UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

August 30, 2017

2017 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES # ACS17-RER-06

MEMORANDUM FOR	Victoria Velkoff Chief, American Community Survey
From:	David Waddington Chief, Social, Economic, and Housing Statistics Division (SEHSD)
Prepared by:	Rose Kreider Social, Economic, and Housing Statistics Division (SEHSD)
Subject:	2016 American Community Survey Content Test Evaluation Report: Relationship

Attached is the final report for the 2016 American Community Survey (ACS) Content Test for Relationship. This report describes the results of the test for a revised version of the Relationship question.

If you have any questions about this report, please contact Rose Kreider at 301-763-6059 or Yeris Mayol Garcia at 301-763-6844.

Attachment

cc:
Jennifer Ortman (ACSO)
Megan Rabe (ACSO)
Jennifer Reichert (ACSO)
Robert Sawyer (ACSO)
Nancy Bates (ADRM)
Michael Bentley (DSSD)
Patrick Cantwell (DSSD)
Elizabeth Poehler (DSSD)
Anthony Tersine (DSSD)
Colleen Keating (POP)
Nicole Scanniello (SEHSD)

Intentionally Blank

American Community Survey Research and Evaluation Program
August 30, 2017

2016 American Community Survey Content Test Evaluation Report: Relationship

FINAL REPORT

Rose M. Kreider and Yerís Mayol-García
Social, Economic, and Housing Statistics Division

R. Chase Sawyer
American Community Survey Office

United States™ Census Bureau

Intentionally Blank

TABLE OF CONTENTS

List of Tables

List of Figures

EXECUTIVE SUMMARY

Overview

From February to June of 2016, the U.S. Census Bureau conducted the 2016 American Community Survey (ACS) Content Test, a field test of new and revised content. The primary objective was to test whether changes to question wording, response categories, and definitions of underlying constructs improve the quality of data collected. Both new and revised versions of existing questions were tested to determine if they could provide data of sufficient quality compared to a control version as measured by a series of metrics including item missing data rates, response distributions, comparisons with benchmarks, and response error. The results of this test will be used to help determine the future ACS content and to assess the expected data quality of revised questions and new questions added to the ACS.

The 2016 ACS Content Test consisted of a nationally representative sample of 70,000 residential addresses in the United States, independent of the production ACS sample. The sample universe did not include group quarters, nor did it include housing units in Alaska, Hawaii, or Puerto Rico. The test was a split-panel experiment with one-half of the addresses assigned to the control treatment and the other half assigned to the test treatment. As in production ACS, the data collection consisted of three main data collection operations: 1) a six-week mailout period, during which the majority of self-response via internet and mailback were received; 2) a one-month Computer-Assisted Telephone Interview period for nonresponse follow-up; and 3) a one-month Computer-Assisted Personal Interview period for a sample of the remaining nonresponse. For housing units that completed the original 2016 ACS Content Test interview, a Content Follow-Up telephone reinterview was conducted to measure response error.

Relationship

For several years, the Census Bureau has been testing a revised relationship question to improve the estimates of couple households. The 1990 Census was the first to include *unmarried partner* as a response category to the relationship to householder question. The 2000 and 2010 Censuses built on this work, changing the processing of the relationship data to more accurately represent same-sex couples. Although the 2010 Census did not include a separate category for married or unmarried same-sex couples, estimates were developed for same-sex and opposite-sex married and unmarried couples by combining couples' responses on relationship and sex. The Census Bureau discovered an error in the 2010 Census data that resulted, in part, from opposite-sex couples mismarking their sex (O'Connell and Feliz, 2011). This error has the potential to affect the estimates of same-sex married couple households. The Census Bureau released a set of modified state-level same-sex household estimates from the 2010 Census due to this error and also began new research efforts to improve the relationship question. The revised relationship question tested in the 2016 ACS Content Test, one that features explicit same-sex and opposite-sex spouse and same-sex and opposite-sex partner response categories, has the potential to improve couple household estimates, especially same-sex couple household estimates.

Two versions of the relationship question were featured in the 2016 ACS Content Test: the 2016 ACS production relationship question (control) and a new version where the two couple relationship categories were expanded to distinguish between opposite-sex and same-sex relationships (test). Images of both versions are shown later in this report.

Research Questions and Results

1. *Are item missing data rates lower in the test treatment than in the control treatment?* Except for the internet mode, no significant differences were found in the item missing data rates between the test and control treatments. The control treatment rate in the internet mode was significantly lower than the test treatment rate; however, both rates were less than one-half of one percent.

2. *Does the distribution of people reported as a spouse or partner of the householder differ between control and test treatments?* The test and control versions show no significant differences between the proportions of responses for spouses or partners of the householder. The only significant difference found in the distributions of response categories was the *missing* category in the internet mode. The category of *missing* responses does not affect the resulting estimates from the relationship question.

3. *Do the measures of response reliability differ between the test and control treatment?* Overall, there was no significant difference between the reliability of test and control treatments, except for one category. The analysis of the *unmarried partner* response showed a moderate inconsistency (between 20-50 percent) with the test question and low inconsistency (less than 20 percent) with the control question.

4. *Does the inclusion of an automated relationship/sex consistency check improve data quality?* Data quality is improved through the automated consistency check. When prompted, many couples changed their sex or relationship responses, which resulted in lower inconsistency rates.

Conclusions

The test relationship question evaluated in the 2016 ACS Content Test did not show overall differences in response rates nor in the distribution of those who reported being a spouse or partner compared to respondents receiving the old version of the question. There was a significant difference in the item missing data rates in the internet response mode, but the magnitude of these differences is small, only 0.2 percentage points. The distribution across categories in the 2016 ACS Content Test distribution conform to the expectations set by the 2015 National Content Test (NCT). These results suggest that the collection of detailed relationship data will not adversely affect data quality and will enable improvements in editing procedures. We also evaluated the function of an automated relationship/sex consistency check in electronic instruments, finding that it helps reduce inconsistent responses, especially among opposite-sex couples, who are the largest source of the error affecting the estimate of same-sex married couples. While final recommendations for the relationship question will be made primarily based

on testing done in the 2015 NCT and other Census Tests, the 2016 ACS Content Test provides further evidence that the revised relationship question functions well, and that an automated relationship/sex consistency check in electronic instruments improves data quality.

Intentionally Blank

1. BACKGROUND

From February to June of 2016, the Census Bureau conducted the 2016 American Community Survey (ACS) Content Test, a field test of new and revised content. The primary objective was to test whether changes to question wording, response categories, and definitions of underlying constructs improve the quality of data collected. Both revised versions of existing questions and new questions were tested to determine if they could provide data of sufficient quality compared to a control version as measured by a series of metrics including item missing data rates, response distributions, comparisons with benchmarks, and response error. The results of this test will be used to help determine the future ACS content and to assess the expected data quality of revised questions and new questions added to the ACS.

The 2016 ACS Content Test included the following topics:

- Relationship
- Race and Hispanic Origin
- Telephone Service
- Computer and Internet Use
- Health Insurance Coverage
- Health Insurance Premium and Subsidy (new questions)
- Journey to Work: Commute Mode
- Journey to Work: Time of Departure for Work
- Number of Weeks Worked
- Class of Worker
- Industry and Occupation
- Retirement, Survivor, and Disability Income

This report discusses the Relationship topic.

1.1. Justification for Inclusion of Relationship in the Content Test

The Census Bureau collects the relationship of each member of the household to the householder (the person who owns or rents the home) in the decennial census and household surveys. The relationship question has been asked on the decennial census since 1880. In 1990, the category *unmarried partner* was added to the relationship item in the decennial census to measure the growing complexity of American households and the increasing tendency for couples to live together before getting married. The *unmarried partner* category was also added to the Current Population Survey (CPS) in 1995, the Survey of Income and Program Participation (SIPP) in 1996, and has been on the ACS since it went into full implementation in 2005.

The increasing social acceptance and legal recognition of same-sex marriages have led to a need for better federal data on same-sex couples. In 2010, as part of the interagency group on Measuring Relationships in Federal Household Surveys, led by the U.S. Office of Management and Budget (OMB), the Census Bureau conducted focus groups and cognitive interviews to see how respondents viewed the relationship question categories. Key findings from that research included: 1) respondents desired new categories to reflect legal unions for same-sex couples (e.g., civil unions and domestic partnerships); 2) respondents desired to move the *unmarried*

partner category next to *spouse* in the list; and 3) while some persons interpreted the term *partner* to apply more to same-sex intimate relationships, opposite-sex unmarried couples were generally comfortable selecting *unmarried partner* as their relationship category (Interagency Working Group on Measuring Relationships in Federal Household Surveys, 2014).

To date, the decennial census and the ACS have identified same-sex couples using the relationship question in conjunction with the sex question. In 1990, couples who reported they were of the same-sex and married were edited and shown as an opposite-sex married couple. In 2000 and 2010, same-sex married couples were edited to be shown as same-sex unmarried couples (Simmons & O'Connell, 2003; O'Connell & Gooding, 2007; O'Connell & Feliz, 2011). Data from Census 2000 reported all same-sex couples as unmarried couples, as no states allowed for same-sex marriages at that time. The Census Bureau has used ACS data to release yearly estimates of same-sex married couple households back to 2005. The 2010 Census marked the first published reports of those who identified themselves as same-sex married couples using decennial data. Beginning in the 2013 ACS, those who reported being same-sex married couples were shown as such in the data.

The proposed change to the relationship question makes the question consistent with questions used by other countries that collect data on same-sex married couples, such as the United Kingdom, Canada, New Zealand, and France. The explicitly listed categories for each couple type make it easier to identify and edit households in which respondents mismark the sex of one of the members of the couple. This kind of mistake, when made by a very small proportion of opposite-sex married couples, who constitute a very large group, can have a large impact on the estimates of a relatively small group like same-sex married couples (O'Connell & Feliz, 2011).

The newly revised relationship question (see Section 1.3) has been tested in the 2013 ACS Questionnaire Design Test (ACS-QDT), as well as the 2013 American Housing Survey. The SIPP implemented the new question in 2014. The new question is also being tested in the decennial program. It was fielded in the 2014 Census Test and the spring 2015 Census Test (Seem & Coombs, 2017) and the 2015 National Content Test (NCT), which was the largest test of the new question with a sample of approximately 1.2 million households. The question was fielded in the 2016 Census Test, was included in the 2017 Census Test, and is planned for inclusion in the 2018 End-to-End Census Test.

1.2. Question Development

Initial versions of the new and revised questions were proposed by federal agencies participating in the OMB Interagency Committee for the ACS. The initial proposals contained a justification for each change and described previous testing of the question wording, the expected impact of revisions to the time series and the single-year as well as five-year estimates, and the estimated net impact on respondent burden for the proposed revision.[1] For proposed new questions, the justification also described the need for the new data, whether federal law or regulation required the data for small areas or small population groups, if other data sources were currently available

[1] The ACS produces both single and five-year estimates annually. Single year estimates are produced for geographies with populations of 65,000 or more and five-year estimates are produced for all areas down to the block-group level, with no population restriction.

to provide the information (and why any alternate sources were insufficient), how policy needs or emerging data needs would be addressed through the new question, an explanation of why the data were needed with the geographic precision and frequency provided by the ACS, and whether other testing or production surveys had evaluated the use of the proposed questions.

The Census Bureau and the OMB, as well as the Interagency Council on Statistical Policy Subcommittee, reviewed these proposals for the ACS. The OMB determined which proposals moved forward into cognitive testing. After OMB approval of the proposals, topical subcommittees were formed from the OMB Interagency Committee for the ACS, which included all interested federal agencies that use the data from the impacted questions. These subcommittees further refined the specific proposed wording that was cognitively tested.

The proposed changes identified through cognitive testing for each question topic were reviewed by the Census Bureau, the corresponding topical subcommittee, and the Interagency Council on Statistical Policy Subcommittee for the ACS. The OMB then provided final overall approval of the proposed wording for field testing.[2]

1.3. Question Content

The relationship question was revised in order to improve the measurement of same-sex couples. For this test, the existing *husband or wife* and *unmarried partner* response categories were each split into two versions, *same-sex …* or *opposite-sex … .* Additionally, the two *unmarried partner* categories were moved from near the end of the list of response options to near the beginning, immediately after the *husband/wife/spouse* options. Control and test versions of each question are shown in Figures 1 and 2, respectively, as they appeared on the paper questionnaire.[3] Automated versions of the questionnaire had the same content formatted as appropriate for each mode. There are no notable differences in the presentation of this question between modes aside from branching in Computer-Assisted Telephone Interview (CATI) mode, which is described further in section 2.2.

[2] A cohabitation question and domestic partnership question were included in cognitive testing but ultimately we decided not to move forward with field testing these questions.
[3] Regarding interview mode, paper and mail are used interchangeably in this report.

Figure 1: Control Version of the Relationship Question, 2016 ACS Content Test

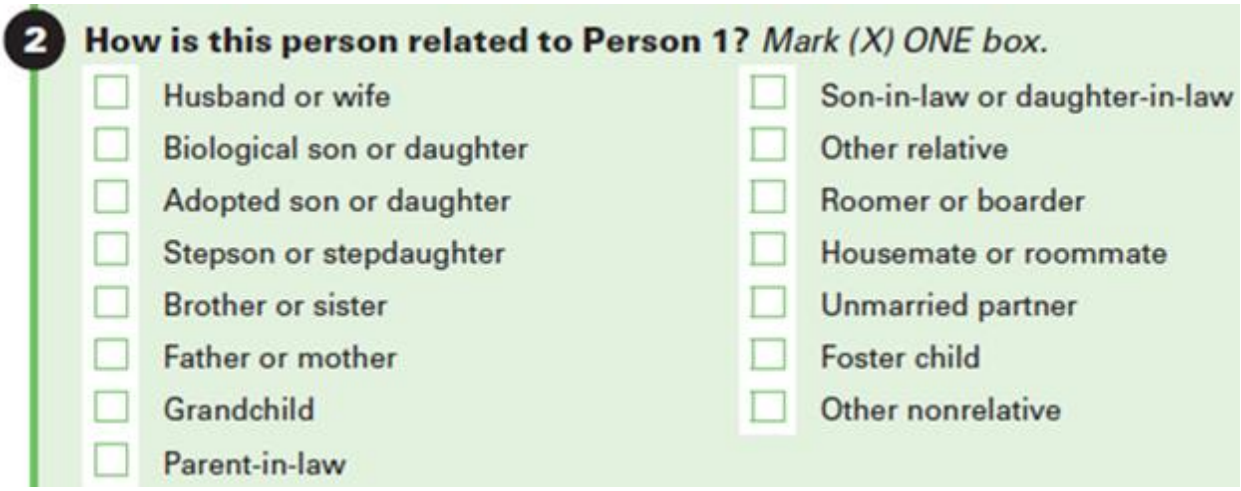
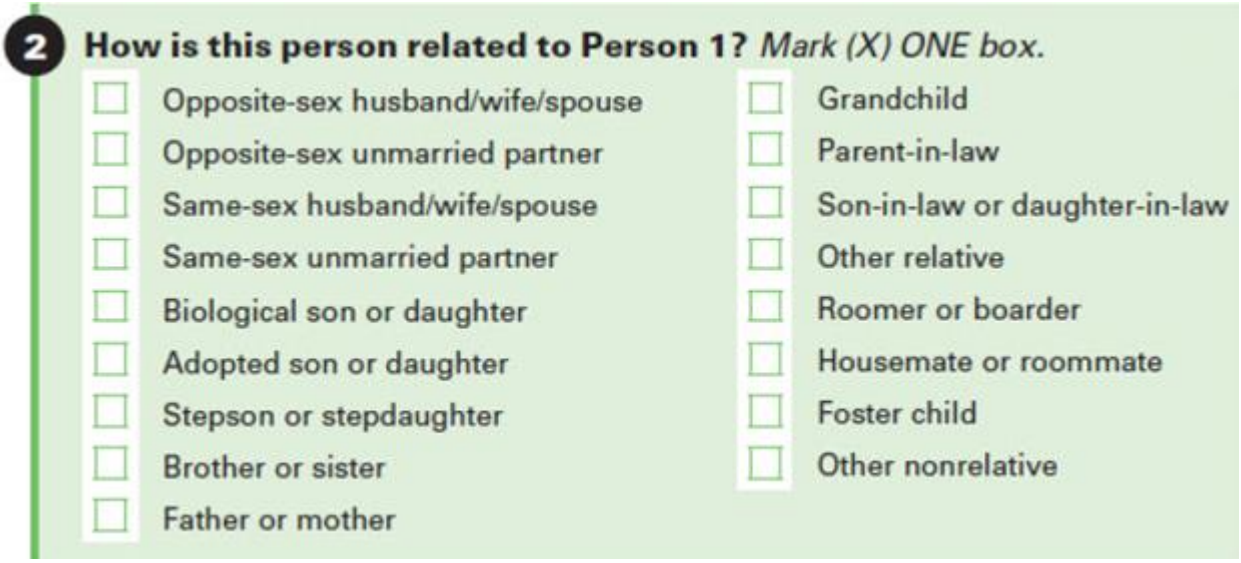


Figure 2: Test Version of the Relationship Question, 2016 ACS Content Test



1.4. Research Questions

The following research questions were formulated to guide the analyses of the relationship question. The analyses assessed how the test version of the question performed compared to the control version in the following ways: how often the respondents answered the question, the consistency and accuracy of the responses, and how the responses affected the resulting estimates. Also, the analyses assess the performance of the automated relationship/sex consistency check. More detailed questions that fit under the broader research questions shown in this section are also addressed in the current report.

1. *Are item missing data rates lower in the test treatment than in the control treatment?*

2. *Does the distribution of people reporting as a spouse or partner of the householder differ between control and test treatments?*

3. *Do the measures of response reliability (gross difference rate, index of inconsistency) differ between the test and control treatment?*

4. *Does the inclusion of an automated relationship/sex consistency check improve data quality?*

The Research and Evaluation Analysis Plan for the 2016 ACS Content Test was developed before the 2015 NCT was conducted. As a result, the original research questions for the relationship topic (provided below) differed from the research questions for the 2015 NCT. Because the 2015 NCT is the largest test fielded in the decennial program leading up to the 2020 Census, the 2016 ACS Content Test research questions were revised in order to conduct the same analysis on the 2016 ACS Content Test data to be able to make meaningful comparisons between the results of the two tests. Questions 1 and 2 above are the same as those used in the final report for 2015 NCT data.[4] Question 3 above was included in this report because of the inclusion of reinterviews in the 2016 ACS Content Test. Question 4 was added in order to evaluate the function of the automated relationship/sex consistency check. Below are the original research questions for the relationship topic. Notes following each question indicate which of the revised questions it corresponds to and in which section of this report the corresponding analysis is found.

<u>Original Research Questions:</u>

1. *How does the distribution of coupled households from each treatment (Control/Test) compare with the results from the 2015 National Content Test (NCT)?* This question is addressed in research question 2 and Section 5.4.

2. *Is the item missing data rate lower for the test treatment than for the control treatment?* This question is addressed in research question 1 and Section 5.2.

3. *Is the percentage of coupled households (married and unmarried) the same with the test question as with the Control?* This question was restructured and fits into the discussion of research question 2 and Section 5.3.

4. *Is the percent of coupled households by type, whether opposite-sex or same-sex (married and unmarried) higher in the test responses than in control?* This question is addressed in research question 2 and Section 5.4.

5. *Are the measures of response reliability (gross difference rate, index of inconsistency) better for the test treatment than for the control treatment?* This question is the same as research question 3 and is answered in Section 5.5.

[4] Note that these questions evaluate essentially the same basic issues as the questions we had in the Research and Evaluation Analysis Plan, but they are worded differently, to coincide with what we did for the 2015 NCT.

2. METHODOLOGY

2.1. Sample Design

The 2016 ACS Content Test consisted of a nationally representative sample of 70,000 residential addresses in the United States, independent of the production ACS sample. The 2016 ACS Content Test sample universe did not include group quarters (GQ), nor did it include housing units in Alaska, Hawaii, or Puerto Rico.[5] The sample design for the 2016 ACS Content Test was largely based on the ACS production sample design with some modifications to better meet the test objectives.[6] The modifications included adding an additional level of stratification by stratifying addresses into high and low self-response areas, oversampling addresses from low self-response areas to ensure equal response from both strata, and sampling units as pairs.[7] The high and low self-response strata were defined based on ACS self-response rates at the tract level. Sampled pairs were formed by first systematically sampling an address within the defined sampling stratum and then pairing that address with the address listed next in the geographically sorted list. Note that the pair was likely not neighboring addresses. One member of the pair was randomly assigned to receive the control version of the question and the other member was assigned to receive the test version of the question, thus resulting in a sample of 35,000 control cases and 35,000 test cases.

As in the production ACS, if efforts to obtain a response by mail or telephone were unsuccessful, attempts were made to interview in person a sample of the remaining nonresponding addresses (see Section 2.2 Data Collection for more details). Addresses were sampled at a rate of 1-in-3, with some exceptions that were sampled at a higher rate.[8] For the 2016 ACS Content Test, the development of workload estimates for the CATI and Computer Assisted Personal Interviews (CAPI) did not take into account the oversampling of low response areas. This oversampling resulted in a higher than expected workload for CATI and CAPI and therefore required more budget than was allocated. To address this issue, the CAPI sampling rate for the 2016 ACS Content Test was adjusted to meet the budget constraint.

2.2. Data Collection

The field test occurred in parallel with the data collection activities for the March 2016 ACS production panel, using the same basic data collection protocol as production ACS with a few differences as noted below. The data collection protocol consisted of three main data collection operations: 1) a six-week mailout period, during which the majority of internet and mailback responses were received; 2) a one-month CATI period for nonresponse follow-up; and 3) a one-

[5] Alaska and Hawaii were excluded for cost reasons. GQs and Puerto Rico were excluded because the sample sizes required to produce reliable estimates would be overly large and burdensome, as well as costly.

[6] The ACS production sample design is described in Chapter 4 of the ACS Design and Methodology report (U.S. Census Bureau, 2014).

[7] Tracts with the highest response rate based on data from the 2013 and 2014 ACS were assigned to the high response stratum in such a way that 75 percent of the housing units in the population (based on 2010 Census estimates) were in the high response areas; all other tracts were designated in the low response strata. Self-response rates were used as a proxy for overall cooperation. Oversampling in low response areas helps to mitigate larger variances due to CAPI subsampling. This stratification at the tract level was successfully used in previous ACS Content Tests, as well as the ACS Voluntary Test in 2003.

[8] The ACS production sample design for CAPI follow-up is described in Chapter 4, Section 4.4 of the ACS Design and Methodology report (U.S. Census Bureau, 2014).

month CAPI period for a sample of the remaining nonresponse. Internet and mailback responses were accepted until three days after the end of the CAPI month.

As indicated earlier, housing units included in the 2016 ACS Content Test sample were randomly assigned to a control or test version of the questions. CATI interviewers were not assigned specific cases; rather, they worked the next available case to be called and therefore conducted interviews for both control and test cases. CAPI interviewers were assigned 2016 ACS Content Test cases based on their geographic proximity to the cases and therefore could also conduct both control and test cases.

The 2016 ACS Content Test's data collection protocol differed from the production ACS in a few significant ways. The 2016 ACS Content Test analysis did not include data collected via the Telephone Questionnaire Assistance (TQA) program since those who responded via TQA used the ACS production TQA instrument. The 2016 ACS Content Test excluded the telephone Failed Edit Follow-Up (FEFU) operation.[9] Furthermore, the Content Test had an additional telephone reinterview operation used to measure response reliability. We refer to this telephone reinterview component as the Content Follow-Up, or CFU. The CFU is described in more detail in Section 2.3.

ACS production provides Spanish-language versions of the internet, CATI, and CAPI instruments, and callers to the TQA number can request to respond in Spanish, Russian, Vietnamese, Korean, or Chinese. The 2016 ACS Content Test had Spanish-language automated instruments; however, there were no paper versions of the 2016 ACS Content Test questionnaires in Spanish.[10] Any case in the 2016 ACS Content Test sample that completed a Spanish-language internet, CATI, or CAPI response was included in analysis. However, if a case sampled for the 2016 ACS Content Test called TQA to complete an interview in Spanish or any other language, the production interview was conducted and the response was therefore excluded from the 2016 ACS Content Test analysis. This was due to the low volume of non-English language cases and the operational complexity of translating and implementing several language instruments for the 2016 ACS Content Test. CFU interviews for the 2016 ACS Content Test were conducted in either Spanish or English. The practical need to limit the language response options for 2016 ACS Content Test respondents is a limitation to the research, as some respondents self-selected out of the test.

Additionally, as is already implemented in the CATI for current ACS production, the relationship question branched for *son/daughter* responses to allow respondents to specify how the child is related to them. In other words, only in CATI, the option *son or daughter* was in the initial answer category list, and then the respondent was prompted to identify the child as their *biological child, stepchild,* or *adopted child*. *Foster child* was also listed as an answer category following the branching, although it is listed separately in the initial answer categories as well. In

[9] In ACS production, paper questionnaires with an indication that there are more than five people in the household or questions about the number of people in the household, and self-response returns that are identified as being vacant or a business or lacking minimal data are included in FEFU. FEFU interviewers call these households to obtain any information the respondent did not provide.

[10] In the 2014 ACS, respondents requested 1,238 Spanish paper questionnaires, of which 769 were mailed back. From that information, we projected that fewer than 25 Spanish questionnaires would be requested in the 2016 ACS Content Test.

mail, internet, and CAPI, *biological son/daughter*, *stepson/stepdaughter*, and *adopted son/adopted daughter* are listed individually as answer categories.

2.3. Content Follow-Up

For housing units that completed the original interview, a CFU telephone reinterview was also conducted to measure response error.[11] A comparison of the original interview responses and the CFU reinterview responses was used to answer research questions about response error and response reliability.

A CFU reinterview was attempted with every household that completed an original interview for which there was a telephone number. A reinterview was conducted no sooner than two weeks (14 calendar days) after the original interview. Once the case was sent to CFU, it was to be completed within three weeks. This timing balanced two competing interests: (1) conducting the reinterview as soon as possible after the original interview to minimize changes in truth between the two interviews, and (2) not making the two interviews so close together that the respondents were simply recalling their previous answers. Interviewers made two call attempts to interview the household member who originally responded, but if that was not possible, the CFU reinterview was conducted with any other eligible household member (15 years or older).

The CFU asked basic demographic questions and a subset of housing and detailed person questions that included all of the topics being tested, with the exception of Telephone Service, and any questions necessary for context and interview flow to set up the questions being tested.[12] All CFU questions were asked in the reinterview, regardless of whether or not a particular question was answered in the original interview. Because the CFU interview was conducted via telephone, the wording of the questions in CFU followed the same format as the CATI nonresponse interviews. Housing units assigned to the control version of the questions in the original interview were asked the control version of the questions in CFU; housing units assigned to the test version of the questions in the original interview were asked the test versions of the question in CFU. The only exception was for retirement, survivor, and disability income, for which a different set of questions was asked in CFU.[13]

2.4. Analysis Metrics

This section describes the metrics used to assess the revised versions of the relationship question, which includes the item missing data rate, response distributions, comparisons to benchmarks, response error, and other metrics. This section also describes the methodology used to calculate unit response rates and standard errors for the test.

All 2016 ACS Content Test data were analyzed without imputation due to our interest in how question changes or differences between versions of new questions affected "raw" responses, not the final edited variables. Some editing of responses was done for analysis purposes, such as

[11] Throughout this report, the "original interview" refers to responses completed via paper questionnaire, internet, CATI, or CAPI.

[12] Because the CFU interview was conducted via telephone the Telephone Service question was not asked. We assume that CFU respondents have telephone service.

[13] Refer to the 2016 ACS Content Test report on Retirement Income for a discussion on CFU questions for survivor, disability, and retirement income.

collapsing response categories or modes together or calculating a person's age based on his or her date of birth.

All estimates from the 2016 ACS Content Test were weighted. Analysis involving data from the original interviews used the final weights that take into account the initial probability of selection (the base weight) and CAPI subsampling. For analysis involving data from the CFU interviews, the final weights were adjusted for CFU nonresponse to create CFU final weights.

The significance level for all hypothesis tests is $\alpha = 0.1$. Since we are conducting numerous comparisons between the control and test treatments, there is a concern about incorrectly rejecting a hypothesis that is actually true (a "false positive" or Type I error). The overall Type I error rate is called the familywise error rate and is the probability of making one or more Type I errors among all hypotheses tested simultaneously. When adjusting for multiple comparisons, the Holm-Bonferroni method was used (Holm, 1979).

2.4.1. Unit Response Rates and Demographic Profile of Responding Households

The unit response rate is generally defined as the proportion of sample addresses eligible to respond that provided a complete or sufficient partial response.[14] Unit response rates from the original interview are an important measure to look at when considering the analyses in this report that compare responses between the control and test versions of the survey questionnaire. High unit response rates are important in mitigating potential nonresponse bias.

For both control and test treatments, we calculated the overall unit response rate (all modes of data collection combined) and unit response rates by mode: internet, mail, CATI, and CAPI. We also calculated the total self-response rate by combining internet and mail modes together. Some 2016 ACS Content Test analyses focused on the different data collection modes for topic-specific evaluations, thus we felt it was important to include each mode in the response rates section. In addition to those rates, we calculated the response rates for high and low response areas because analysis for some Content Test topics was done by high and low response areas. Using the Census Bureau's Planning Database (U.S. Census Bureau, 2016), we defined these areas at the tract level based on the low response score.

The universe for the overall unit response rates consists of all addresses in the initial sample (70,000 addresses) that were eligible to respond to the survey. Some examples of addresses ineligible for the survey were a demolished home, a home under construction, a house or trailer that was relocated, or an address determined to be a permanent business or storage facility. The universe for self-response (internet and mail) rates consists of all mailable addresses that were eligible to respond to the survey. The universe for the CATI response rate consists of all nonrespondents at the end of the mailout month from the initial survey sample that were eligible to respond to the survey and for whom we possessed a telephone number. The universe for the CAPI response rates consists of a subsample of all remaining nonrespondents (after CATI) from the initial sample that were eligible to respond to the survey. Any nonresponding addresses that were sampled out of CAPI were not included in any of the response rate calculations.

[14] A response is deemed a "sufficient partial" when the respondent gets to the first question in the detailed person questions section for the first person in the household.

We also calculated the CFU interview unit response rate overall and by mode of data collection of the original interview and compared the control and test treatments because response error analysis (discussed in Section 2.4.5) relies upon CFU interview data. Statistical differences between CFU response rates for the control and test treatments will not be taken as evidence that one version is better than the other. For the CFU response rates, the universe for each mode consists of housing units that responded to the original questionnaire in the given mode (internet, mail, CATI, or CAPI) and were eligible for the CFU interview. We expected the response rates to be similar between treatments; however, we calculated the rates to verify that assumption.

Another important measure to look at in comparing experimental treatments is the demographic profile of the responding households in each treatment. The 2016 ACS Content Test sample was designed with the intention of having respondents in both control and test treatments exhibit similar distributions of socioeconomic and demographic characteristics. Similar distributions allow us to compare the treatments and conclude that any differences are due to the experimental treatment instead of underlying demographic differences. Thus, we analyzed distributions for data from the following response categories: *age, sex, educational attainment,* and *tenure.* In addition to *relationship,* the topics of *race* and *Hispanic origin* are also typically used for demographic analysis; however, those questions were also modified as part of the 2016 ACS Content Test, so we could not include them in the demographic profile. Additionally, we calculated *average household size* and the *language of response* for the original interview.[15]

For response distributions, we used chi-square tests of independence to determine statistical differences between control and test treatments. If the distributions were significantly different, we performed additional testing on the differences for each response category. To control for the overall Type I error rate for a set of hypotheses tested simultaneously, we performed multiple-comparison procedures with the Holm-Bonferroni method (Holm, 1979). A family for our response distribution analysis was the set of p-values for the overall characteristic categories (*age*, *sex*, *educational attainment,* and *tenure*) and the set of p-values for a characteristic's response categories if the response distributions were found to have statistically significant differences. To determine statistical differences for *average household size* and the *language of response* of the original interview we performed two-tailed hypothesis tests.

For all response-related calculations mentioned in this section, addresses that were either sampled out of the CAPI data collection operation or that were deemed ineligible for the survey were not included in any of the universes for calculations. Unmailable addresses were also excluded from the self-response universe. For all unit response rate estimates, differences, and demographic response analysis, we used replicate base weights adjusted for CAPI sampling (but not adjusted for CFU nonresponse).

2.4.2. Item Missing Data Rates

Respondents leave items blank for a variety of reasons including not understanding the question (clarity), their unwillingness to answer a question as presented (sensitivity), and their lack of knowledge of the data needed to answer the question. The item missing data rate (for a given

[15] Language of response analysis excludes paper questionnaire returns because there was only an English questionnaire.

item) is the proportion of eligible units, housing units for household-level items or persons for person-level items, for which a required response (based on skip patterns) is missing.

For each item, it is important to define carefully both the universe of eligible units and the criteria that determine whether a response to that item is missing or not missing. The definition of *missing* includes *don't know* and *refused to answer* from CATI and CAPI interviews, as well as paper and internet questionnaires where no answer was provided. Since the paper questionnaire does not preclude respondents from marking more than one response category, if more than one box was marked then the answer was considered *missing* for this analysis, since we cannot assume which answer was the correct one.

The universe of eligible persons for the relationship question is all persons who were not the reference person. The percent of eligible persons who did not provide a valid response (valid responses are those where one and only one answer category is marked) to this question in the control treatment were compared to the corresponding percent from the test treatment. Two-tailed t-tests were used to determine significant differences between item missing data rates for the control and test treatments.

2.4.3. Response Distributions

Comparing the response distributions between the control version of a question and the test version of a question allows us to assess whether the question change affected the resulting estimates. Comparisons were made using Rao-Scott chi-squared tests (Rao & Scott, 1987) for distribution or t-tests for single categories when the corresponding distributions were found to be statistically different.

Proportion estimates were calculated as:

$$\text{Category proportion} = \frac{\text{weighted count of valid responses in category}}{\text{weighted count of all valid responses}}$$

All persons, including householder, were included in the universe for the calculations. Because some response categories had very small cell sizes, it was necessary to combine categories to make comparisons between the control and test versions of the question. The *other relative* category was created by combining the categories of *biological son or daughter, adopted son or daughter, stepson or stepdaughter, brother or sister, father or mother, grandchild, parent-in-law, son-in-law or daughter-in-law*, and *other relative*. The *other non-relative* category was created by combining the categories of *roomer or boarder, housemate or roommate, foster child*, and *other non-relative* responses.

To make one-to-one category proportion comparison it was necessary to combine some of the new categories being tested. The test categories of *opposite-sex husband/wife/spouse* and *same-sex husband/wife/spouse* were combined so they could be compared with the control category of *husband or wife*. The test categories of *opposite-sex unmarried partner* and *same-sex unmarried* partner were combined so they could be compared with the control category of *unmarried partner*.

2.4.4. Benchmarks

For the topic of relationship, we compared data from both control and test treatments in the 2016 ACS Content Test and the 2015 NCT. No statistical testing was performed because of differences in the methodology used to create the two datasets. However, since the 2015 NCT is the largest nationally representative test of the revised relationship question to date, it is the best benchmark available to set reasonable expectations for distributions of the relationship categories.

While the revised relationship question has already been tested several times, the best test to use as a benchmark is the 2015 NCT, which was the largest test conducted as part of the decennial testing program leading up to the 2020 Census. The 2015 NCT utilized a sample of 1.2 million housing units to test content modifications, different contact strategies designed to optimize self-response, and different approaches to offering in-language materials. Two versions of the relationship question were tested in the 2015 NCT: (1) a slightly modified version of the 2010 Census question (control) and (2) a new version where the two couple relationship categories were expanded to distinguish between opposite-sex and same-sex relationships (test). The control and test versions of the question were the same in the 2016 ACS Content Test as in the 2015 NCT. The 2015 NCT allowed self-response via paper questionnaire, internet, or Telephone Questionnaire Assistance (TQA). There was no nonresponse follow-up (NRFU) operation. Since the 2015 NCT included only self-response, we compared only the combined (all modes) and internet responses with the 2015 NCT results. We do not include a straight comparison of paper modes responses because in 2015 NCT households might not have received the same version of the relationship question between internet and paper modes. All 2015 NCT households had the option to respond by internet.

2.4.5. Response Error

Response error occurs for a variety of reasons, such as flaws in the survey design, misunderstanding of the questions, misreporting by respondents, or interviewer effects. There are two components of response error: response bias and simple response variance. Response bias is the degree to which respondents consistently answer a question incorrectly. Simple response variance is the degree to which respondents answer a question inconsistently. A question has good response reliability if respondents tend to answer the question consistently. Re-asking the same question of the same respondent (or housing unit) allows us to measure response variance.

We measured simple response variance by comparing valid responses to the CFU reinterview with valid responses to the corresponding original interview.[16] The Census Bureau has frequently used content reinterview surveys to measure simple response variance for large demographic data collection efforts, including the 2010 ACS Content Test, and the 1990, 2000, and 2010 decennial censuses (Dusch & Meier, 2012).

[16] A majority of the CFU interviews were conducted with the same respondent as the original interview (see the Limitations section for more information).

The following measures were used to evaluate consistency:

- Gross difference rate (GDR)
- Index of inconsistency (IOI)
- L-fold index of inconsistency (IOI_L)

The first two measures – GDR and IOI – were calculated for individual response categories. The L-fold index of inconsistency was calculated for questions that had three or more mutually exclusive response categories, as a measure of overall reliability for the question.

The GDR, and subsequently the simple response variance, are calculated using the following table and formula.

Table 1. Interview and Reinterview Counts for Each Response Category Used for Calculating the Gross Difference Rate and Index of Inconsistency

	Original Interview "Yes"	Original Interview "No"	**Reinterview Totals**
CFU Reinterview "Yes"	a	b	**a + b**
CFU Reinterview "No"	c	d	**c + d**
Original Interview Totals	**a + c**	**b + d**	**n**

Where a, b, c, d, and n are defined as follows:

a = weighted count of units in the category of interest for both the original interview and reinterview
b = weighted count of units NOT in the category of interest for the original interview, but in the category for the reinterview
c = weighted count of units in the category of interest for the original interview, but NOT in the category for the reinterview
d = weighted count of units NOT in the category of interest for either the original interview or the reinterview
n = total units in the universe = a + b + c + d.

The GDR for a specific response category is the percent of inconsistent answers between the original interview and the reinterview (CFU). We calculate the GDR for a response category as

$$\text{GDR} = \frac{(b + c)}{n} \times 100$$

Statistical significance between the GDR for a specific response category between the control and test treatments is determined using a two-tailed t-test.

In order to define the IOI, we must first discuss the variance of a category proportion estimate. If we are interested in the true proportion of a total population that is in a certain category, we can use the proportion of a survey sample in that category as an estimate. Under certain reasonable assumptions, it can be shown that the total variance of this proportion estimate is the sum of two

components, sampling variance (SV) and simple response variance (SRV). It can also be shown that an unbiased estimate of SRV is half of the GDR for the category (Flanagan, 1996).

SV is the part of total variance resulting from the differences among all the possible samples of size *n* one might have selected. SRV is the part of total variance resulting from the aggregation of response error across all sample units. If the responses for all sample units were perfectly consistent, then SRV would be zero, and the total variance would be due entirely to SV. As the name suggests, the IOI is a measure of how much of the total variance is due to inconsistency in responses, as measured by SRV and is calculated as:

$$\mathrm{IOI} = \frac{\mathrm{n(b+c)}}{\mathrm{(a+c)(c+d)+(a+b)(b+d)}} \times 100$$

Per the Census Bureau's general rule, index values of less than 20 percent indicate low inconsistency, 20 to 50 percent indicate moderate inconsistency, and over 50 percent indicate high inconsistency.

An IOI is computed for each response category and an overall index of inconsistency, called the L-fold index of inconsistency, is reported for the entire distribution. The L-fold index is a weighted average of the individual indexes computed for each response category.

When the sample size is small, the reliability estimates are unstable. Therefore, we do not report the IOI and GDR values for categories with a small sample size, as determined by the following formulas: $2a + b + c < 40$ or $2d + b + c < 40$, where a, b, c, and d are unweighted counts as shown in Table 1 above (see Flanagan 1996, p. 15).

The measures of response error assume that those characteristics in question did not change between the original interview and the CFU interview. To the extent that this assumption is incorrect, we assume that it is incorrect at similar rates between the control and test treatments. An example of this could be a household that originally identified a couple as unmarried but might have married before the CFU interview and then accurately reported a different response than in the original interview.

In calculating the IOI reliability measures, the assumption is that the expected value of the error in the original interview is the same as in the CFU reinterview. This assumption of parallel measures is necessary for the SRV and IOI to be valid. In calculating the IOI measures for this report, we found this assumption was not met for the response categories specified in the limitations section (see Section 4).

Biemer (2011, pp. 56-58) provides an example where the assumption of parallel measures is not met, but does not provide definitive guidelines for addressing it. In Biemer's concluding remarks, he states, "...both estimates of reliability are biased to some extent because of the failure of the parallel assumptions to hold." Flanagan (2001) addresses this bias problem and offers the following adjustment to the IOI formula:

$$IOI_{testimate} = \frac{\frac{n^2(b+c) - n(c-b)^2}{n-1}}{(a+c)(c+d) + (a+b)(b+d)} \times 100$$

This formula was tested on selected topics in the 2016 ACS Content Test. The $IOI_{testimate}$ resulted in negligible reduction in the IOI values. For this reason, we did not recalculate the IOI values using $IOI_{testimate}$. Similar to Biemer (2011, p. 58), we acknowledge that for some cases, the estimate of reliability is biased to some extent.

The universe of eligible persons for the relationship question are all persons who were not the reference person in the original interview and in the CFU interview. We also excluded cases with a different reference person in CFU than in the original interview.

A person could have changed their relationship status between the original response and the CFU response, such as by getting married in the time between each data collection. We assumed this would happen at the same rate in the control universe as in the test universe, so the resulting error or inconsistency seen in the reliability measures should be comparable and should not affect the conclusions. Statistical significance between the GDRs and IOIs of each version were determined using a two-tailed t-test.

2.4.6. Analysis of the Relationship/Sex Consistency Check

One motivation for the revision of the relationship question was the discovery of an error that particularly affects estimates of same-sex married couple households. If a very small proportion of a very large group accidentally mismarks responses, it can cause them to appear to have reported being part of a much smaller related group. If just 0.5 percent of the roughly 56 million opposite-sex married couple households mistakenly report as same-sex married couples, this could add 280,000 households to the estimate of same-sex married couples, which may only be as large as 500,000 total (Kreider, Bates and Lofquist, 2016).

In the 2016 ACS Content Test, we included a relationship/sex consistency check in internet and computer assisted (CATI/CAPI) data collection modes. The purpose of the check is to reduce inconsistent responses in which the value reported for relationship in a coupled household does not agree with the sex values reported for the householder and her or his partner or spouse. For example, if John is the *householder* and Mary is reported as John's *same-sex spouse*, but John's sex is reported as *male* and Mary's sex as *female*, then the automated check would be triggered.

In the 2016 ACS Content Test internet instrument, the check functioned much as it did in the 2015 NCT. In the electronic modes, the check was triggered if the values of relationship and sex were inconsistent. Respondents were first asked to confirm the relationship value (e.g., *same-sex spouse*) they reported. If they confirmed that relationship was correct, then they were asked to confirm the sex value (e.g., *female*) they reported for the spouse/partner. If they confirmed that value was correct, then they were asked to confirm the sex value (e.g., *male*) they reported for the householder. If they reported that any of these values was not correct, they were re-asked the question and had an opportunity to change their response, though they were not required to do so. All respondents who triggered the check received the three confirmation questions. A very similar check sequence was triggered in CATI/CAPI if the relationship and sex values were

inconsistent for householders and their spouses or partners. Since this check has the potential to reduce mismarking errors that affect the estimates of same-sex married couple households, we also present results for the rate of inconsistent responses and the operation of the check in this report.

The analyses were conducted considering one couple per household, instead of two separate responses for each individual. No statistical testing was done for the analysis of the relationship/sex consistency check.

2.4.7. Standard Error Calculations

We estimated the variances of the estimates using the Successive Differences Replication (SDR) method with replicate weights, the standard method used in the ACS (see U.S. Census Bureau, 2014, Chapter 12). We calculated the variance for each rate and difference using the formula below. The standard error of the estimate (X_0) is the square root of the variance:

$$\mathrm{Var}(X_0) = \frac{4}{80}\sum_{r=1}^{80}(X_r - X_0)^2$$

where:

X_0 = the estimate calculated using the full sample,
X_r = the estimate calculated for replicate r.

3. KEY RESEARCH FOR RELATIONSHIP

Before fielding the 2016 ACS Content Test, we identified which of the metrics would be given higher importance in determining which version of the question would be recommended for inclusion in the ACS moving forward. Table 2 identifies the research questions and associated metrics in priority order.

Table 2. Key Research for the Relationship Question

Research Questions	Key Research, in order of priority
2	The distributions of the control and test versions should have minimal to no differences.
1	The item missing data rates for the test version should be the same or lower than the control version.
3	Response reliability (gross difference rate, index of inconsistency) should not differ between the test and control treatments.
4	Upon receiving automated relationship/sex consistency check, respondents will use this opportunity to change responses.

4. LIMITATIONS

CATI and CAPI interviewers were assigned control and test treatment cases, as well as production cases. The potential risk of this approach is the introduction of a cross-contamination or carry-over effect due to the same interviewer administering multiple versions of the same question item. Interviewers are trained to read the questions verbatim to minimize this risk, but there still exists the possibility that an interviewer may deviate from the scripted wording of one question version to another. This could potentially mask a treatment effect from the data collected.

2016 ACS Content Test interviews were only conducted in English and Spanish. Respondents who needed language assistance in another language were not able to participate in the test. Additionally, the Content Test was not conducted in Alaska, Hawaii, or Puerto Rico. Any conclusions drawn from this test may not apply to these areas or populations.

For statistical analysis specific to the mail mode, there may be bias in the results because of unexplained unit response rate differences between the control and test treatments.

We were not able to conduct demographic analysis by race or ethnicity because these topics were tested as part of the Content Test.

The CFU reinterview was not conducted in the same mode of data collection for households that responded by internet, mail, or CAPI in the original interview since CFU interviews were only administered using a CATI mode of data collection. As a result, the data quality measures derived from the reinterview may include some bias due to the differences in mode of data collection.

To be eligible for a CFU reinterview, respondents needed to either provide a telephone number in the original interview or have a telephone number available to the Census Bureau through reverse address look up. As a result, 2,284 of the responding households (11.8 percent with a standard error of 0.2) from the original control interviews and 2,402 of the responding households (12.4 percent with a standard error of 0.2) from the original test interviews were not eligible for the CFU reinterview. The difference between the control and test treatments was statistically significant (p-value=0.06).

The 2016 ACS Content Test does not include the production weighting adjustments for seasonal variations in ACS response patterns, nonresponse bias, and under-coverage bias. As a result, any estimates derived from the 2016 ACS Content Test data do not provide the same level of inference as the production ACS and cannot be compared to production estimates.

In developing initial workload estimates for CATI and CAPI, we did not take into account the fact that we oversampled low response areas as part of the 2016 ACS Content Test sample design. Therefore, workload and budget estimates were too low. In order to stay within budget, the CAPI workload was subsampled more than originally planned. This caused an increase in the variances for the analysis metrics used.

An error in addressing and assembling the materials for the 2016 ACS Content Test caused some cases to be mailed production ACS questionnaires instead of 2016 ACS Content Test questionnaires. There were 49 of these cases that returned completed questionnaires, and they were all from the test treatment. These cases were excluded from the analysis. Given the small number of cases affected by this error, there is very little effect on the results.

Questionnaire returns were expected to be processed and keyed within two weeks of receipt. Unfortunately, a check-in and keying backlog prevented this requirement from being met, thereby delaying eligible cases from being sent to CFU on a schedule similar to the other modes. Additionally, the control treatment questionnaires were processed more quickly in keying than the test treatment questionnaires resulting in a longer delay for test mail cases to be eligible for CFU. On average, it took 18 days for control cases to become eligible for CFU; it took 20 days for test cases. The difference is statistically significant. This has the potential to impact the response reliability results.

The assumption of parallel measures for the GDR and IOI calculations was not met for the following relationship category: *relative* or a person who is related to the householder in some way but is not a *spouse, unmarried partner* or *non-relative*. For the *relative* category, the GDR and IOI estimates are biased to some extent.

Limitations specific to the relationship question include the fact that some respondents may not identify with the answer categories provided, or may have a different understanding of *related* and *not related* than the question designers. While these distinctions are explained in the help text, we know that few respondents actually access the help text. Understanding of the new couple categories may be affected for those whose first language is not English. Respondents who are not familiar with the U.S. foster care system may have trouble understanding exactly what is meant by the *foster child* category (Goerman, Meyers, Simmons, forthcoming; Meyers et al., forthcoming; Goerman et al., 2014). Another limitation is that the electronic modes included an automated check for consistency between the relationship and sex reports of the householder and their spouse/partner, while paper could not.

5. RESEARCH QUESTIONS AND RESULTS

This section presents the results from the analyses of the 2016 ACS Content Test data for the relationship question. An analysis of unit response rates is presented first followed by topic-specific analyses. For the topic-specific analyses, each research question is restated first, followed by corresponding data and a brief summary of the results.

5.1. Unit Response Rates and Demographic Profile of Responding Households

This section provides results for unit response rates for both control and test treatments for the original interview and for the CFU interview. It also provides results of a comparison of socioeconomic and demographic characteristics of respondents in both control and test treatments.

5.1.1. Unit Response Rates for the Original Content Test Interview

The unit response rate is generally defined as the proportion of sample addresses eligible to respond that provided a complete or sufficient partial response. We did not expect the unit response rates to differ between treatments. This is important because the number of unit responses should also affect the number of item responses we receive for analyses done on specific questions on the survey. Similar item response universe sizes allow us to compare the treatments and conclude that any differences are due to the experimental treatment instead of differences in the populations sampled for each treatment.

Table 3 shows the unit response rates for the original interview for each mode of data collection (internet, mail, CATI, and CAPI), all modes combined, and both self-response modes (internet and mail combined) for the control and test treatments. When looking at the overall unit response rate (all modes combined) the difference between control (93.5 percent) and test (93.5 percent) is less than 0.1 percentage points and is not statistically significant.

Table 3. Original Interview Unit Response Rates for Control and Test Treatments, Overall and by Mode

Mode	Test Interviews	Test Percent	Control Interviews	Control Percent	Test minus Control	P-Value
All Modes	19,400	93.5 (0.3)	19,455	93.5 (0.3)	<0.1 (0.4)	0.98
Self-Response	13,131	52.9 (0.5)	13,284	53.7 (0.5)	-0.8 (0.6)	0.23
Internet	8,168	34.4 (0.4)	8,112	34.1 (0.4)	0.4 (0.6)	0.49
Mail	4,963	18.4 (0.3)	5,172	19.6 (0.3)	-1.2 (0.5)	0.01*
CATI	872	8.7 (0.4)	880	9.2 (0.4)	-0.4 (0.6)	0.44
CAPI	5,397	83.5 (0.7)	5,291	83.6 (0.6)	<0.1 (0.9)	0.96

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. P-values with an asterisk (*) indicate a significant difference based on a two-tailed t-test at the α=0.1 level. The weighted response rates account for initial sample design as well as CAPI subsampling.

When analyzing the unit response rates by mode of data collection, the only modal comparison that shows a statistically significant difference is the mail response rate. The control treatment had a higher mail response (19.6 percent) than the test treatment (18.4 percent) by 1.2 percentage points. As a result of this difference, we looked at how mail responses differed in the high and low response areas. Table 4 shows the mail response rates for both treatments in high and low response areas.[17] The difference in mail response rates appears to be driven by the difference of rates in the high response areas.

It is possible that the difference in the mail response rates between control and test is related to the content changes made to the test questions. There are some test questions that could be perceived as being too sensitive by some respondents (such as the test question relating to same-sex relationships) and some test questions that could be perceived to be too burdensome by some respondents (such as the new race questions with added race categories). In the automated modes (internet, CATI, and CAPI) there is a higher likelihood of obtaining a sufficient partial response

[17] Table A-1 (including all modes) can be found in Appendix A.

(obtaining enough information to be deemed a response for calculations before the respondent stops answering questions) than in the mail mode. If a respondent is offended by the questionnaire or feels that the questions are too burdensome, they may just throw the questionnaire away and not respond by mail. This could be a possible explanation for the unit response rate being lower for test than control in the mail mode.

We note that differences between overall and total self-response response rates were not statistically significant. As most analysis was conducted at this level, we are confident the response rates were sufficient to conduct topic-specific comparisons between the control and test treatments and that there are no underlying response rate concerns that would impact those findings.

Table 4. Mail Response Rates by Designated High (HRA) and Low (LRA) Response Areas

	Test Interviews	Test Percent	Control Interviews	Control Percent	Test minus Control	P-Value
HRA	2,082	20.0 (0.4)	2,224	21.5 (0.4)	-1.5 (0.6)	0.02*
LRA	2,881	13.8 (0.3)	2,948	14.1 (0.3)	-0.3 (0.4)	0.43
Difference		6.2 (0.5)		7.4 (0.4)	-1.1 (0.7)	0.11

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. P-values with an asterisk (*) indicate a significant difference based on a two-tailed t-test at the α=0.1 level. The weighted response rates account for initial sample design as well as CAPI subsampling.

5.1.2. Unit Response Rates for the Content Follow-Up Interview

Table 5 shows the unit response rates for the CFU interview by mode of data collection of the original interview and for all modes combined, for control and test treatments. Overall, the differences in CFU response rates between the treatments are not statistically significant. The rate at which CAPI respondents from the original interview responded to the CFU interview is lower for test (34.8 percent) than for control (37.7 percent) by 2.9 percentage points. While the protocols for conducting CAPI and CFU were the same between the test and control treatments, we could not account for personal interactions that occur in these modes between the respondent and interviewer. This can influence response rates. We do not believe that the difference suggests any underlying CFU response issues that would negatively affect topic-specific response reliability analysis for comparing the two treatments.

Table 5. Content Follow-Up Interview Unit Response Rates for Control and Test Treatments, Overall and by Mode of Original Interview

Original Interview Mode	Test Interviews	Test Percent	Control Interviews	Control Percent	Test minus Control	P-Value
All Modes	7,867	44.8 (0.5)	7,903	45.7 (0.6)	-0.8 (0.8)	0.30
Internet	4,078	51.9 (0.6)	4,045	52.5 (0.7)	-0.6 (0.8)	0.49
Mail	2,202	46.4 (0.9)	2,197	44.2 (0.9)	2.1 (1.3)	0.11
CATI	369	48.9 (1.9)	399	51.5 (2.5)	-2.5 (2.9)	0.39
CAPI	1,218	34.8 (1.2)	1,262	37.7 (1.1)	-2.9 (1.6)	0.07*

Source: U.S. Census Bureau, 2016 American Community Survey Content Test

Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. P-values with an asterisk (*) indicate a significant difference based on a two-tailed t-test at the α=0.1 level.

5.1.3. Demographic and Socioeconomic Profile of Responding Households

One of the underlying assumptions of our analyses in this report is that the sample for the Content Test was selected in such a way that responses from both treatments would be comparable. We did not expect the demographics of the responding households for control and test treatments to differ. To test this assumption, we calculated distributions for respondent data for the following response categories: *age, sex, educational attainment,* and *tenure.*[18] The response distribution calculations can be found in Table 6. Items with missing data were not included in the calculations. After adjusting for multiple comparisons, none of the differences in the categorical response distributions shown below is statistically significant.

Table 6. Response Distributions: Test versus Control Treatment

Item	Test Percent	Control Percent	Adjusted P-Value
AGE	(n=43,236)	(n=43,325)	0.34
Under 5 years old	5.7 (0.2)	6.1 (0.2)	
5 to 17 years old	17.8 (0.3)	17.6 (0.3)	
18 to 24 years old	8.6 (0.3)	8.1 (0.3)	
25 to 44 years old	25.1 (0.3)	26.2 (0.3)	
45 to 64 years old	26.8 (0.4)	26.6 (0.4)	
65 years old or older	16.0 (0.3)	15.4 (0.3)	
SEX	(n=43,374)	(n=43,456)	1.00
Male	48.8 (0.3)	49.1 (0.3)	
Female	51.2 (0.3)	50.9 (0.3)	
EDUCATIONAL ATTAINMENT#	(n=27,482)	(n=27,801)	1.00
No schooling completed	1.3 (0.1)	1.2 (0.1)	
Nursery to 11th grade	8.1 (0.3)	8.0 (0.3)	
12th grade (no diploma)	1.7 (0.1)	1.6 (0.1)	
High school diploma	21.7 (0.4)	22.3 (0.4)	
GED† or alternative credential	3.5 (0.2)	3.6 (0.2)	
Some college	21.0 (0.4)	20.2 (0.4)	
Associate's degree	8.8 (0.3)	9.1 (0.3)	
Bachelor's degree	20.9 (0.4)	20.3 (0.4)	
Advanced degree	13.1 (0.3)	13.7 (0.3)	
TENURE	(n=17,190)	(n=17,236)	1.00
Owned with a mortgage	43.1 (0.6)	43.2 (0.5)	
Owned free and clear	21.1 (0.4)	21.2 (0.4)	
Rented	33.8 (0.6)	34.0 (0.5)	
Occupied without payment of rent	1.9 (0.2)	1.7 (0.1)	

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
#For ages 25 and older
†General Educational Development
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance testing done at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method.

[18] We were not able to conduct demographic analysis by race or ethnicity because these topics were tested as part of the Content Test.

We also analyzed two other demographic characteristics shown by the responses from the survey: *average household size* and *language of response.* The results for the remaining demographic analyses can be found in Table 7 and Table 8 below.

Table 7. Comparison of Average Household Size

	Test (n=17,608)	Control (n=17,694)	Test minus Control	P-value
Average Household Size (Number of People)	2.51 (<0.1)	2.52 (<0.1)	>-0.01 (<0.1)	0.76

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Standard errors are shown in parentheses. Significance was tested based on a two-tailed t-test at the α=0.1 level.

Table 8. Comparison of Language of Response

Language of Response	Test Percent (n=17,608)	Control Percent (n=17,694)	Test minus Control	P-value
English	96.1 (0.2)	96.2 (0.2)	<0.1 (0.3)	0.52
Spanish	2.7 (0.2)	2.6 (0.2)	<0.1 (0.2)	0.39
Undetermined	1.2 (0.1)	1.2 (0.1)	<0.1 (0.2)	0.62

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a two-tailed t-test at the α=0.1 level.

The 2016 ACS Content Test was available in two languages, English and Spanish, for all modes except the mail mode. However, the language of response variable was missing for some responses, so we created a category called *undetermined* to account for those cases.

There are no detectable differences between control and test for *average household size* or *language of response*. There are also no differences for any of the response distributions that we calculated. As a result of these analyses, it appears that respondents in both treatments do exhibit comparable demographic characteristics since none of the resulting findings is significant, which verifies our assumption of demographic similarity between treatments.

5.2. Item Missing Data Rates

Are item missing data rates lower in the test treatment than in the control treatment?

Table 9 shows the percentage of missing responses in the relationship question in the control and test treatment. "Overall" combines information from the internet, mail and computer-assisted interview modes, which are also provided separately. The results show there were no statistically significant differences overall or in most of the modes. We did find that there was a significant difference for the internet mode, but the magnitude of the difference (0.3 percent) was small and the missing data rates were low for both treatments.

Table 9. Item Missing Data Rates for Control and Test Treatments, by Mode

Mode	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	P-Value
Overall	25,997	0.4 (0.1)	25,983	0.3 (0.1)	0.1 (0.1)	0.23
Internet	12,997	0.3 (0.1)	12,788	<0.1 (<0.1)	0.3 (0.3)	0.02*
Mail	5,279	1.5 (0.2)	5,567	1.1 (0.2)	0.3 (0.1)	0.29
CATI/CAPI	7,721	0.4 (0.1)	7,628	0.2 (0.1)	-0.1 (0.1)	0.54

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. P-values with an asterisk (*) indicate a significant difference based on a two-tailed t-test at the α=0.1 level

5.3. Response Distributions

Does the distribution of people reported as a spouse or partner differ between control and test treatments?

Tables 10, 11, 12, and 13 show that the only mode with a significant difference between response distributions was the internet response mode (Table 11). For internet, no relationship response categories were significantly different. The only statistically significant difference was the *missing* response category. The percentage point difference was 0.2 percent with a standard error of 0.1 and a p-value of 0.02.

Table 10. Response Distributions for Control and Test Treatments: All Modes Combined

Response Category	Test Percent (n=43,593)	Control Percent (n=43,671)
Householder	39.9 (0.2)	39.8 (0.2)
Spouse vs. Husband or Wife	19.4 (0.2)	19.4 (0.2)
Unmarried Partner	2.7 (0.1)	2.6 (0.1)
Other Relative[1]	34.5 (0.4)	34.7 (0.3)
Other Non-relative[2]	3.2 (0.2)	3.2 (0.2)
Missing	0.2 (<0.1)	0.1 (<0.1)
Multiple Marks	<0.1 (<0.1)	<0.1 (<0.1)

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 3.4$, p-value=0.76, Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding.
[1] Includes the following: biological son or daughter, adopted son or daughter, stepson or stepdaughter, brother or sister, father or mother, grandchild, parent-in-law, son-in-law or daughter-in-law, and other relative.
[2] Includes the following: roomer or boarder, housemate or roommate, foster child, and other non-relative.

Table 11. Response Distributions for Control and Test Treatments: Internet Mode

Response Category	Test Percent (n=21,102)	Control Percent (n=20,861)
Householder	38.5 (0.2)	38.9 (0.2)
Spouse vs. Husband or Wife	22.8 (0.2)	23.1 (0.2)
Unmarried Partner	2.5 (0.1)	2.4 (0.1)
Other Relative[1]	33.0 (0.4)	32.7 (0.4)
Other Non-relative[2]	3.0 (0.2)	2.9 (0.2)
Missing*	0.2 (0.1)	<0.1 (<0.1)

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 33.18$, p-value=<0.01. Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. An asterisk (*) indicates a significant difference between the test and control percent based on a two-tailed t-test (test ≠ control) at the α=0.1 level.
[1] Includes the following: biological son or daughter, adopted son or daughter, stepson or stepdaughter, brother or sister, father or mother, grandchild, parent-in-law, son-in-law or daughter-in-law, and other relative.
[2] Includes the following: roomer or boarder, housemate or roommate, foster child, and other non-relative.

Table 12. Response Distributions for Control and Test Treatments: Mail Mode

Response Category	Test Percent (n=10,126)	Control Percent (n=10,623)
Householder	48.0 (0.4)	47.7 (0.5)
Spouse vs. Husband or Wife	21.8 (0.4)	21.4 (0.4)
Unmarried Partner	2.1 (0.2)	1.9 (0.2)
Other Relative[1]	25.7 (0.5)	26.2 (0.6)
Other Non-relative[2]	1.7 (0.2)	2.2 (0.2)
Missing	0.7 (0.1)	0.5 (0.1)
Multiple Marks	0.1 (<0.1)	0.1 (<0.1)

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 9.53$, p-value=0.15. Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding.
[1] Includes the following: biological son or daughter, adopted son or daughter, stepson or stepdaughter, brother or sister, father or mother, grandchild, parent-in-law, son-in-law or daughter-in-law, and other relative.
[2] Includes the following: roomer or boarder, housemate or roommate, foster child, and other non-relative.

Table 13. Response Distributions for Control and Test Treatments: CATI and CAPI Modes Combined

Response Category	Test Percent (n=12,365)	Control Percent (n=12,187)
Householder	37.8 (0.5)	37.0 (0.4)
Spouse vs. Husband or Wife	14.8 (0.4)	14.7 (0.4)
Unmarried Partner	3.2 (0.3)	3.2 (0.2)
Other Relative[1]	40.0 (0.7)	41.0 (0.6)
Other Non-relative[2]	4.1 (0.5)	4.0 (0.4)
Missing	0.1 (<0.1)	0.1 (0.1)

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 1.52$, p-value=0.9. Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding.
[1] Includes the following: biological son or daughter, adopted son or daughter, stepson or stepdaughter, brother or sister, father or mother, grandchild, parent-in-law, son-in-law or daughter-in-law, and other relative.
[2] Includes the following: roomer or boarder, housemate or roommate, foster child, and other non-relative.

When looking specifically at the distribution of coupled categories, the estimated percentage of respondents who report as *husband or wife* was not significantly different between the control and test treatments. Likewise, the percentage of household members reported as *unmarried partner* was not significantly different between control and test treatments.

5.4. Benchmarks

While no statistical testing was performed because of differences in the methodology of the tests, the 2016 ACS Content Test estimates compare well with those in the 2015 NCT estimates. The 2016 ACS Content Test shows about 19 percent of household residents who were identified as the *spouse* of the householder in the test treatment (see Table 14). People reported as the *same-sex spouse* of the householder comprised 0.2 percent of all people. Those reported as the *unmarried partner* of the householder comprised 2.7 percent of people in the 2016 ACS Content Test. Looking just at the internet mode (see Table 15), about 23 percent of people were reported as the *spouse* of the householder in the test treatment, while about 2.5 percent of people were reported as the *unmarried partner* of the householder.

Table 14. Response Distributions for 2016 ACS Content Test and 2015 National Content Test: All Modes Combined

Response Category	2016 ACS Test Percent (n=43,593)	2015 NCT Test Percent (n=658,115)	2016 ACS Control Percent (n=43,671)	2015 NCT Control Percent (n=671,829)
Householder	39.9 (0.2)	41.1 (<0.1)	39.8 (0.2)	41.0 (<0.1)
Spouse or Husband or Wife	19.4 (0.2)	21.2 (0.1)	19.4 (0.2)	21.2 (0.1)
Opposite-sex Spouse	19.3 (0.2)	21.0 (0.2)	NA	NA
Same-Sex Spouse	0.2 (<0.1)	0.2 (<0.1)	NA	NA
Unmarried Partner	2.7 (0.1)	2.4 (<0.1)	2.6 (0.1)	2.4 (<0.1)
Opposite-sex Partner	2.6 (0.1)	2.2 (<0.1)	NA	NA
Same-Sex Partner	0.1 (<0.1)	0.2 (<0.1)	NA	NA
Other Relative[1]	34.5 (0.4)	32.0 (0.0)	34.7 (0.3)	32.0 (0.0)
Other Non-relative[2]	3.2 (0.2)	2.9 (<0.1)	3.2 (0.2)	2.9 (<0.1)
Missing	0.2 (<0.1)	0.4 (<0.1)	0.1 (<0.1)	0.4 (<0.1)
Multiple Marks	<0.1 (0.0)	<0.1 (<0.1)	<0.1 (<0.1)	0.0 (<0.1)

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, 2015 National Content Test
Note: NA indicates not applicable. Estimates across surveys are not statistically comparable.
[1] Includes the following: biological son or daughter, adopted son or daughter, stepson or stepdaughter, brother or sister, father or mother, grandchild, parent-in-law, son-in-law or daughter-in-law, and other relative.
[2] Includes the following: roomer or boarder, housemate or roommate, foster child, and other non-relative.

The 2015 NCT test treatment shows 21 percent as *spouses*, about 0.2 percent as the *same-sex spouse* of the householder, and about 2.4 percent as *unmarried partners* (see Table 14). Focusing only on the internet mode, 22 percent of people were reported as the *spouse* of the householder, while 2.5 percent were reported as *unmarried partners*. Overall, the 2016 ACS Content Test estimates are in line with the estimates generated by the 2015 NCT.

Table 15. Response Distributions for 2016 ACS Content Test and 2015 National Content Test: Internet Mode

Response Category	2016 ACS Test Percent (n=21,102)	2015 NCT Test Percent (n=468,804)	2016 ACS Control Percent (n=20,861)	2015 NCT Control Percent (n=471,440)
Householder	38.5 (0.2)	39.1 (<0.1)	38.9 (0.2)	39.0 (<0.1)
Spouse or Husband or Wife	22.8 (0.2)	22.4 (0.1)	23.1 (0.2)	22.4 (0.1)
Opposite-sex Spouse	22.6 (0.2)	22.1 (0.2)	NA	NA
Same-Sex Spouse	0.2 (<0.1)	0.2 (<0.1)	NA	NA
Unmarried Partner	2.5 (0.1)	2.5 (<0.1)	2.4 (0.1)	2.5 (<0.1)
Opposite-sex Partner	2.3 (0.1)	2.2 (<0.1)	NA	NA
Same-Sex Partner	0.2 (<0.1)	0.2 (<0.1)	NA	NA
Other Relative[1]	33.0 (0.4)	32.8 (0.0)	32.7 (0.4)	33.0 (0.0)
Other Non-relative[2]	3.0 (0.2)	3.0 (<0.1)	2.9 (0.2)	2.9 (<0.1)
Missing	0.2 (<0.1)	0.3 (<0.1)	<0.1 (0.0)	0.3 (<0.1)
Multiple Marks	NA	NA	NA	NA

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, 2015 National Content Test
Note: NA indicates not applicable. Estimates across surveys are not statistically comparable.
[1] Includes the following: biological son or daughter, adopted son or daughter, stepson or stepdaughter, brother or sister, father or mother, grandchild, parent-in-law, son-in-law or daughter-in-law, and other relative.
[2] Includes the following: roomer or boarder, housemate or roommate, foster child, and other non-relative.

5.5. Response Error

Do the measures of response reliability (gross difference rate, index of inconsistency) differ between the test and control treatment?

When measuring reliability we grouped responses into four categories for comparisons: *spouse, unmarried partner, relative*, and *non-relative*. After analyzing the reinterview data, we found only *unmarried partner* had a statistically different level of reliability (see Table 16).

Table 16. Difference in Gross Difference Rates (GDR) between Control and Test Treatments

Category	Test GDR Percent	Control GDR Percent	Test minus Control	P-Value
Spouse	0.9 (0.2)	0.8 (0.1)	0.1 (0.2)	0.61
Unmarried Partner	2.2 (0.4)	1.5 (0.2)	0.7 (0.4)	<0.10*
Relative	1.6 (0.4)	1.4 (0.2)	0.3 (0.5)	0.61
Non-Relative	2.3 (0.3)	2.1 (0.3)	0.1 (0.4)	0.75

Source: U.S. Census Bureau, American Community Survey, 2016 Content Test
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. P-values with an asterisk (*) indicate a significant difference based on a two-tailed t-test at the α=0.1 level

The index of inconsistency (IOI) in responses to the relationship question is shown in Table 17. *Unmarried partner* was the only category with a significant difference in this analysis. The IOI for *unmarried partner* in the test group is moderate (28 percent) and the IOI for the control group

in this category is low (15 percent). Overall, the IOI L-fold analysis found no significant difference in the reliability of the question as a whole.

Table 17. Difference in Index of Inconsistency (IOI) between Control and Test Treatments

Category	Test IOI Percent	Control IOI Percent	Test minus Control	P-Value
Spouse	2.0 (0.3)	1.7 (0.3)	0.3 (0.5)	0.58
Unmarried Partner	28.0 (4.0)	15.1 (2.1)	13.0 (4.3)	<0.01*
Relative	3.3 (0.9)	2.8 (0.5)	0.6 (1.1)	0.61
Non-Relative	20.8 (3.5)	21.1 (2.8)	-0.3 (4.4)	0.93
IOI L-fold	6.1 (0.9)	5.0 (0.5)	1.1 (1.0)	0.87

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. P-values with an asterisk (*) indicate a significant difference based on a two-tailed t-test at the α=0.1 level

5.6. Results for the Analysis of the Relationship/Sex Consistency Check

Does the inclusion of an automated relationship/sex consistency check improve data quality?

The results shown did not undergo statistical testing due to a small number of couples receiving the checks; however, they do provide useful information on the function of the automated check.[19] Data in Tables 18 and 19 come from the test version of the survey, are unweighted and the differences were not statistically tested.

Table 18. Percent with Inconsistent Relationship and Sex, by Couple Type

Relationship	Sample Size	% Consistent	% Inconsistent
Total	9,264	99.4	0.6
Married, opposite-sex	8,072	99.9	0.1
Married, same-sex	76	57.9	42.1
Unmarried, opposite-sex	1,050	99.2	0.8
Unmarried, same-sex	66	92.4	7.6

Source: U.S. Census Bureau, American Community Survey, 2016 Content Test
Note: Data are unweighted. Differences were not statistically tested.

[19] Note that the small number of cases receiving the check is to be expected, given that only a very small percentage of cases report an inconsistency between sex and relationship for coupled households. However, even this small number of cases has a disproportionately large effect on estimates of the relatively small population of same-sex couples, especially married couples.

Table 19. Percent Receiving Relationship/Sex Consistency Check, by Couple Type: Internet Mode

Relationship	Total	Married, opposite-sex	Married, same-sex	Unmarried, opposite-sex	Unmarried, same-sex
Total (n)	**5,016**	**4,423**	**40**	**512**	**41**
Percent of total that received the consistency check (%)	0.7	0.5	22.5	0.2	2.4
Received the consistency check (n)	**34**	**23**	**9**	**1**	**1**
Percent that changed their response, of those who received check (%)	64.7	91.3	0.0	100.0	0.0

Source: U.S. Census Bureau, American Community Survey, 2016 Content Test
Note: Data are unweighted. Differences were not statistically tested.

Sex and relationship inconsistencies remain a problem affecting estimates of same-sex couples in particular, even after including the relationship/sex consistency check. Not only were inconsistent responses apparently higher in proportion among this group after the check, but upon receiving the sex and relationship check prompts, they also appear to be less likely to change responses. For example, a householder may identify another person as his or her *same-sex spouse* in the relationship question, but also report the spouse as having a different sex than his or her own, and confirm this response during the check. These couples show up in our analysis as having inconsistent responses after the check. In contrast, the majority of couples who received checks ended up as opposite-sex couples who did change their responses, resolving the inconsistencies. The results indicate that the automated check provides an opportunity for respondents to make corrections and improve the accuracy of the estimates while reducing error in the estimate of coupled households, particularly for same-sex couples.

There is evidence that fielding the test relationship question with a relationship/sex consistency check improves accuracy. Although few couples who responded via the internet mode received the check (only 34 cases, see Table 19), nearly 65 percent of these did change their responses. In other words, most respondents made a correction when given the opportunity to review and change responses. The majority of changes ended up as opposite-sex married couples. Had the original responses not been changed and been distributed across same-sex couple categories, the net proportion of these smaller couple types would have been greatly inflated. Although not shown, most couples who changed responses did not have a sex or relationship inconsistency by the end of the survey. Notably, same-sex couples who received the check did not change any of their responses. However, it is important to keep in mind that the number of these couples who received the check is too small to conclude that no couples initially reported as same sex will change their answers when given the opportunity via the automated check. Due to small sample size and the limited number of field representatives involved in this test, we do not show the results for the automated sex and relationship check in the CATI/CAPI mode.

6. CONCLUSIONS AND RECOMMENDATIONS

This test was designed to answer four research questions about the use of new response options in the relationship question:

1. Are item missing data rates lower in the test treatment than in the control treatment?
2. Does the distribution of people reported as a *spouse* or *partner* of the householder differ between control and test treatments?
3. Do the measures of response reliability (gross difference rate, index of inconsistency) differ between the test and control treatment?
4. Does the inclusion of an automated relationship/sex consistency check improve data quality?

The expansion of the *husband or wife* and *unmarried partner* categories to distinguish between opposite-sex and same-sex couples did not result in statistically different item nonresponse rates for the relationship question when comparing the overall results. This finding holds when examining item nonresponse rates for each mode individually, except for the internet mode. This result is not concerning though because the magnitude of the difference is small, only 0.2 percentage points.

Likewise, the inclusion of the new categories was not associated with a significant difference in the overall distribution of the response categories. The only difference that was found in the distribution involved the internet mode and the *missing* category. The magnitude of this difference, while statistically significant, is negligible. Finally, the distribution of categories was what we expected to see from using the 2015 NCT as a benchmark.

Similarly, the response reliability measures show no statistical differences overall in responses to the relationship question between test and control treatments. For only one category, *unmarried partner*, the reliability of the test version was significantly lower than the control version, but this finding did not affect the results for the overall IOI L-fold reliability metric of the relationship question indicating that the test version is as reliable as the control version of the question.

Collecting detailed data about relationships can improve editing procedures for demographic data after data collection as well as increase the accuracy of estimates of coupled households. The automated check in electronic instruments allows respondents to correct responses, improving data quality. In addition, cases that have inconsistent relationship and sex reports will be flagged in the editing process, which is not possible with the control version of the relationship question.[20]

While significant differences were found in some of the analyses that were performed, the magnitude of these differences should not affect data quality and do not outweigh the benefits

[20] Note that the CATI/CAPI version of the production ACS relationship question does include an automated confirmation question for relationship if the sex of the householder and spouse are reported to be the same. However, since the relationship question only has the category *spouse* we cannot currently check it against the reported sex of the householder and spouse to see if the reports are consistent.

that will be realized from the new editing procedures. The recommendation, based on this research, is to implement the test version of the relationship question. The final wording for the relationship question in the ACS will be determined in the decennial testing program though, which is performing similar analysis that involves similar changes to the question, including the relationship/sex consistency check for electronic instruments.

7. ACKNOWLEDGEMENTS

The 2016 ACS Content Test would not have been possible without the participation and assistance of many individuals from the Census Bureau and other agencies. Their contributions are sincerely appreciated and gratefully acknowledged.

* Census Bureau staff in the American Community Survey Office, Application Development and Services Division, Decennial Information Technology Division, Decennial Statistical Studies Division, Field Division, National Processing Center, Population Division, and Social, Economic, and Housing Statistics Division.

* Representatives from other agencies in the Federal statistical system serving on the Office of Management and Budget's Interagency Working Group for the ACS and the Topical Subcommittees formed by the Working Group for each topic tested on the 2016 ACS Content Test.

* Staff in the Office of Management and Budget's Statistical and Science Policy Office.

We would like to acknowledge Eli Poehler and Michael Risley for their contributions to this report.

8. REFERENCES

Biemer, P. (2011). *Latent Class Analysis of Survey Error*. Wiley, New York.

Dusch, G. and Meier, F. (2012). *2010 Census Content Reinterview Survey Evaluation Report*, U.S. Census Bureau, June 13, 2012. Retrieved May 17, 2016 from http://www.census.gov/2010census/pdf/2010_Census_Content_Reinterview_Survey_Evaluation_Report.pdf

Flanagan, P. (1996). *Survey Quality & Response Variance* (Unpublished Internal Document). U.S. Census Bureau. Demographic Statistical Methods Division. Quality Assurance and Evaluation Branch.

Flanagan, P. (2001). *Measurement Errors in Survey Response*. University of Maryland Baltimore County, Baltimore, Maryland.

Goerman, P., Meyers, M., and Simmons, D. (Forthcoming). "Usability Results for the 2014 Census Test in Spanish." *Survey Methodology Research Report Series*.

Goerman, P., Quiroz, M.., McAvinchey, G., Reed, L., and Rodriguez, S. (2014). "Census American Community Survey Spanish CAPI/CATI Instrument Testing Phase 1, Round 2 Final Report." *Survey Methodology Research Report Series*. SSM 2014/06. Available online http://www.census.gov/srd/papers/pdf/ssm2014-06.pdf.

Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure," Scandinavian Journal of Statistics, Vol. 6, No. 2: 65-70. Retrieved on January 31, 2017 from https://www.jstor.org/stable/4615733?seq=1#page_scan_tab_contents

Interagency Working Group on Measuring Relationships in Federal Household Surveys (2014). "Paper 1: Measuring Same-Sex Co-Residential Relationships." Statistical Policy Working Paper.

Kreider, R. M., Bates, N. and Lofquist, D. 2016. "Improving Measurement of Same-Sex Couple Households," Presentation at the AAPOR annual meetings in Austin, TX, May 13, 2016, available online at: https://www.census.gov/library/working-papers/2016/demo/AAPOR_Kreideretal.html

Meyers, M. Holliday, N., Lykke, L., García Trejo, Y., & L. Fernandez. (Forthcoming). "Cognitive & Usability Results from Spanish Pre-Testing of the 2015 National Content Test." *Survey Methodology Research Report Series*.

O'Connell, M., and Feliz, S. (2011). "Same-Sex Couple Households from the 2010 Census." SEHSD Working Paper, U.S. Census Bureau. Retrieved May 30, 2017 from https://www.census.gov/library/working-papers/2011/demo/SEHSD-WP2011-26.html

O'Connell, M., and Gooding. G. (2006). "The Use of First Names to Evaluate Reports of Gender and Its Effect on the Distribution of Married and Unmarried Couple Households." Paper presented at the Population Association of America, Los Angeles, CA.

Rao, J. N. K. and Scott, A. J. (1987). "On Simple Adjustments to Chi-Square Tests with Sample Survey Data," The Annal of Statistics, Vol. 15, No. 1, 385-397. Retrieved on January 31, 2017 from http://projecteuclid.org/euclid.aos/1176350273

Simmons, T., and O'Connell, M. (2003). *Married-Couple and Unmarried-Partner Households: 2000.* Census 2000 Special Reports. Retrieved on June 21, 2017 from https://www.census.gov/prod/2003pubs/censr-5.pdf

Seem, E. and Coombs, J. (2017). 2020 Research and Testing: 2015 National Content Test Relationship Question Experiment Analysis Report. U.S. Census Bureau, Suitland, Maryland. Retrieved on March 7, 2017 from https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/final-analysis/2015nct-relationship-question-experiment.html

Stapleton, M., and Steiger, D. (2015). *Cognitive Testing of the 2016 American Community Survey Content Test Items: Summary Report for Round 1 and Round 2 Interviews.* Westat, Rockville, Maryland, January 2015.

Steiger, D., Anderson, J., Folz, J., Leonard, M., and Stapleton, M. (2015). *Cognitive Testing of the 2016 American Community Survey Content Test Items: Briefing Report for Round 3 Interviews.* Westat, Rockville, Maryland, June, 2015.

U.S. Census Bureau. (2014). *American Community Survey Design and Methodology (January 2014).* Retrieved February 1, 2017 from http://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html

U.S. Census Bureau (2016). *2015 Planning Database Tract Data* [Data file]. Retrieved on January 31, 2017 from http://www.census.gov/research/data/planning_database/2015/

Appendix A: Unit Response Rates Supplemental Table

Table A-1. Unit Response Rates by Designated High (HRA) and Low (LRA) Response Areas

Mode	Test Interviews	Test Percent	Control Interviews	Control Percent	Test minus Control	P-Value
Total Response	19,400		19,455			
HRA	7,556	94.3 (0.4)	7,608	94.5 (0.3)	-0.2 (0.6)	0.72
LRA	11,844	91.5 (0.3)	11,847	91.0 (0.3)	0.5 (0.5)	0.29
Difference		2.7 (0.5)		3.5 (0.5)	-0.7 (0.7)	0.33
Self-Response	13,131		13,284			
HRA	6,201	59.7 (0.7)	6,272	60.6 (0.7)	-0.9 (0.9)	0.31
LRA	6,930	33.2 (0.4)	7,012	33.6 (0.4)	-0.4 (0.6)	0.55
Difference		26.5 (0.8)		27.0 (0.8)	-0.5 (1.2)	0.66
Internet	8,168		8,112			
HRA	4,119	39.6 (0.6)	4,048	39.1 (0.6)	0.5 (0.8)	0.51
LRA	4,049	19.4 (0.3)	4,064	19.5 (0.3)	0.1 (0.4)	0.87
Difference		20.2 (0.6)		19.6 (0.7)	0.6 (0.9)	0.52
Mail	4,963		5,172			
HRA	2,082	20.0 (0.4)	2,224	21.5 (0.4)	-1.5 (0.6)	0.02*
LRA	2,881	13.8 (0.3)	2,948	14.1 (0.3)	-0.3 (0.4)	0.43
Difference		6.2 (0.5)		7.4 (0.4)	-1.1 (0.7)	0.11
CATI	872		880			
HRA	296	9.0 (0.5)	301	9.6 (0.6)	-0.6 (0.8)	0.44
LRA	576	7.9 (0.4)	579	8.0 (0.3)	-0.1 (0.5)	0.85
Difference		1.1 (0.6)		1.6 (0.7)	-0.5 (0.9)	0.58
CAPI	5,397		5,291			
HRA	1,059	82.2 (1.0)	1,035	82.7 (0.9)	-0.5 (1.3)	0.69
LRA	4,338	85.8 (0.5)	4,256	85.0 (0.4)	0.8 (0.7)	0.23
Difference		-3.7 (1.1)		-2.3 (1.0)	-1.3 (1.5)	0.36

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Standard errors are in parentheses. Minor additive discrepancies are due to rounding. P-values with an asterisk (*) indicate a significant difference based on a two-tailed t-test at the α=0.1 level. The weighted response rates account for initial sample design as well as CAPI subsampling.