly higher overall QUIS scores than the three–level table. A one–way ANOVA showed that there was at least one significant difference among the tables on this variable. Pairwise post-hoc Tukey's tests showed that both the baseline and four–level "reliable" tables had significantly higher QUIS scores than the three–level table. So, satisfaction as measured by the QUIS instrument was significantly higher for these two tables than for the three–level table.

Eye–tracking analyses revealed that participants do not look at the MOE column on the three–level table as frequently or as long as they look at the MOE column for the other tables. Future testing will explore the psychological relevance of the three–level "stoplight" color coding system and the possibility that it carries more implicit information than the four–level coding system.

A third round of usability testing is planned to examine the issues discussed in this report further.

7 Acknowledgements

Many thanks to Debbie Griffin, Nicole Scanniello, Tony Tersine, Michael Springer, Erica Olmsted-Hawala, Gregg Robinson, and Betty Murphy for their comments on this paper.

References

Ashenfelter, K. T., Beck, J., & Murphy, E. D. (2009). Final report for first-round usability testing of data-reliability indicator prototypes. *Statistical Research Division Report Series, Report SSM2009/01*. Available from `http://www.census.gov/srd/papers/pdf/ssm2009-01.pdf`

Cahoon, L., Donnalley, G., Gore, E., Kostanich, D., Runyan, R., Detlefsen, R., et al. (2007). Quality requirements for releasing data products. *U.S. Census Bureau official documentation.*

Chin, J. P., Diehl, V., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of CHI 88: Human Factors in Computing Systems*, 213-218.

Greitzer, F. (2005). Toward the development of cognitive task difficulty metrics to support intelligence analysis research. *Proceedings of the IEEE 2005: International Conference of Cognitive Informatics (ICCI 2005).*

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122-149.

Just, M. A., Carpenter, P. A., & Miyake, A. (2003). Neuroindices of cognitive workload: Neuroimaging, pupillometic, and event-related potential studies of brain work. *Theoretical Issues in Ergonomical Science*, *4*, 56-88.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice Hall.

Navon, D., & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review*, *86*, 214-255.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*, 44-64.

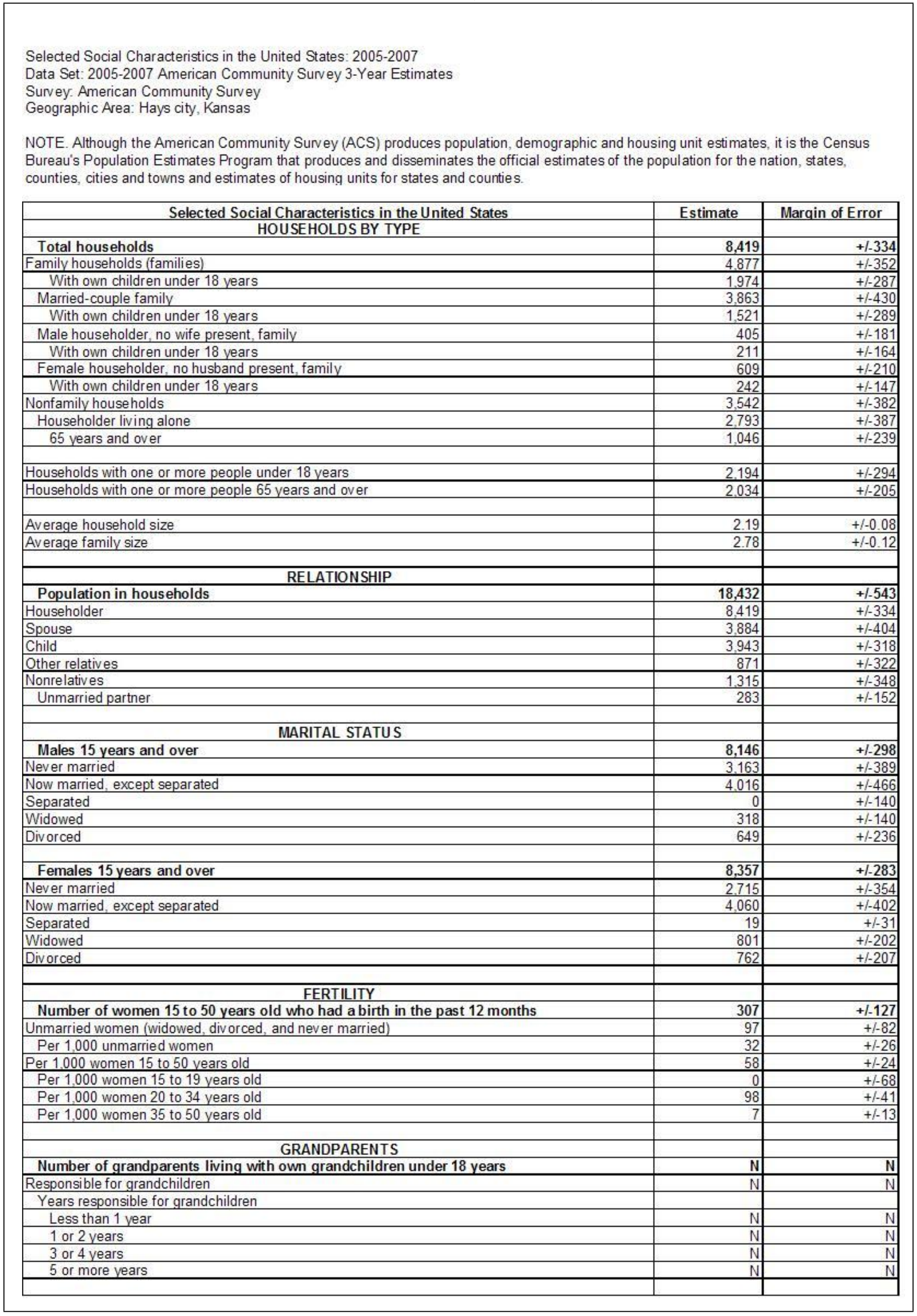
Rubio, S., Daz, E., Martn, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology: An International Review*, *53*, 6186.

Tobii Technology, I. (2008). *Tobii studio enterprise edition software.*

Whitford, D., & Weinberg, D. (2008). Proposal to highlight american community survey data with a data confidence indicator. *U.S. Census Bureau Document*.

8 Appendix A: Tables Shown During Testing

Figure 21. Baseline (Previous) ACS Table

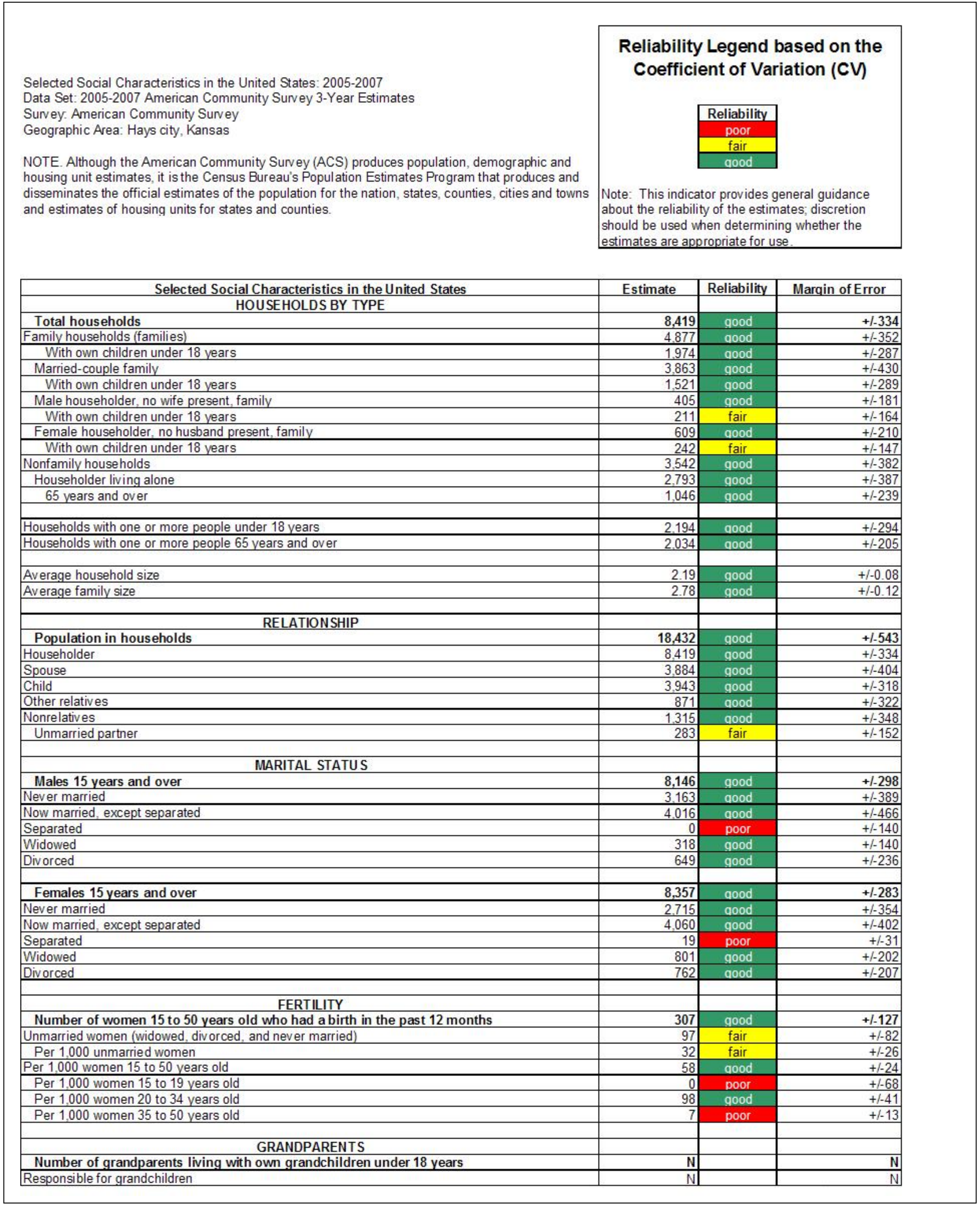


Selected Social Characteristics in the United States: 2005-2007
Data Set: 2005-2007 American Community Survey 3-Year Estimates
Survey: American Community Survey
Geographic Area: Hays city, Kansas

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Selected Social Characteristics in the United States	Estimate	Margin of Error
HOUSEHOLDS BY TYPE		
Total households	**8,419**	**+/-334**
Family households (families)	4,877	+/-352
With own children under 18 years	1,974	+/-287
Married-couple family	3,863	+/-430
With own children under 18 years	1,521	+/-289
Male householder, no wife present, family	405	+/-181
With own children under 18 years	211	+/-164
Female householder, no husband present, family	609	+/-210
With own children under 18 years	242	+/-147
Nonfamily households	3,542	+/-382
Householder living alone	2,793	+/-387
65 years and over	1,046	+/-239
Households with one or more people under 18 years	2,194	+/-294
Households with one or more people 65 years and over	2,034	+/-205
Average household size	2.19	+/-0.08
Average family size	2.78	+/-0.12
RELATIONSHIP		
Population in households	**18,432**	**+/-543**
Householder	8,419	+/-334
Spouse	3,884	+/-404
Child	3,943	+/-318
Other relatives	871	+/-322
Nonrelatives	1,315	+/-348
Unmarried partner	283	+/-152
MARITAL STATUS		
Males 15 years and over	**8,146**	**+/-298**
Never married	3,163	+/-389
Now married, except separated	4,016	+/-466
Separated	0	+/-140
Widowed	318	+/-140
Divorced	649	+/-236
Females 15 years and over	**8,357**	**+/-283**
Never married	2,715	+/-354
Now married, except separated	4,060	+/-402
Separated	19	+/-31
Widowed	801	+/-202
Divorced	762	+/-207
FERTILITY		
Number of women 15 to 50 years old who had a birth in the past 12 months	**307**	**+/-127**
Unmarried women (widowed, divorced, and never married)	97	+/-82
Per 1,000 unmarried women	32	+/-26
Per 1,000 women 15 to 50 years old	58	+/-24
Per 1,000 women 15 to 19 years old	0	+/-68
Per 1,000 women 20 to 34 years old	98	+/-41
Per 1,000 women 35 to 50 years old	7	+/-13
GRANDPARENTS		
Number of grandparents living with own grandchildren under 18 years	**N**	**N**
Responsible for grandchildren	N	N
Years responsible for grandchildren		
Less than 1 year	N	N
1 or 2 years	N	N
3 or 4 years	N	N
5 or more years	N	N

Figure 22. Three–Level Prototype Table



Selected Social Characteristics in the United States: 2005-2007
Data Set: 2005-2007 American Community Survey 3-Year Estimates
Survey: American Community Survey
Geographic Area: Hays city, Kansas

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

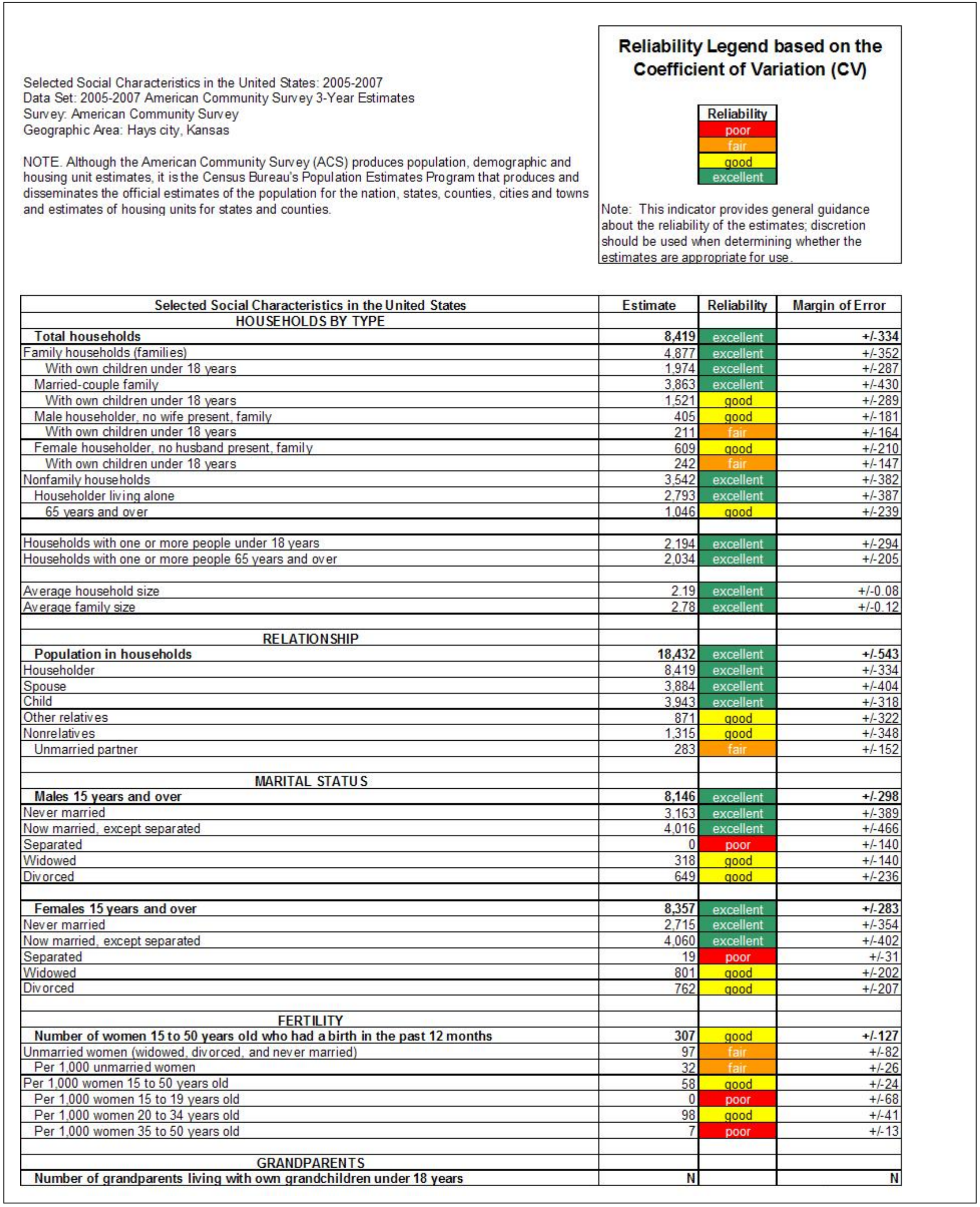
Reliability Legend based on the Coefficient of Variation (CV)

Reliability
poor
fair
good

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

Selected Social Characteristics in the United States	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	**8,419**	good	**+/-334**
Family households (families)	4,877	good	+/-352
With own children under 18 years	1,974	good	+/-287
Married-couple family	3,863	good	+/-430
With own children under 18 years	1,521	good	+/-289
Male householder, no wife present, family	405	good	+/-181
With own children under 18 years	211	fair	+/-164
Female householder, no husband present, family	609	good	+/-210
With own children under 18 years	242	fair	+/-147
Nonfamily households	3,542	good	+/-382
Householder living alone	2,793	good	+/-387
65 years and over	1,046	good	+/-239
Households with one or more people under 18 years	2,194	good	+/-294
Households with one or more people 65 years and over	2,034	good	+/-205
Average household size	2.19	good	+/-0.08
Average family size	2.78	good	+/-0.12
RELATIONSHIP			
Population in households	**18,432**	good	**+/-543**
Householder	8,419	good	+/-334
Spouse	3,884	good	+/-404
Child	3,943	good	+/-318
Other relatives	871	good	+/-322
Nonrelatives	1,315	good	+/-348
Unmarried partner	283	fair	+/-152
MARITAL STATUS			
Males 15 years and over	**8,146**	good	**+/-298**
Never married	3,163	good	+/-389
Now married, except separated	4,016	good	+/-466
Separated	0	poor	+/-140
Widowed	318	good	+/-140
Divorced	649	good	+/-236
Females 15 years and over	**8,357**	good	**+/-283**
Never married	2,715	good	+/-354
Now married, except separated	4,060	good	+/-402
Separated	19	poor	+/-31
Widowed	801	good	+/-202
Divorced	762	good	+/-207
FERTILITY			
Number of women 15 to 50 years old who had a birth in the past 12 months	**307**	good	**+/-127**
Unmarried women (widowed, divorced, and never married)	97	fair	+/-82
Per 1,000 unmarried women	32	fair	+/-26
Per 1,000 women 15 to 50 years old	58	good	+/-24
Per 1,000 women 15 to 19 years old	0	poor	+/-68
Per 1,000 women 20 to 34 years old	98	good	+/-41
Per 1,000 women 35 to 50 years old	7	poor	+/-13
GRANDPARENTS			
Number of grandparents living with own grandchildren under 18 years	**N**		**N**
Responsible for grandchildren	N		N

Figure 23. Four–Level "Good" Prototype Table



Selected Social Characteristics in the United States: 2005-2007
Data Set: 2005-2007 American Community Survey 3-Year Estimates
Survey: American Community Survey
Geographic Area: Hays city, Kansas

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

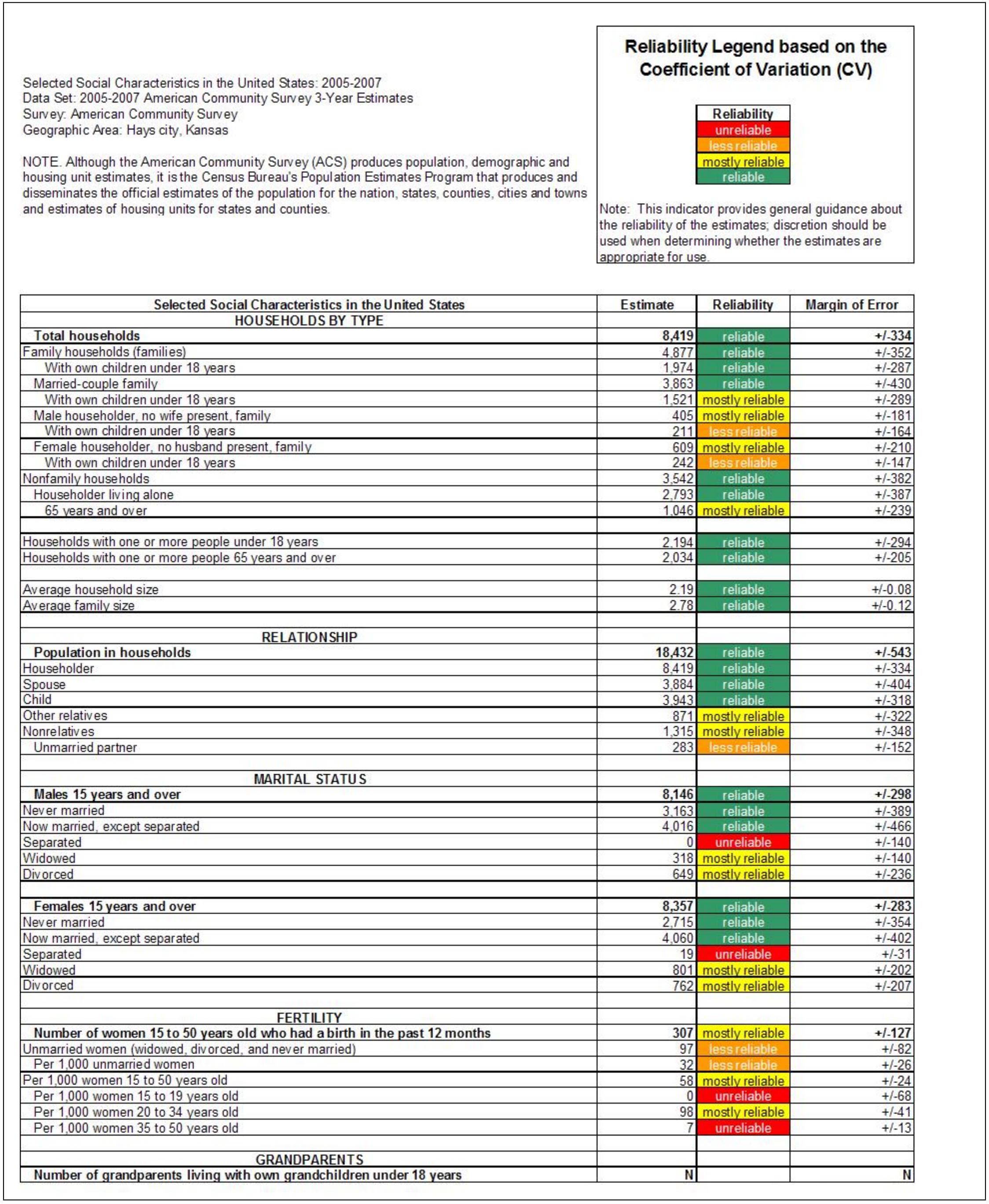
Reliability Legend based on the Coefficient of Variation (CV)

Reliability
poor
fair
good
excellent

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

Selected Social Characteristics in the United States	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	**8,419**	excellent	**+/-334**
Family households (families)	4,877	excellent	+/-352
With own children under 18 years	1,974	excellent	+/-287
Married-couple family	3,863	excellent	+/-430
With own children under 18 years	1,521	good	+/-289
Male householder, no wife present, family	405	good	+/-181
With own children under 18 years	211	fair	+/-164
Female householder, no husband present, family	609	good	+/-210
With own children under 18 years	242	fair	+/-147
Nonfamily households	3,542	excellent	+/-382
Householder living alone	2,793	excellent	+/-387
65 years and over	1,046	good	+/-239
Households with one or more people under 18 years	2,194	excellent	+/-294
Households with one or more people 65 years and over	2,034	excellent	+/-205
Average household size	2.19	excellent	+/-0.08
Average family size	2.78	excellent	+/-0.12
RELATIONSHIP			
Population in households	**18,432**	excellent	**+/-543**
Householder	8,419	excellent	+/-334
Spouse	3,884	excellent	+/-404
Child	3,943	excellent	+/-318
Other relatives	871	good	+/-322
Nonrelatives	1,315	good	+/-348
Unmarried partner	283	fair	+/-152
MARITAL STATUS			
Males 15 years and over	**8,146**	excellent	**+/-298**
Never married	3,163	excellent	+/-389
Now married, except separated	4,016	excellent	+/-466
Separated	0	poor	+/-140
Widowed	318	good	+/-140
Divorced	649	good	+/-236
Females 15 years and over	**8,357**	excellent	**+/-283**
Never married	2,715	excellent	+/-354
Now married, except separated	4,060	excellent	+/-402
Separated	19	poor	+/-31
Widowed	801	good	+/-202
Divorced	762	good	+/-207
FERTILITY			
Number of women 15 to 50 years old who had a birth in the past 12 months	**307**	good	**+/-127**
Unmarried women (widowed, divorced, and never married)	97	fair	+/-82
Per 1,000 unmarried women	32	fair	+/-26
Per 1,000 women 15 to 50 years old	58	good	+/-24
Per 1,000 women 15 to 19 years old	0	poor	+/-68
Per 1,000 women 20 to 34 years old	98	good	+/-41
Per 1,000 women 35 to 50 years old	7	poor	+/-13
GRANDPARENTS			
Number of grandparents living with own grandchildren under 18 years	**N**		**N**

Figure 24. Four–Level “Reliable” Prototype Table



Selected Social Characteristics in the United States: 2005-2007
Data Set: 2005-2007 American Community Survey 3-Year Estimates
Survey: American Community Survey
Geographic Area: Hays city, Kansas

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Reliability Legend based on the Coefficient of Variation (CV)

Reliability
unreliable
less reliable
mostly reliable
reliable

Note: This indicator provides general guidance about the reliability of the estimates; discretion should be used when determining whether the estimates are appropriate for use.

Selected Social Characteristics in the United States	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	**8,419**	reliable	**+/-334**
Family households (families)	4,877	reliable	+/-352
With own children under 18 years	1,974	reliable	+/-287
Married-couple family	3,863	reliable	+/-430
With own children under 18 years	1,521	mostly reliable	+/-289
Male householder, no wife present, family	405	mostly reliable	+/-181
With own children under 18 years	211	less reliable	+/-164
Female householder, no husband present, family	609	mostly reliable	+/-210
With own children under 18 years	242	less reliable	+/-147
Nonfamily households	3,542	reliable	+/-382
Householder living alone	2,793	reliable	+/-387
65 years and over	1,046	mostly reliable	+/-239
Households with one or more people under 18 years	2,194	reliable	+/-294
Households with one or more people 65 years and over	2,034	reliable	+/-205
Average household size	2.19	reliable	+/-0.08
Average family size	2.78	reliable	+/-0.12
RELATIONSHIP			
Population in households	**18,432**	reliable	**+/-543**
Householder	8,419	reliable	+/-334
Spouse	3,884	reliable	+/-404
Child	3,943	reliable	+/-318
Other relatives	871	mostly reliable	+/-322
Nonrelatives	1,315	mostly reliable	+/-348
Unmarried partner	283	less reliable	+/-152
MARITAL STATUS			
Males 15 years and over	**8,146**	reliable	**+/-298**
Never married	3,163	reliable	+/-389
Now married, except separated	4,016	reliable	+/-466
Separated	0	unreliable	+/-140
Widowed	318	mostly reliable	+/-140
Divorced	649	mostly reliable	+/-236
Females 15 years and over	**8,357**	reliable	**+/-283**
Never married	2,715	reliable	+/-354
Now married, except separated	4,060	reliable	+/-402
Separated	19	unreliable	+/-31
Widowed	801	mostly reliable	+/-202
Divorced	762	mostly reliable	+/-207
FERTILITY			
Number of women 15 to 50 years old who had a birth in the past 12 months	**307**	mostly reliable	**+/-127**
Unmarried women (widowed, divorced, and never married)	97	less reliable	+/-82
Per 1,000 unmarried women	32	less reliable	+/-26
Per 1,000 women 15 to 50 years old	58	mostly reliable	+/-24
Per 1,000 women 15 to 19 years old	0	unreliable	+/-68
Per 1,000 women 20 to 34 years old	98	mostly reliable	+/-41
Per 1,000 women 35 to 50 years old	7	unreliable	+/-13
GRANDPARENTS			
Number of grandparents living with own grandchildren under 18 years	**N**		**N**

9 Appendix B: Tasks

Initial Probe Question: What is the first thing that you noticed about this table?

1. Your supervisor asks you to find some information about the number of women ages 15 to 50 who gave birth in the past 12 months for your hometown of Hays, KS. What information would you report to your supervisor?

   Task 1 Difficulty: Medium (find information; make a judgment about acceptability of data reliability)

   Average Participant Difficulty Rating: 2.8

2. You are researching background information for a paper and need to find the number of people of West Indian descent in Hays, KS. What do you report in the paper based on your findings in the tables?

   Task 2 Difficulty: Medium (find information; make a judgment about acceptability of data reliability)

   Average Participant Difficulty Rating: 2.5

3. Find the total number of people with German ancestry and the total number of people with Slovak ancestry for Hays, KS. Which category of ancestry do you think is a better estimate in terms of data quality? Please explain why you think this is a better estimate of data quality.

   Task 3 Difficulty: Hard (find information, compare 2 estimates and their associated reliability and make a judgment about acceptability of data reliability)

   Average Participant Difficulty Rating: 2.3

4. You work for a major corporation that sells children's products, music, and videos. Your job is to organize a concert in the Hays, KS area and your boss wants you to find out:

   (a) How many family households in this area have children under 18 years old? Would you report this estimate? Why or why not?

   (b) What is the average family size in Hays, KS? Would you report this estimate? Why or why not?

   (c) How many nursery school, kindergarten, and elementary school students are enrolled in this area? Would you report this estimate? Why or why not?

(d) Based on the information you found, your boss wants to know whether you think Hays, KS is a good place to hold this concert.

Task 4 Difficulty: Hard (find multiple pieces of information, and make a judgment about acceptability of data reliability, integrate information)

Average Participant Difficulty Rating: 4.4

5. You are asked to report to state leaders the number of people of Italian descent living in Hays. What answer would you give them?
   Task 5 Difficulty: Medium (find information; make a judgment about acceptability of data reliability)

   Average Participant Difficulty Rating: 2.2

6. The mayor of Hays said that if there are more than 300 people ages 5 to 15 with disabilities in Hays, the city might be eligible to receive some government funding to develop programs for the disabled. He asks you if the there are at least 300 people in this age group with disabilities in Hays. What would you tell him using ACS data?

   Task 6 Difficulty: Hard (find information, and make a judgment about acceptability of data reliability and context of problem that involves money and impacts people (socially complex), integrate information)

   Average Participant Difficulty Rating: 3.3

7. The Danish embassy wants a listing of all cities with more than 200 people of Danish descent. Would you include the city of Hays based on the ACS data?

   Task 7 Difficulty: Medium (find information, and make a judgment about acceptability of data reliability)

   Average Participant Difficulty Rating: 2.3

8. Cities with less than 200 people of French Canadian descent will engage in an outreach program designed to attract more people of French Canadian descent. Does Hays qualify based on ACS data?

   Task 8 Difficulty: Medium (find information, and make a judgment about acceptability of data reliability)

   Average Participant Difficulty Rating: 2.5

9. You are writing a news article about voter turnout in the 2008 presidential election and want to find out how many civilians are 18 or older in your home town of Hays, KS. What results do you find in the table?

   Task 9 Difficulty: Hard (find information, and make a judgment about acceptability of data reliability, consider the year that the data were collected; integrate all of this information)

   Average Participant Difficulty Rating: 5.0

9.1 Task Difficulty Rating Metric

The proposed metric for assessing task difficulty incorporates the research findings from the field of cognitive science, which indicate that "difficult" tasks require more mental/cognitive work (Just, Carpenter, & Miyake, 2003). Cognitive theory posits that the more cognitive work is required for a task, the more mental capacity or resources must be used in order to complete this task (Just & Carpenter, 1992; Just et al., 2003; Kahneman, 1973; Norman & Bobrow, 1975; Navon & Gopher, 1979). Similar research in the field of intelligence analysis research also suggests some guidelines for what constitutes a difficult task (Greitzer, 2005). NASA also developed the NASA-TLX instrument (Rubio, Daz, Martn, & Puente, 2004) for evaluating mental effort, but it is mainly used for evaluating physical tools such as airplane cockpits. However, no set of metrics has been constructed specifically for use in usability studies. Since efficiency score standards are usually based on the difficulty of the task, it is important that an objective metric be used to rate these tasks. The tasks assigned to participants in a usability test usually require them to perform problem-solving tasks using working memory. The a priori difficulty rating assigned to the current tasks will be evaluated for validation purposes based on the results of the study (e.g., subjective rating scale, eye-tracking data, pupillometrics, etc.). For the current test plan, a task will be considered "hard" if it requires the participant to perform two or more of the following cognitive tasks:

- Compare and contrast concepts (especially if one or more concepts need to be retrieved from long-term memory) (Greitzer, 2005).
- Find and interpret content of the Web site/data table
- Perform deep/complex navigation (following a series of more than 2 links, or a complex or unintuitive series of links)
- Answer multiple sub-questions for one main question (e.g., Task 1, parts A, B, and C).
- Answer a question that requires advanced experience or knowledge (e.g., a challenging statistical question)
- Perform spatial comparisons or rotations (Just et al., 2003).

- Answer a question based on what participant thinks would be best for another person or group of people (or most people); socially complex thinking required (Greitzer, 2005).

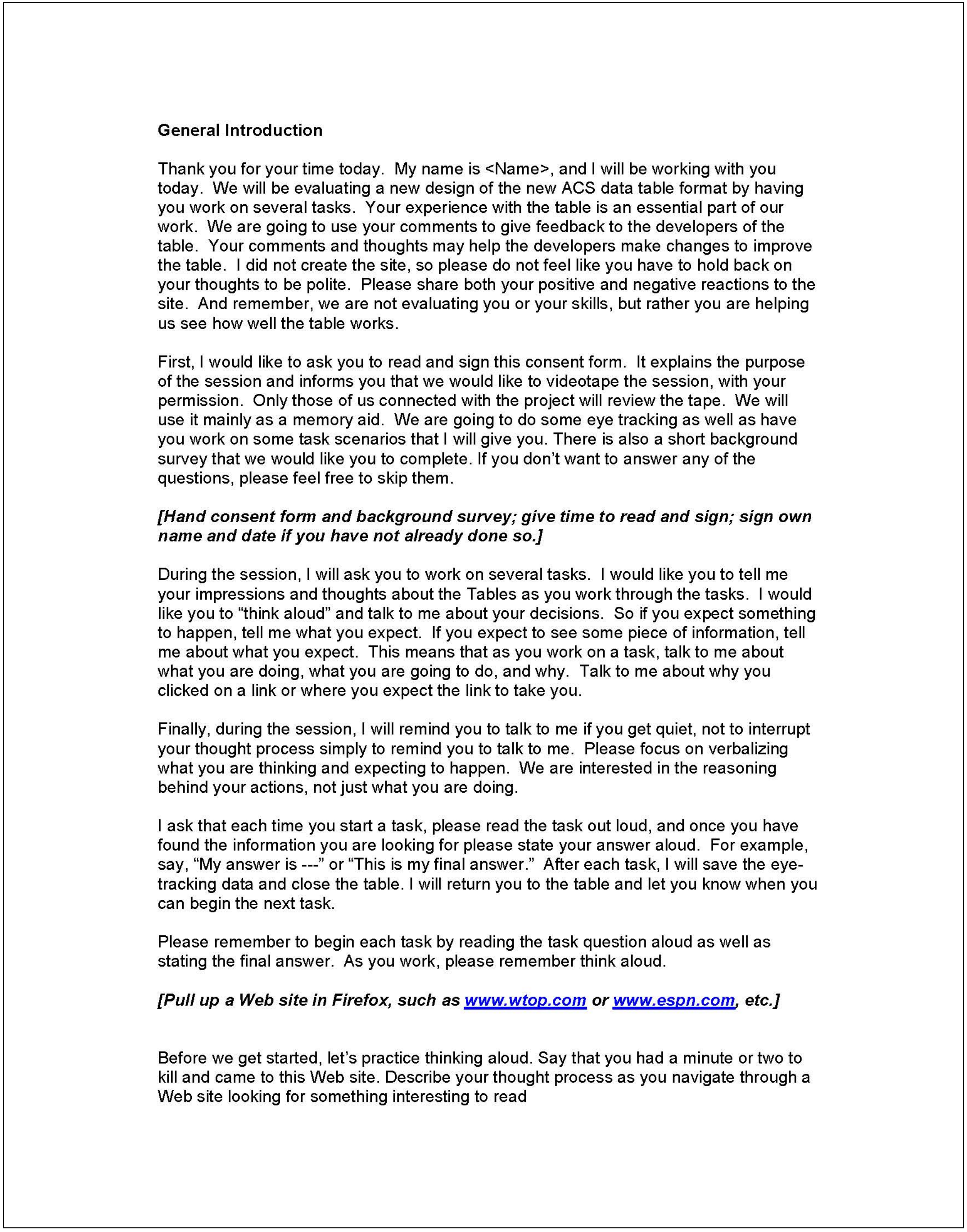
A task will be considered of medium difficulty if it requires the participant to perform only one of the above cognitive tasks.

Typically, an easy task using a Web site or data table will involve the following:

- Visually searching for key words
- Shallow navigation (one or two links deep)
- Reporting numbers from a table without interpretation

10 Appendix C: General Protocol

Figure 25. Data Reliability Indicator General Protocol Page 1



General Introduction

Thank you for your time today. My name is <Name>, and I will be working with you today. We will be evaluating a new design of the new ACS data table format by having you work on several tasks. Your experience with the table is an essential part of our work. We are going to use your comments to give feedback to the developers of the table. Your comments and thoughts may help the developers make changes to improve the table. I did not create the site, so please do not feel like you have to hold back on your thoughts to be polite. Please share both your positive and negative reactions to the site. And remember, we are not evaluating you or your skills, but rather you are helping us see how well the table works.

First, I would like to ask you to read and sign this consent form. It explains the purpose of the session and informs you that we would like to videotape the session, with your permission. Only those of us connected with the project will review the tape. We will use it mainly as a memory aid. We are going to do some eye tracking as well as have you work on some task scenarios that I will give you. There is also a short background survey that we would like you to complete. If you don't want to answer any of the questions, please feel free to skip them.

[Hand consent form and background survey; give time to read and sign; sign own name and date if you have not already done so.]

During the session, I will ask you to work on several tasks. I would like you to tell me your impressions and thoughts about the Tables as you work through the tasks. I would like you to "think aloud" and talk to me about your decisions. So if you expect something to happen, tell me what you expect. If you expect to see some piece of information, tell me about what you expect. This means that as you work on a task, talk to me about what you are doing, what you are going to do, and why. Talk to me about why you clicked on a link or where you expect the link to take you.

Finally, during the session, I will remind you to talk to me if you get quiet, not to interrupt your thought process simply to remind you to talk to me. Please focus on verbalizing what you are thinking and expecting to happen. We are interested in the reasoning behind your actions, not just what you are doing.

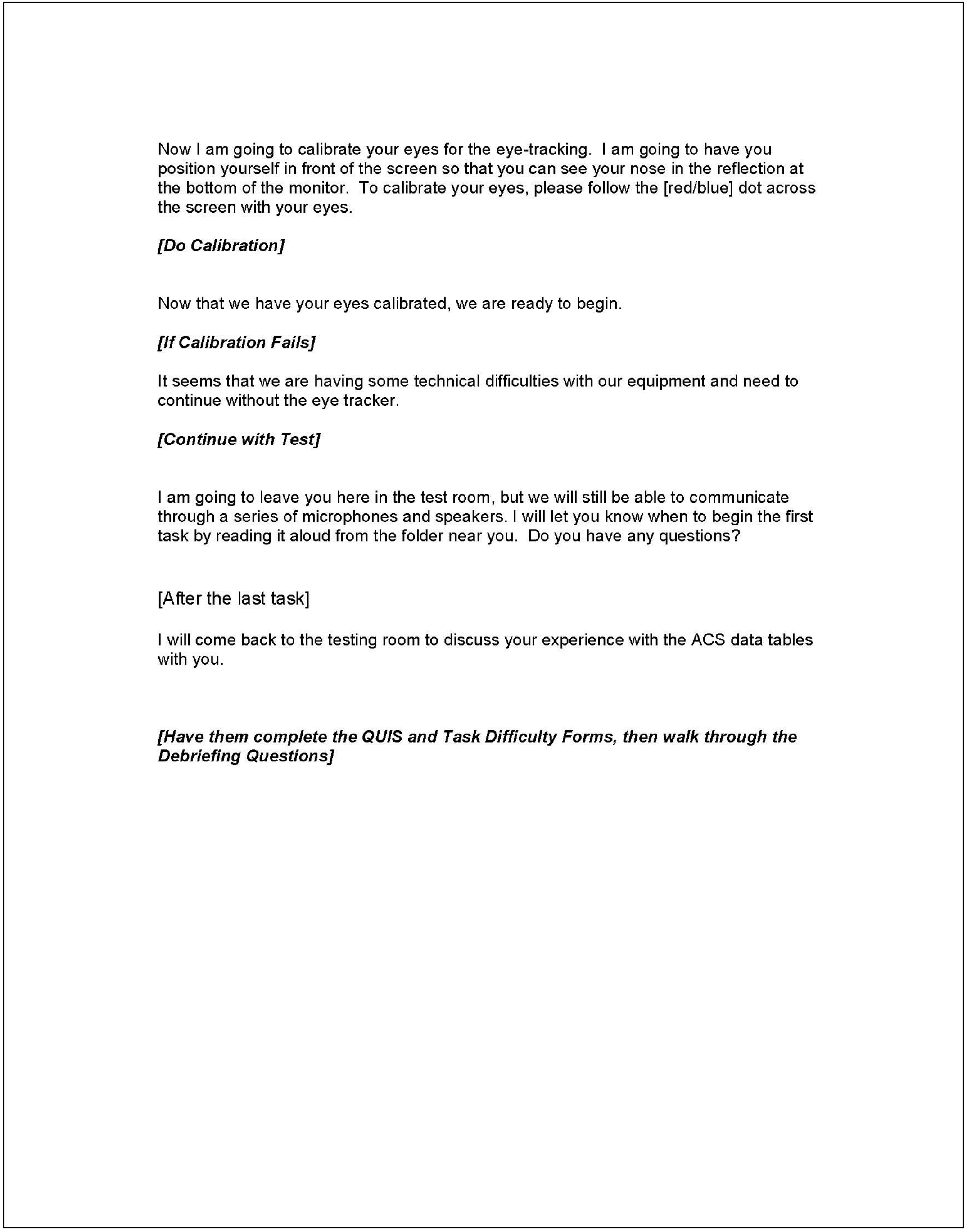
I ask that each time you start a task, please read the task out loud, and once you have found the information you are looking for please state your answer aloud. For example, say, "My answer is ---" or "This is my final answer." After each task, I will save the eye-tracking data and close the table. I will return you to the table and let you know when you can begin the next task.

Please remember to begin each task by reading the task question aloud as well as stating the final answer. As you work, please remember think aloud.

[Pull up a Web site in Firefox, such as www.wtop.com or www.espn.com, etc.]

Before we get started, let's practice thinking aloud. Say that you had a minute or two to kill and came to this Web site. Describe your thought process as you navigate through a Web site looking for something interesting to read

Figure 26. Data Reliability Indicator General Protocol Page 2



Now I am going to calibrate your eyes for the eye-tracking. I am going to have you position yourself in front of the screen so that you can see your nose in the reflection at the bottom of the monitor. To calibrate your eyes, please follow the [red/blue] dot across the screen with your eyes.

[Do Calibration]

Now that we have your eyes calibrated, we are ready to begin.

[If Calibration Fails]

It seems that we are having some technical difficulties with our equipment and need to continue without the eye tracker.

[Continue with Test]

I am going to leave you here in the test room, but we will still be able to communicate through a series of microphones and speakers. I will let you know when to begin the first task by reading it aloud from the folder near you. Do you have any questions?

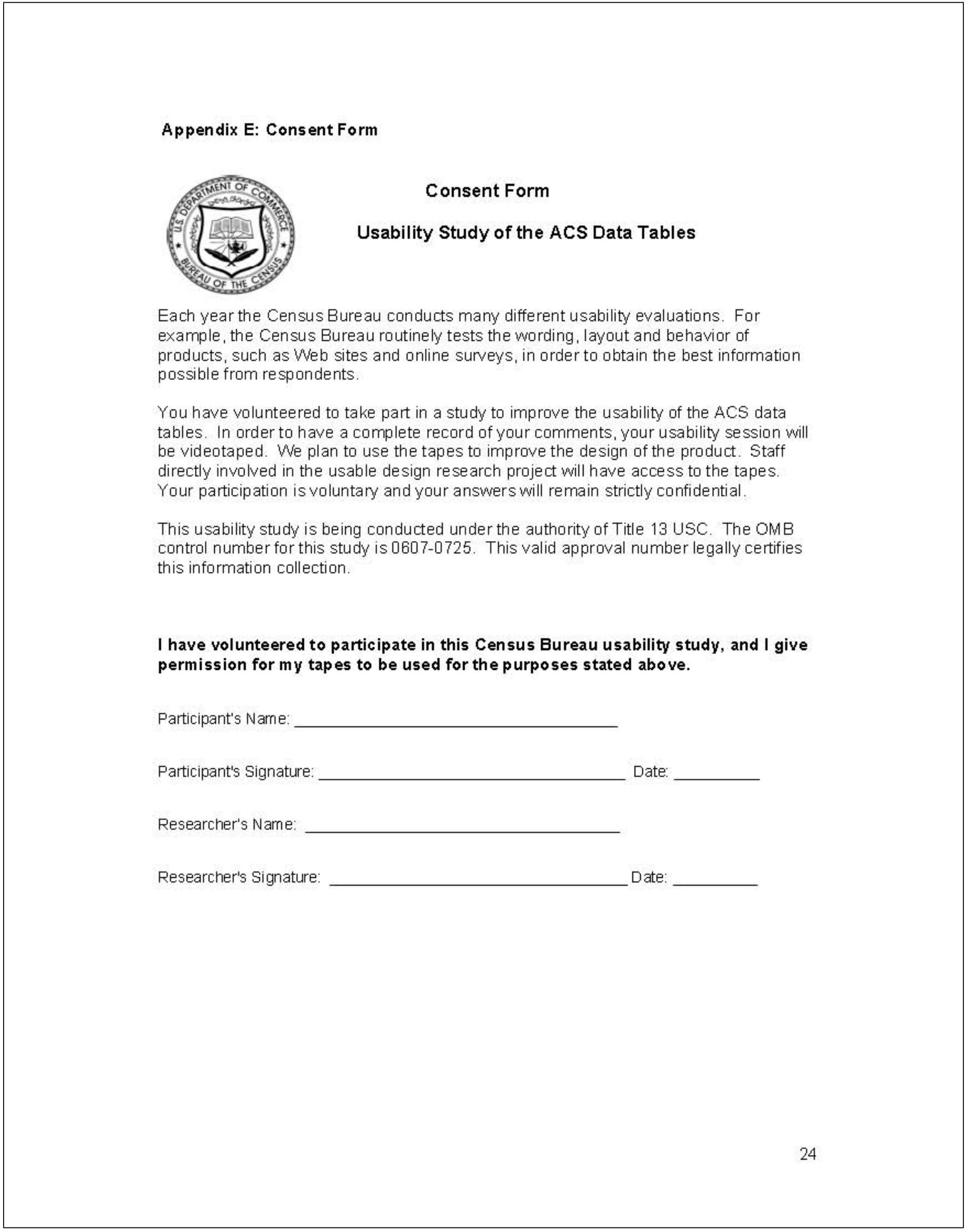
[After the last task]

I will come back to the testing room to discuss your experience with the ACS data tables with you.

[Have them complete the QUIS and Task Difficulty Forms, then walk through the Debriefing Questions]

11 Appendix D: Consent Form

Figure 27. Consent Form for Current Study



Appendix E: Consent Form

Consent Form

Usability Study of the ACS Data Tables

Each year the Census Bureau conducts many different usability evaluations. For example, the Census Bureau routinely tests the wording, layout and behavior of products, such as Web sites and online surveys, in order to obtain the best information possible from respondents.

You have volunteered to take part in a study to improve the usability of the ACS data tables. In order to have a complete record of your comments, your usability session will be videotaped. We plan to use the tapes to improve the design of the product. Staff directly involved in the usable design research project will have access to the tapes. Your participation is voluntary and your answers will remain strictly confidential.

This usability study is being conducted under the authority of Title 13 USC. The OMB control number for this study is 0607-0725. This valid approval number legally certifies this information collection.

I have volunteered to participate in this Census Bureau usability study, and I give permission for my tapes to be used for the purposes stated above.

Participant's Name: ______________________________

Participant's Signature: ____________________________ Date: ________

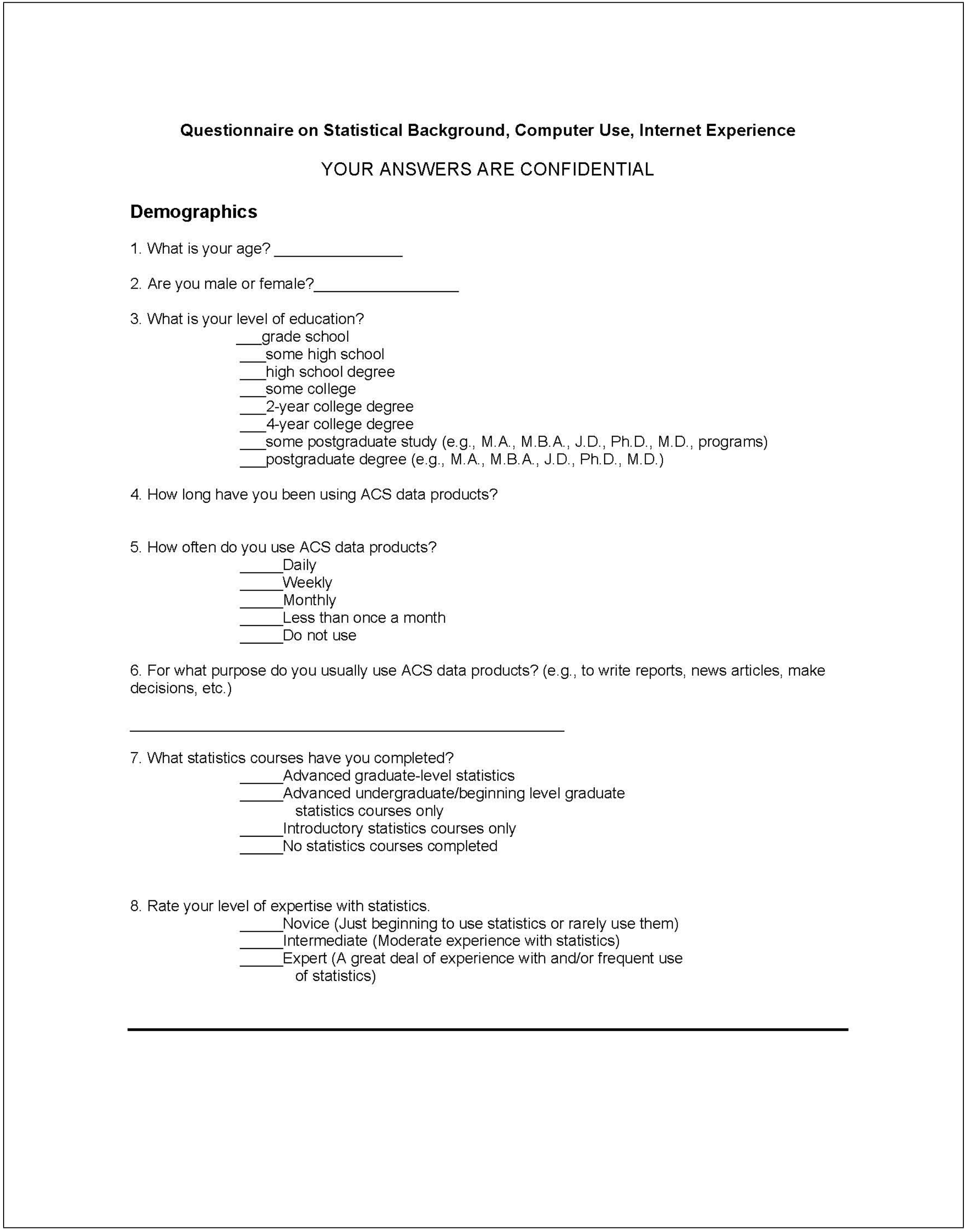
Researcher's Name: ______________________________

Researcher's Signature: ____________________________ Date: ________

24

12 Appendix E: Background Survey

Figure 28. Background Survey Page 1



Questionnaire on Statistical Background, Computer Use, Internet Experience

YOUR ANSWERS ARE CONFIDENTIAL

Demographics

1. What is your age? _______________

2. Are you male or female?_________________

3. What is your level of education?
___grade school
___some high school
___high school degree
___some college
___2-year college degree
___4-year college degree
___some postgraduate study (e.g., M.A., M.B.A., J.D., Ph.D., M.D., programs)
___postgraduate degree (e.g., M.A., M.B.A., J.D., Ph.D., M.D.)

4. How long have you been using ACS data products?

5. How often do you use ACS data products?
_____Daily
_____Weekly
_____Monthly
_____Less than once a month
_____Do not use

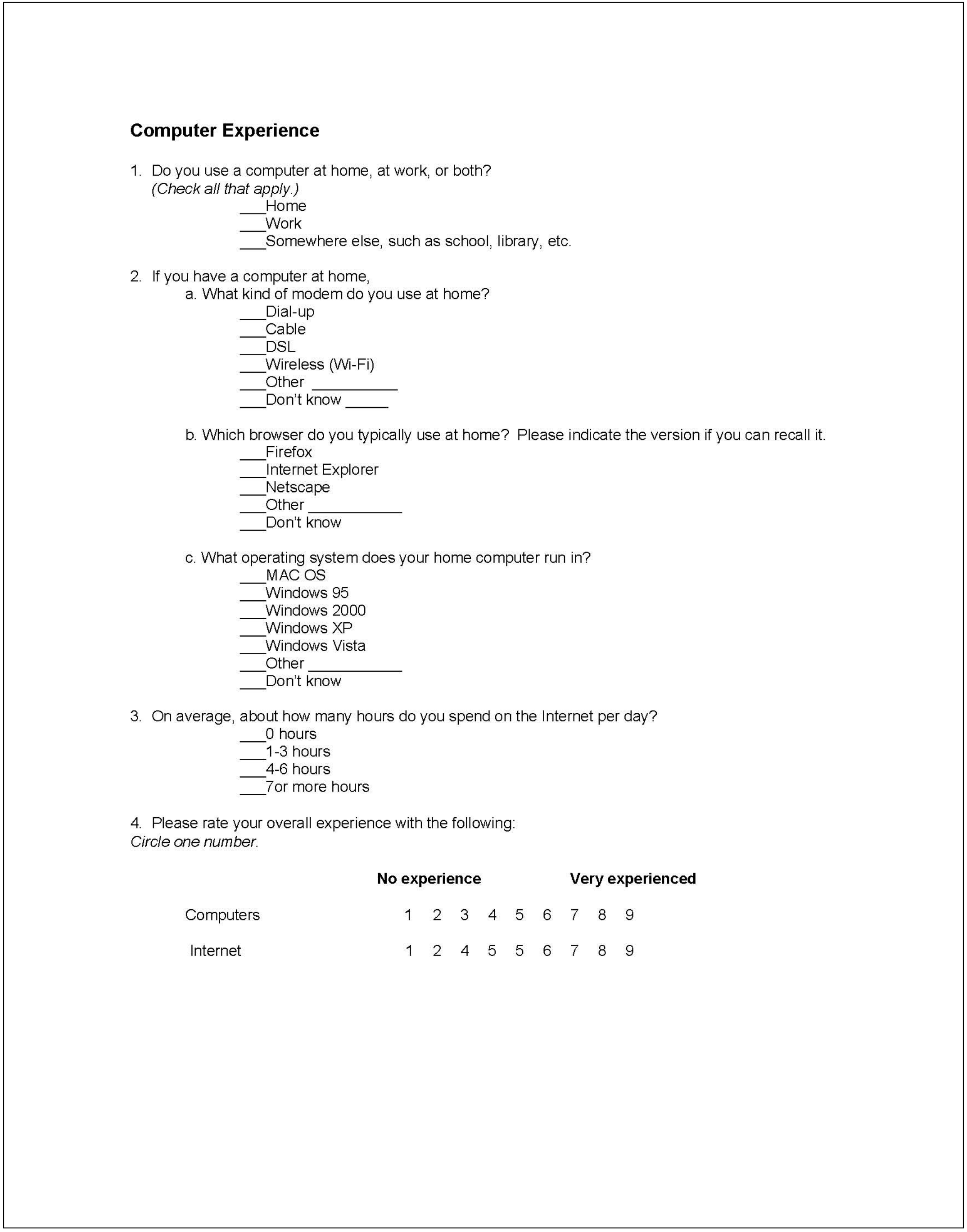
6. For what purpose do you usually use ACS data products? (e.g., to write reports, news articles, make decisions, etc.)

____

7. What statistics courses have you completed?
_____Advanced graduate-level statistics
_____Advanced undergraduate/beginning level graduate statistics courses only
_____Introductory statistics courses only
_____No statistics courses completed

8. Rate your level of expertise with statistics.
_____Novice (Just beginning to use statistics or rarely use them)
_____Intermediate (Moderate experience with statistics)
_____Expert (A great deal of experience with and/or frequent use of statistics)

Figure 29. Background Survey Page 2



Computer Experience

1. Do you use a computer at home, at work, or both?
(Check all that apply.)
___Home
___Work
___Somewhere else, such as school, library, etc.

2. If you have a computer at home,
a. What kind of modem do you use at home?
___Dial-up
___Cable
___DSL
___Wireless (Wi-Fi)
___Other __________
___Don't know _____

b. Which browser do you typically use at home? Please indicate the version if you can recall it.
___Firefox
___Internet Explorer
___Netscape
___Other ___________
___Don't know

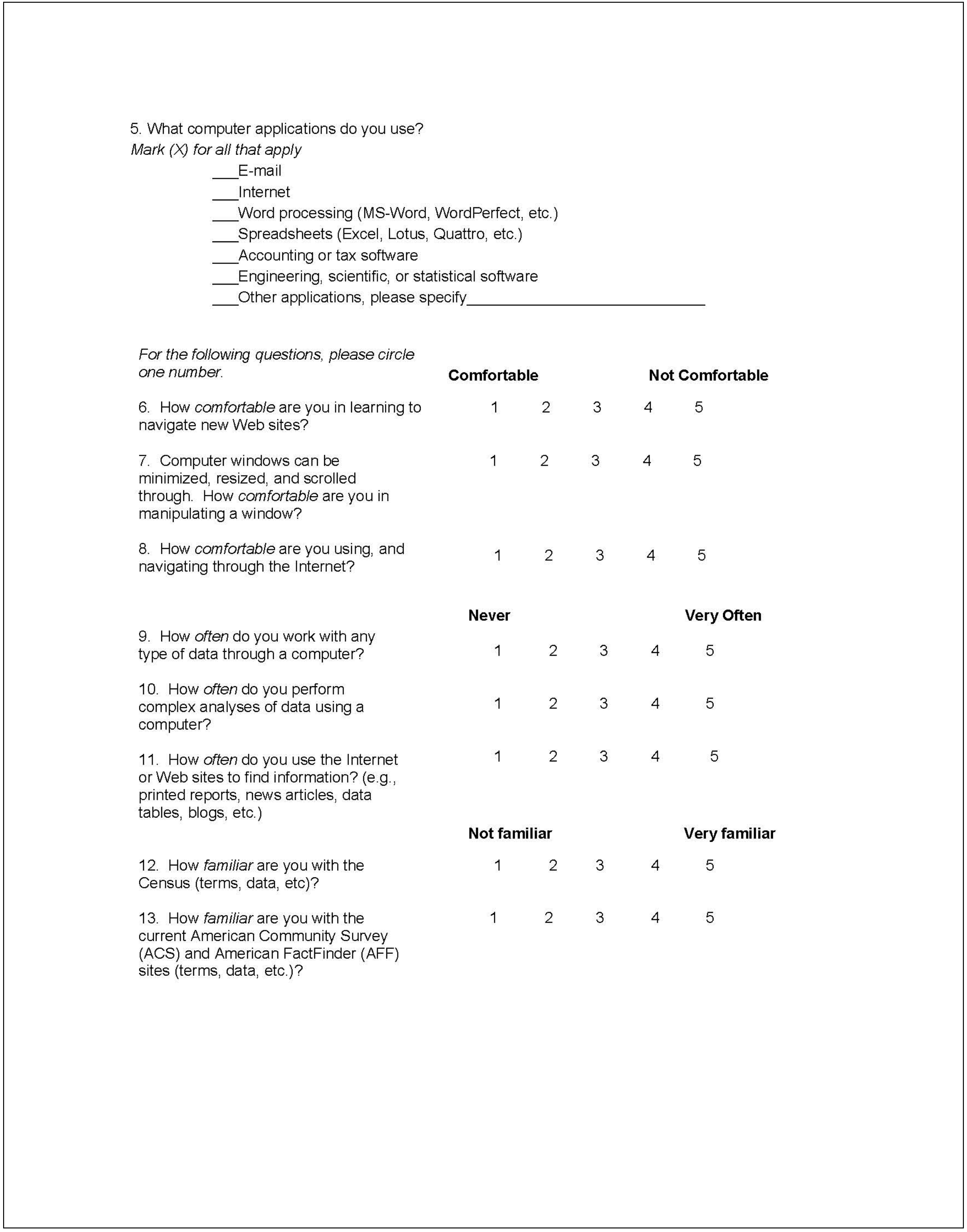
c. What operating system does your home computer run in?
___MAC OS
___Windows 95
___Windows 2000
___Windows XP
___Windows Vista
___Other ___________
___Don't know

3. On average, about how many hours do you spend on the Internet per day?
___0 hours
___1-3 hours
___4-6 hours
___7or more hours

4. Please rate your overall experience with the following:
Circle one number.

	No experience						**Very experienced**		
Computers	1	2	3	4	5	6	7	8	9
Internet	1	2	4	5	5	6	7	8	9

Figure 30. Background Survey Page 3



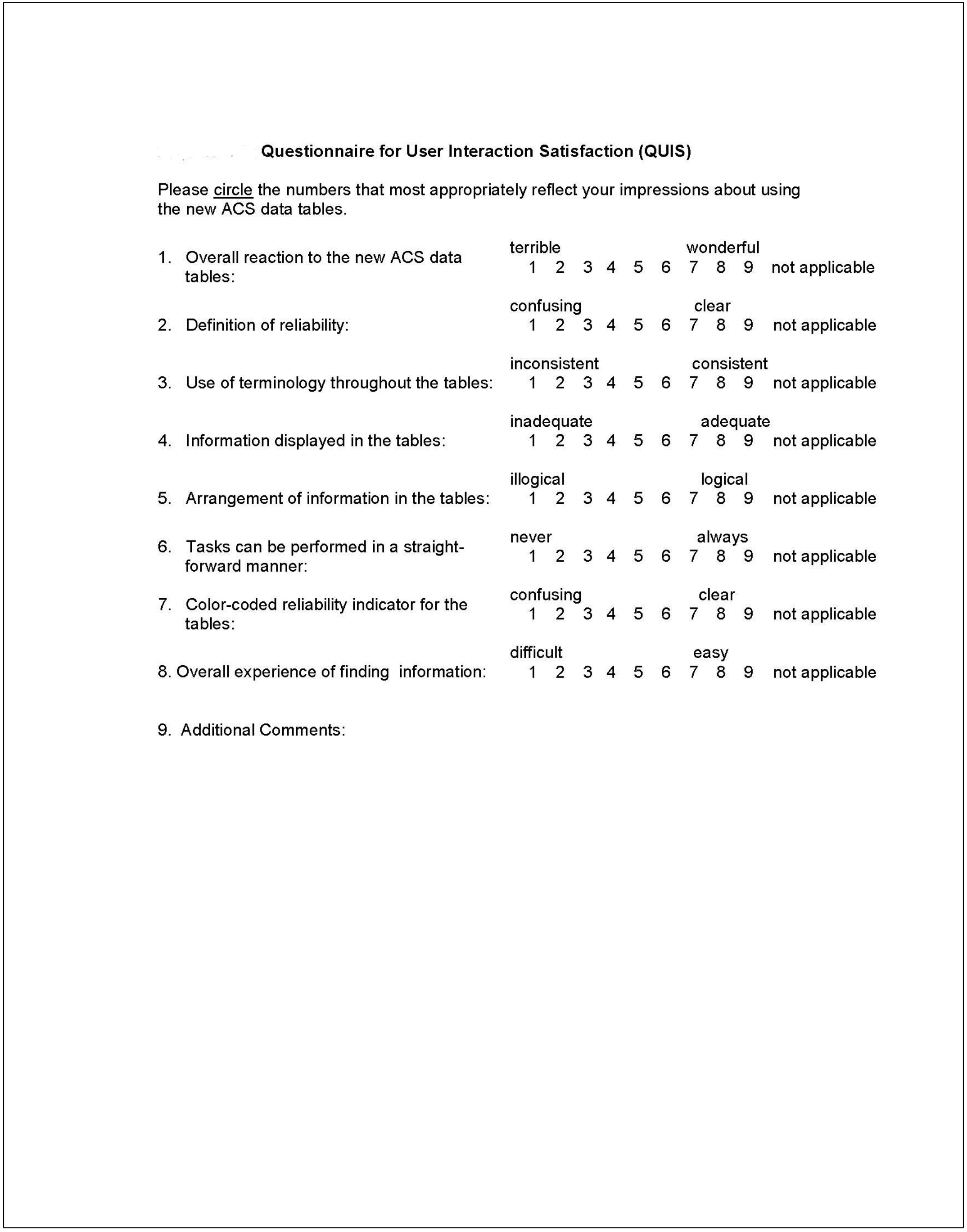
5. What computer applications do you use?
Mark (X) for all that apply

___E-mail
___Internet
___Word processing (MS-Word, WordPerfect, etc.)
___Spreadsheets (Excel, Lotus, Quattro, etc.)
___Accounting or tax software
___Engineering, scientific, or statistical software
___Other applications, please specify___________________________

For the following questions, please circle one number.	**Comfortable**			**Not Comfortable**	
6. How *comfortable* are you in learning to navigate new Web sites?	1	2	3	4	5
7. Computer windows can be minimized, resized, and scrolled through. How *comfortable* are you in manipulating a window?	1	2	3	4	5
8. How *comfortable* are you using, and navigating through the Internet?	1	2	3	4	5
	Never				**Very Often**
9. How *often* do you work with any type of data through a computer?	1	2	3	4	5
10. How *often* do you perform complex analyses of data using a computer?	1	2	3	4	5
11. How *often* do you use the Internet or Web sites to find information? (e.g., printed reports, news articles, data tables, blogs, etc.)	1	2	3	4	5
	Not familiar				**Very familiar**
12. How *familiar* are you with the Census (terms, data, etc)?	1	2	3	4	5
13. How *familiar* are you with the current American Community Survey (ACS) and American FactFinder (AFF) sites (terms, data, etc.)?	1	2	3	4	5

13 Appendix F: Questionnaire for User Interface Satisfaction (QUIS)

Figure 31. QUIS Instrument



Questionnaire for User Interaction Satisfaction (QUIS)

Please circle the numbers that most appropriately reflect your impressions about using the new ACS data tables.

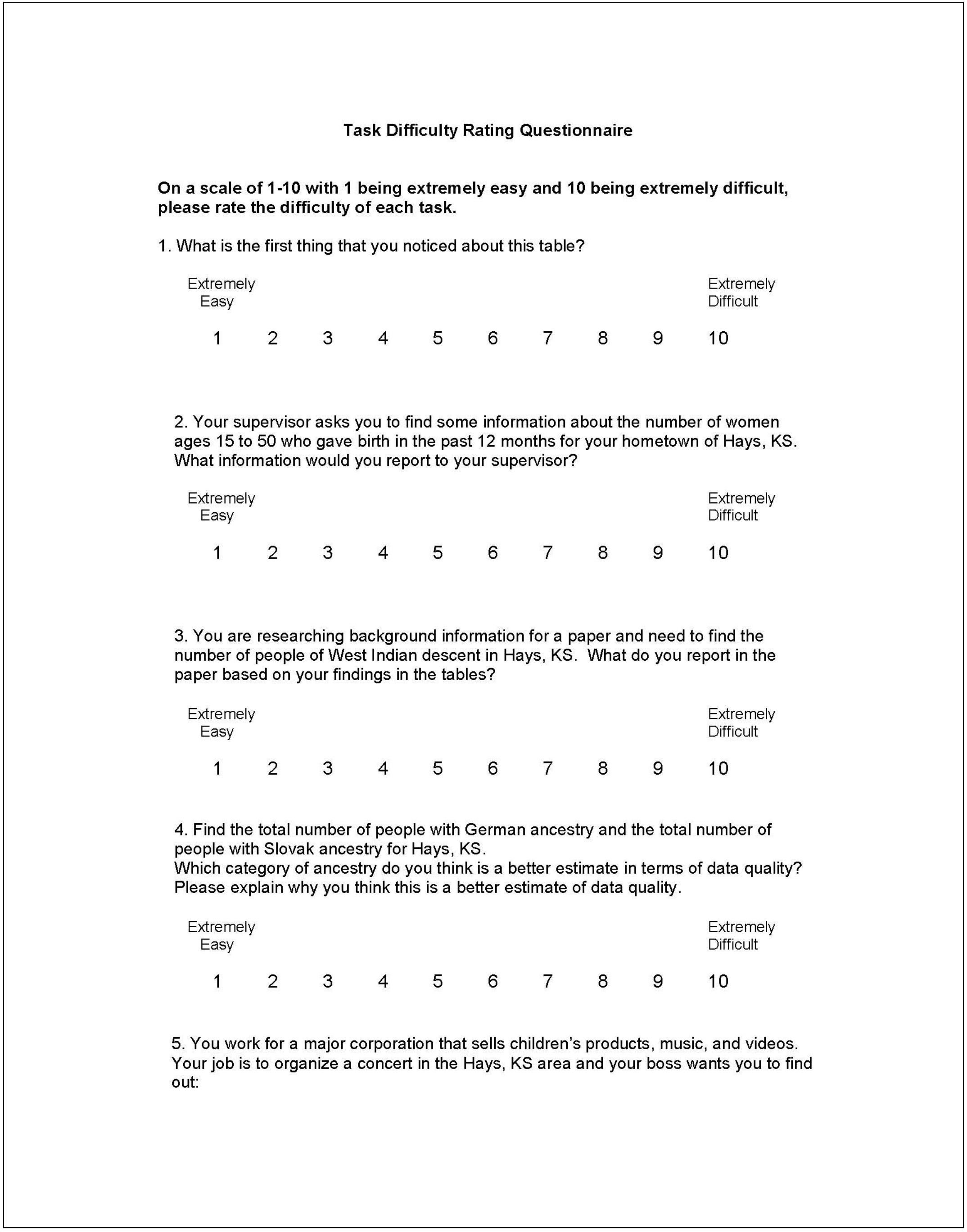
1. Overall reaction to the new ACS data tables:	terrible	wonderful	1 2 3 4 5 6 7 8 9	not applicable
2. Definition of reliability:	confusing	clear	1 2 3 4 5 6 7 8 9	not applicable
3. Use of terminology throughout the tables:	inconsistent	consistent	1 2 3 4 5 6 7 8 9	not applicable
4. Information displayed in the tables:	inadequate	adequate	1 2 3 4 5 6 7 8 9	not applicable
5. Arrangement of information in the tables:	illogical	logical	1 2 3 4 5 6 7 8 9	not applicable
6. Tasks can be performed in a straight-forward manner:	never	always	1 2 3 4 5 6 7 8 9	not applicable
7. Color-coded reliability indicator for the tables:	confusing	clear	1 2 3 4 5 6 7 8 9	not applicable
8. Overall experience of finding information:	difficult	easy	1 2 3 4 5 6 7 8 9	not applicable

9. Additional Comments:

14 Appendix G: Task Difficulty Rating Scale

This scale was given after the testing session itself was complete at the same time as the QUIS form above.

Figure 32. Stoplight Task Difficulty Scale Page 1



Task Difficulty Rating Questionnaire

On a scale of 1-10 with 1 being extremely easy and 10 being extremely difficult, please rate the difficulty of each task.

1. What is the first thing that you noticed about this table?

Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

2. Your supervisor asks you to find some information about the number of women ages 15 to 50 who gave birth in the past 12 months for your hometown of Hays, KS. What information would you report to your supervisor?

Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

3. You are researching background information for a paper and need to find the number of people of West Indian descent in Hays, KS. What do you report in the paper based on your findings in the tables?

Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

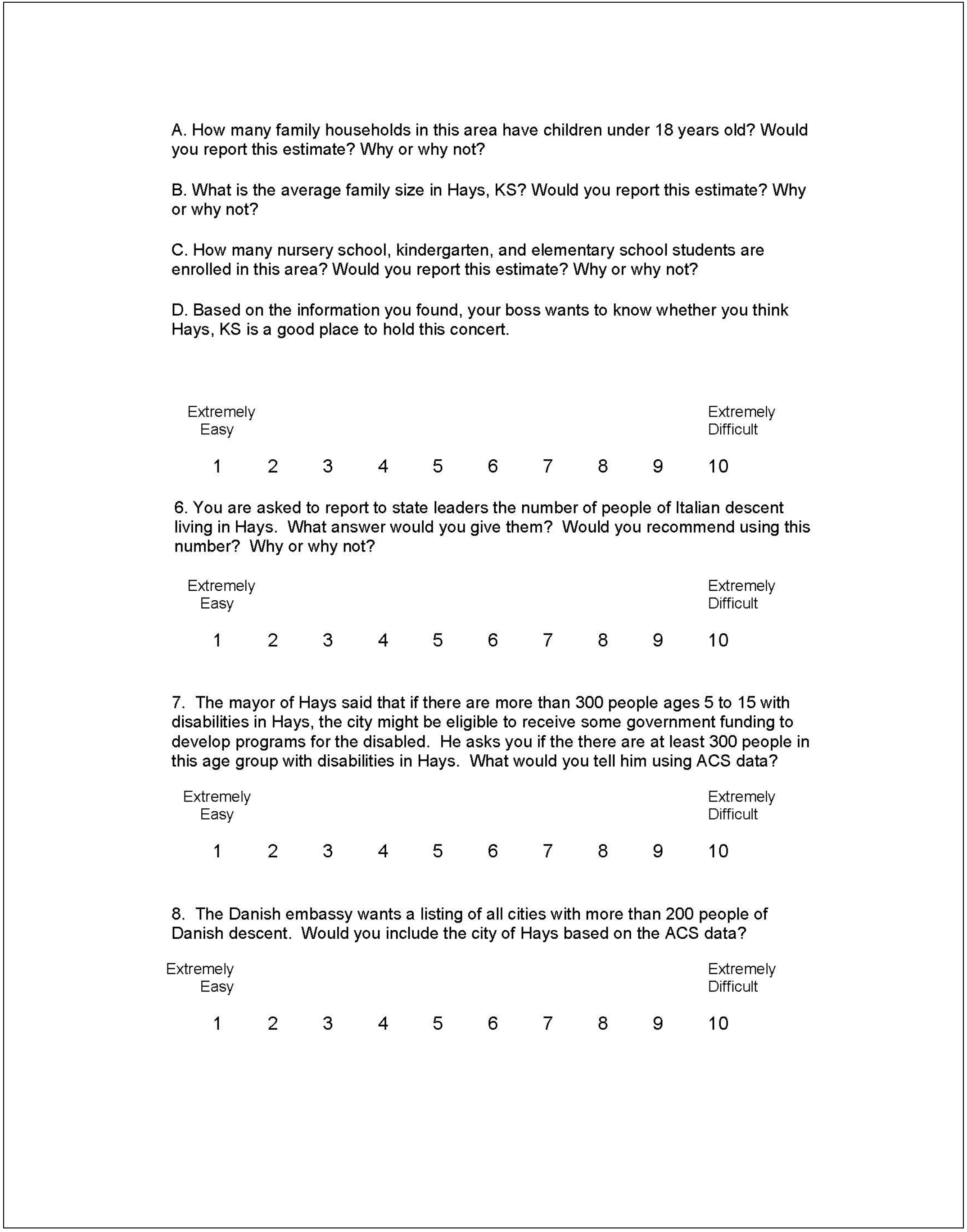
4. Find the total number of people with German ancestry and the total number of people with Slovak ancestry for Hays, KS.
Which category of ancestry do you think is a better estimate in terms of data quality? Please explain why you think this is a better estimate of data quality.

Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

5. You work for a major corporation that sells children's products, music, and videos. Your job is to organize a concert in the Hays, KS area and your boss wants you to find out:

Figure 33. Stoplight Task Difficulty Scale Page 2



A. How many family households in this area have children under 18 years old? Would you report this estimate? Why or why not?

B. What is the average family size in Hays, KS? Would you report this estimate? Why or why not?

C. How many nursery school, kindergarten, and elementary school students are enrolled in this area? Would you report this estimate? Why or why not?

D. Based on the information you found, your boss wants to know whether you think Hays, KS is a good place to hold this concert.

Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

6. You are asked to report to state leaders the number of people of Italian descent living in Hays. What answer would you give them? Would you recommend using this number? Why or why not?

Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

7. The mayor of Hays said that if there are more than 300 people ages 5 to 15 with disabilities in Hays, the city might be eligible to receive some government funding to develop programs for the disabled. He asks you if the there are at least 300 people in this age group with disabilities in Hays. What would you tell him using ACS data?

Extremely Easy — Extremely Difficult

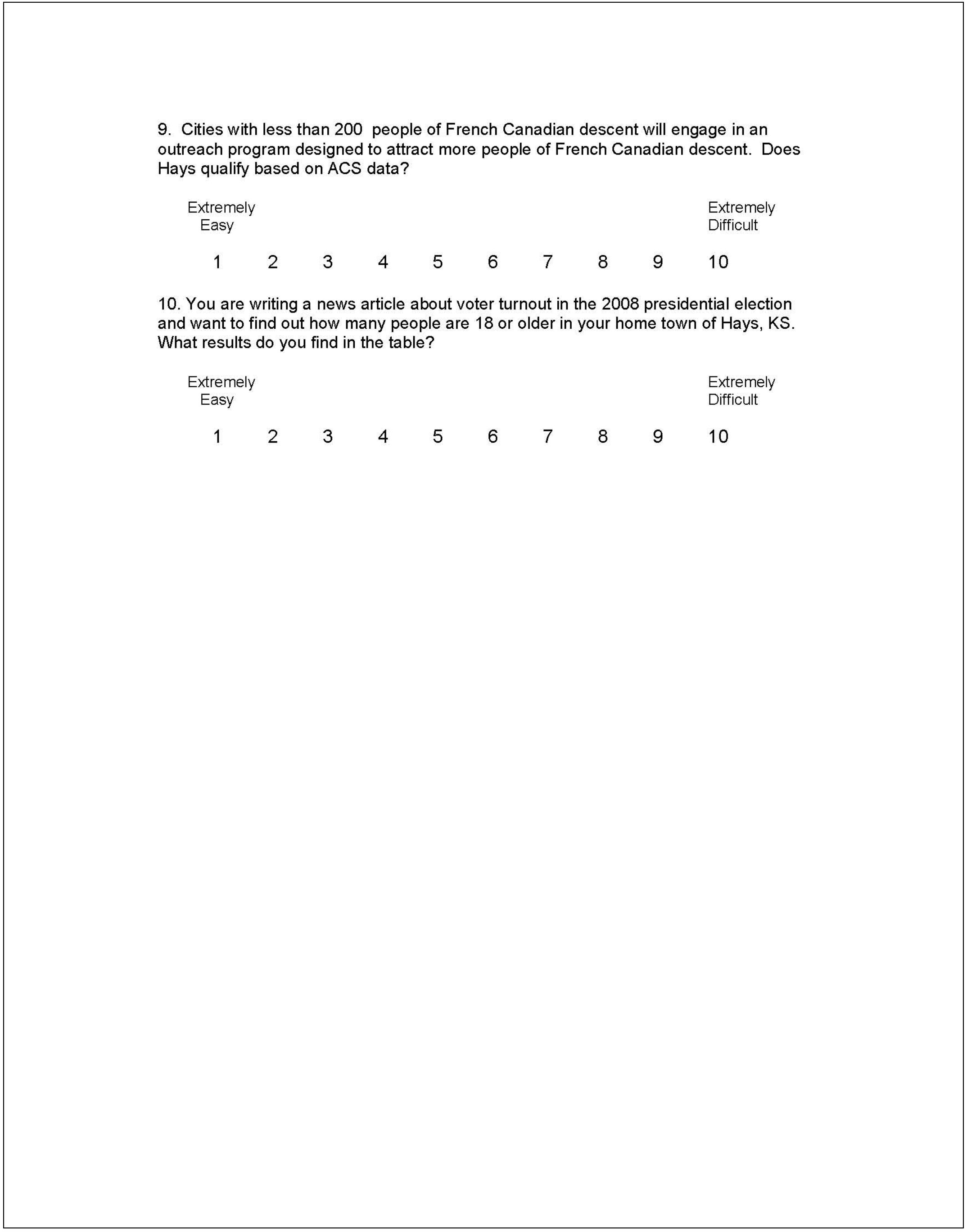
1 2 3 4 5 6 7 8 9 10

8. The Danish embassy wants a listing of all cities with more than 200 people of Danish descent. Would you include the city of Hays based on the ACS data?

Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

Figure 34. Stoplight Task Difficulty Scale Page 3



9. Cities with less than 200 people of French Canadian descent will engage in an outreach program designed to attract more people of French Canadian descent. Does Hays qualify based on ACS data?

Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

10. You are writing a news article about voter turnout in the 2008 presidential election and want to find out how many people are 18 or older in your home town of Hays, KS. What results do you find in the table?

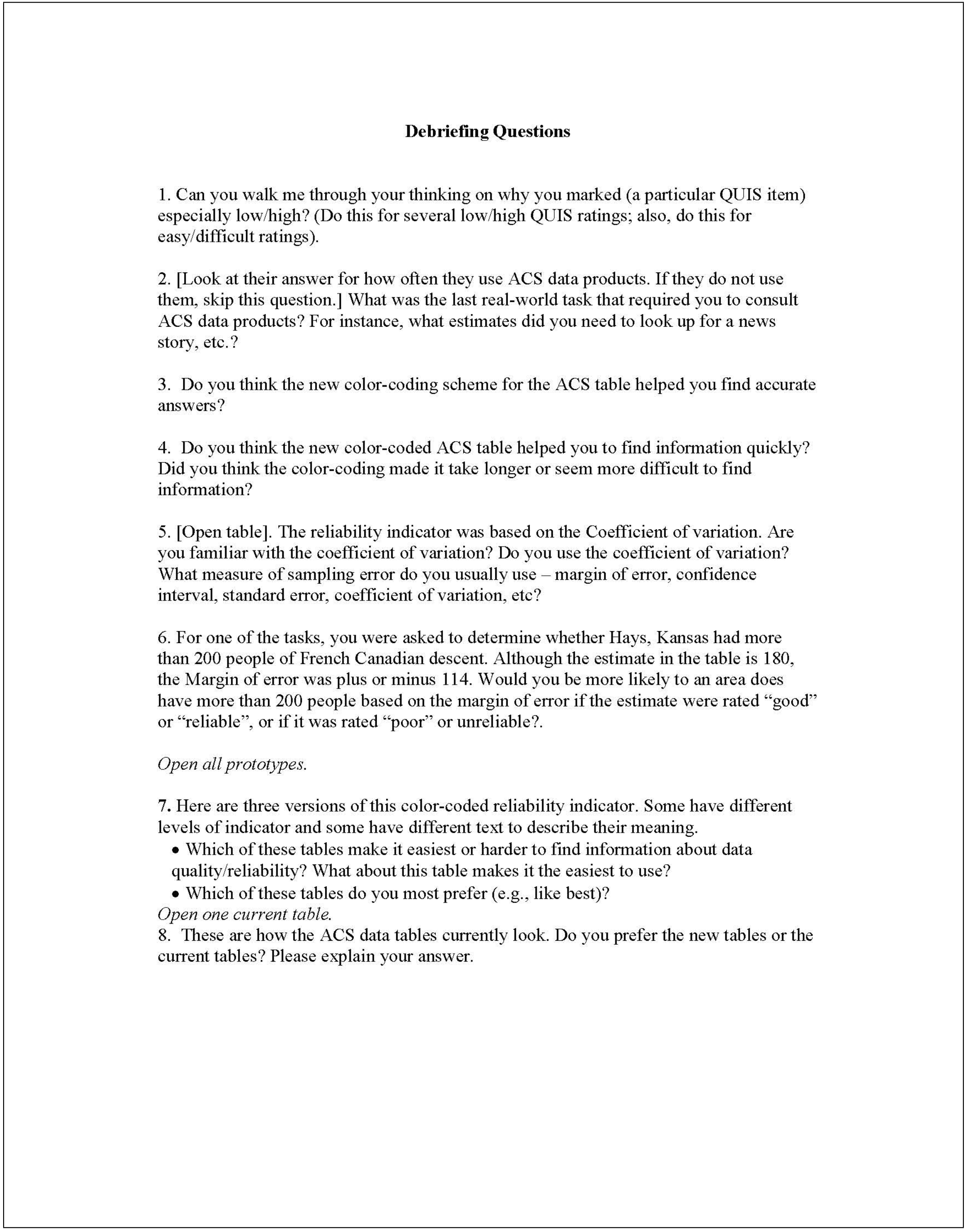
Extremely Easy — Extremely Difficult

1 2 3 4 5 6 7 8 9 10

15 Appendix H: Debriefing Interview Questions

These questions were asked after the testing session itself was complete in order to gain a more complete understanding of the user's experience with the ACS data table.

Figure 35. Debriefing Interview Questions



Debriefing Questions

1. Can you walk me through your thinking on why you marked (a particular QUIS item) especially low/high? (Do this for several low/high QUIS ratings; also, do this for easy/difficult ratings).

2. [Look at their answer for how often they use ACS data products. If they do not use them, skip this question.] What was the last real-world task that required you to consult ACS data products? For instance, what estimates did you need to look up for a news story, etc.?

3. Do you think the new color-coding scheme for the ACS table helped you find accurate answers?

4. Do you think the new color-coded ACS table helped you to find information quickly? Did you think the color-coding made it take longer or seem more difficult to find information?

5. [Open table]. The reliability indicator was based on the Coefficient of variation. Are you familiar with the coefficient of variation? Do you use the coefficient of variation? What measure of sampling error do you usually use – margin of error, confidence interval, standard error, coefficient of variation, etc?

6. For one of the tasks, you were asked to determine whether Hays, Kansas had more than 200 people of French Canadian descent. Although the estimate in the table is 180, the Margin of error was plus or minus 114. Would you be more likely to an area does have more than 200 people based on the margin of error if the estimate were rated "good" or "reliable", or if it was rated "poor" or unreliable?.

Open all prototypes.

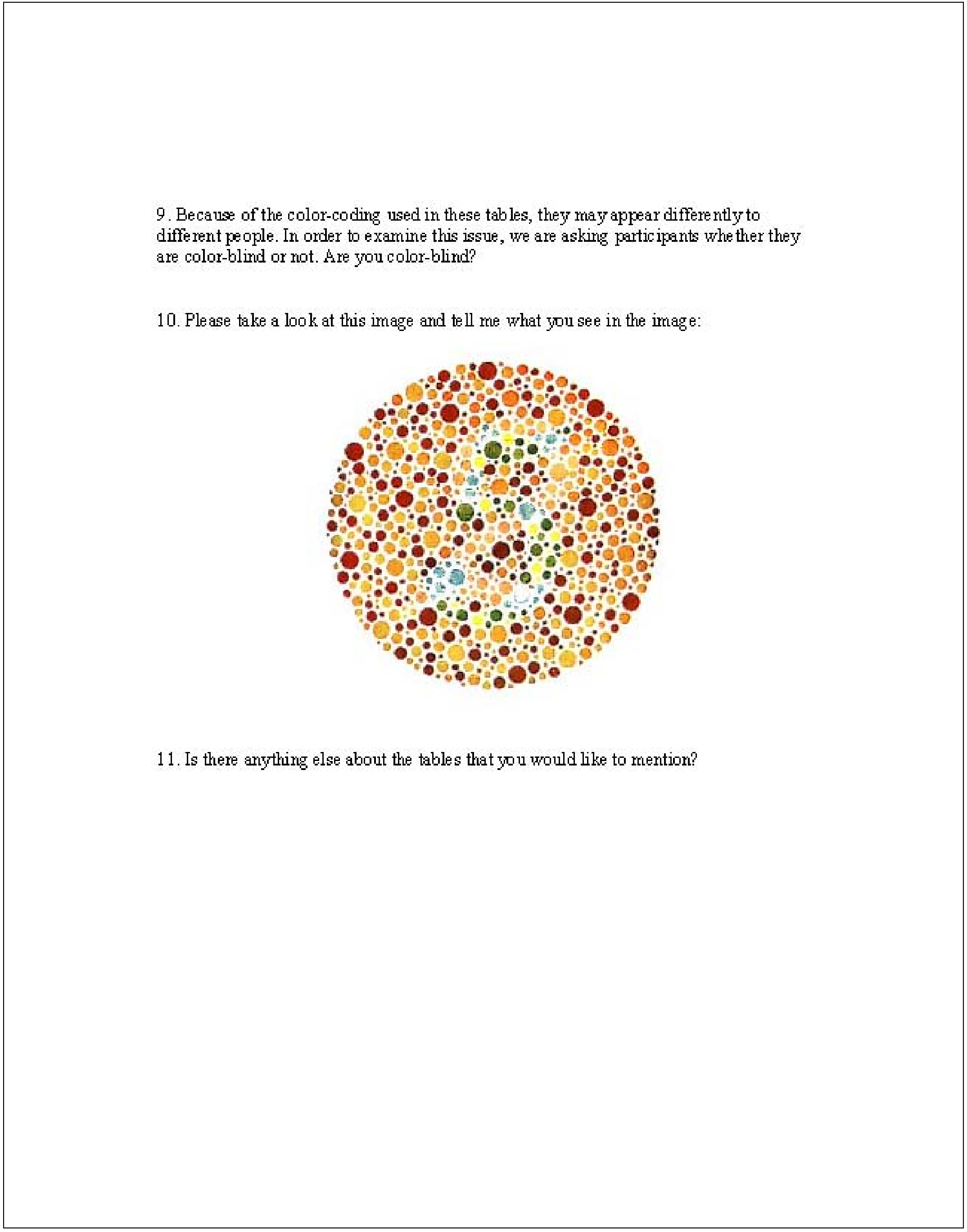
7. Here are three versions of this color-coded reliability indicator. Some have different levels of indicator and some have different text to describe their meaning.

- Which of these tables make it easiest or harder to find information about data quality/reliability? What about this table makes it the easiest to use?
- Which of these tables do you most prefer (e.g., like best)?

Open one current table.

8. These are how the ACS data tables currently look. Do you prefer the new tables or the current tables? Please explain your answer.

Figure 36. Debriefing Interview Questions



9. Because of the color-coding used in these tables, they may appear differently to different people. In order to examine this issue, we are asking participants whether they are color-blind or not. Are you color-blind?

10. Please take a look at this image and tell me what you see in the image:

11. Is there anything else about the tables that you would like to mention?