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

September 12, 2017

2017 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES # ACS17-RER-10

MEMORANDUM FOR	Victoria Velkoff Chief, American Community Survey Office
From:	David Waddington Chief, Social, Economic, and Housing Statistics Division (SEHSD)
Prepared by:	Kirby G. Posey Social, Economic, and Housing Statistics Division (SEHSD)
Subject:	2016 American Community Survey Content Test Evaluation Report: Retirement, Survivor, and Disability Income

Attached is the final report for the 2016 American Community Survey (ACS) Content Test for Retirement, Survivor and Disability Income. This report describes the results of the test for a revised version of the Retirement Income question.

If you have any questions about this report, please contact Jonathan Rothbaum at 301-763-9681 or Kirby Posey at 301-763-5548.

Attachment

cc:
Agnes Kee (ACSO)
Jennifer Ortman (ACSO)
David Raglin (ACSO)
Patrick Cantwell (DSSD)
Asaph Young Chun (DSSD)
Sarah Heimel (DSSD)
Elizabeth Poehler (DSSD)
Anthony Tersine (DSSD)
Trudi Renwick (SEHSD)
Jonathan Rothbaum (SEHSD)
Nicole Scanniello (SEHSD)

Intentionally Blank

September 12, 2017

2016 American Community Survey Content Test Evaluation Report: Retirement, Survivor, and Disability Income

FINAL DRAFT

Kirby G. Posey
Social, Economic, and Housing Statistics Division

Andrew W. Roberts
Population Division

Intentionally Blank

TABLE OF CONTENTS

List of Tables

List of Figures

Intentionally Blank

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 Content Test interview, a Content Follow-Up telephone reinterview was conducted to measure response error.

Retirement, Survivor, and Disability Income

Over the last 40 years, defined contribution retirement plans have become increasingly common while defined benefits plans (such as pensions) have become less so (Butrica et al., 2009).[1] Federal surveys have lagged in addressing these newer forms of retirement income, thus underreport retirement income (Munnell and Chen, 2014). Changes to the Retirement, Survivor and Disability Income question were tested in the 2016 ACS Content Test with the goals of improving income reporting, reducing item missing data rates, reducing reporting errors, and updating questions on retirement income and the income generated from retirement accounts and all other assets in order to better measure retirement income data. The current version of the question on the ACS makes no reference to these specific types of retirement income, asking only about "*retirement, survivor, or disability pensions*." For this test, the question was expanded to ask about "*retirement income, pensions, survivor or disability income*." In addition, the instructions that accompany the question were expanded to note that income from "*a previous employer or union, or any regular withdrawals or distributions from IRA, Roth IRA, 401(k), 403(b) or other accounts specifically designed for retirement*" should be included. This report summarizes the test of a new version of the question on retirement, survivor, and disability income.

[1] Defined contribution plans are a type of retirement plan in which the employer, employee or both make regular contributions to an IRA, ROTH IRA, 401(k), etc.

Research Questions and Results

This research was guided by several research questions concerning comparisons to Current Population Survey Annual Social and Economic Supplement (CPS ASEC) data, missing data rates, and differences in the reports of recipiency and income amount by treatment and other characteristics.

The results showed several positive changes:

- The overall distribution of retirement, survivor, and disability income for the test and control versions was similar to that of the CPS ASEC.
- The number of eligible respondents receiving retirement, survivor, and disability income was significantly higher in the test treatment than the control.
- The overall retirement, survivor, and disability aggregate income amount was significantly higher in the test treatment than the control.
- The income distribution for the test and control treatments was significantly different. There was a significantly higher share of responses in the income category '$1 to $2,499' in the test treatment and a significantly lower share of responses in the category '$40,000 to $64,999' in the test treatment. Note that in every income category, there were more respondents in the test treatment than control.
- The absolute value of the difference between the proportion of positive responses in the original interview and the content follow-up for combined retirement, survivor, and disability income recipiency was significantly smaller for the test treatment than for the control treatment.

Conclusion

There were several positive changes indicating that we will better measure retirement income by implementing these revisions.

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 Retirement, Survivor, and Disability Income topic.

1.1 Justification for Inclusion of Retirement, Survivor, and Disability Income in the Content Test

Questions collecting information on a person's income appeared on the census forms from 1940 through 2000 and on the American Community Survey questionnaires from 2005 forward. The census specifically asked about retirement, survivor, and disability income for the first time in 1990. The current version of the question has been in place since 1990.

Federal surveys, including the ACS, have lagged in addressing newer forms of retirement income, namely defined contribution plans resulting in the underreporting of retirement income.

Positive results of a question redesign on the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) led to testing the Retirement, Survivor and Disability Income question on the 2016 ACS Content Test. The income questions on the CPS ASEC were redesigned in 2014 with the goals of improving income reporting, increasing response rates, reducing response errors by taking better advantage of an automated questionnaire environment, and updating questions on retirement income and the income generated from retirement accounts

and all other assets. To capture more relevant retirement income, the redesigned CPS ASEC specifically asks if anyone in the household has a pension, and then if anyone has a retirement account (such as a 401(k), 403(b), IRA, or other account designed specifically for retirement savings). In contrast, the traditional CPS ASEC asked one broad question on receipt of pension and retirement income. If the respondent has a retirement account, the redesigned CPS ASEC instrument asks the respondent to identify the specific type of account. The instrument proceeds to inquire if there was a withdrawal or distribution from that retirement account. For recipients over 70 years old, the question text changes to add, "including distributions you may have been required to take." To make sure the value of the withdrawal correctly counts as household income, a follow-up question asks if the money was "rolled over" or reinvested to another account.

CPS ASEC test results showed total recipiency, mean income, and aggregate income was higher in the redesigned CPS ASEC. The change in the question resulted in a 419.5 percent increase in people that received income from an IRA, Keogh (retirement plan for self-employed people and small businesses), and/or 401(k) and a 230 percent increase in aggregate income from these retirement accounts (Semega & Welniak, 2015).

1.2 Question Development

Initial versions of the new and revised questions were proposed by federal agencies participating in the U.S. Office of Management and Budget (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.[2] 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 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 Census Bureau contracted with Westat to conduct three rounds of cognitive testing. The results of the first two rounds of cognitive testing informed decisions on specific revisions to the proposed content for the stateside Content Test (Stapleton and Steiger, 2015). In the first round,

[2] 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.

208 cognitive interviews were conducted in English and Spanish and in two modes (self-administered on paper and interviewer-administered on paper). In the second round of testing, 120 cognitive interviews were conducted for one version of each of the tested questions, in English and Spanish, using the same modes as in the first round.

A third round of cognitive testing involved only the Puerto Rico Community Survey (PRCS) and Group Quarters (GQ) versions of the questionnaire (Steiger, Anderson, Folz, Leonard, & Stapleton, 2015). Cognitive interviews in Puerto Rico were conducted in Spanish; GQ cognitive interviews were conducted in English. The third round of cognitive testing was carried out to assess the revised versions of the questions in Spanish and identify any issues with questionnaire wording unique to Puerto Rico and GQ populations.[3] 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.[4]

Two paper versions of the Retirement, Survivor, and Disability Income question were tested in Round 1 of cognitive testing.

Paper Version 1:

> Retirement, survivor, disability payments, or retirement account withdrawals or distributions. *DO NOT include Social Security or amounts rolled over into other retirement accounts. Retirement accounts include employer plans and IRA, Roth IRA, 401(k), 403(b), SEP, KEOGH, SIMPLE accounts.*
>
> Yes
> No
>
> *TOTAL AMOUNT for past 12 months* _____

Paper Version 2:

> Retirement, survivor, disability payments, or retirement account withdrawals or distributions. *DO NOT include amounts rolled over into other retirement accounts.*
>
> Yes
> No
>
> *TOTAL AMOUNT for past 12 months* _____

[3] Note that the field testing of the content was not conducted in Puerto Rico or in GQs. See the Methodology section for more information.

[4] 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.

Only one version of the Computer Assisted Telephone Interview (CATI) / Computer Assisted Personal Interview (CAPI) version of the question was tested in Round 1 but the question was asked in two separate parts:

CATI/CAPI Part 1:

> Did <Name/you> receive any survivor or disability income DURING THE PAST 12 MONTHS?
>
> Yes
> No
>
> *What was the amount?*
> (Do not include Social Security)

CATI/CAPI Part 2:

> Did (Name/you) receive any retirement income from a previous employer or union or income from retirement accounts such as a 401(k), 403(b), IRA, or other accounts designed specifically for retirement savings DURING THE PAST 12 MONTHS? DO NOT include amounts rolled over into other retirement accounts.
>
> Yes
> No
>
> *What was the amount?*

Responses to these questions in Round 1 of cognitive testing served as the basis for changes to the current question content for Round 2. Round 2 was the final wording tested. That wording is provided below in Section 1.3.

Finally, the briefing report for Round 3 interviews recommended retaining the wording in Round 2, the current wording on the content test. Most respondents were able to easily and accurately answer the revised questions about their retirement income and appeared to interpret the instructions correctly.

1.3 Question Content

The Retirement, Survivor, and Disability Income question was revised to elicit responses for certain types of income that are currently not being captured or being partially captured, such as 401(k) accounts and Roth IRAs.

Control and test versions of each question are shown as they appeared on the paper questionnaire (see Figure 1 and Figure 2). For control, the internet, CATI, and CAPI versions of this question are very similar to the paper version.

For the test treatment, the question is split into two parts in the CATI/CAPI and internet instruments. One part asks about survivor and disability income, and a separate part asks about retirement income (pensions, retirement accounts such as 401(k), 403(b), IRA, etc.).

Figure 1. Paper Control Version

Figure 2. Paper Test Version

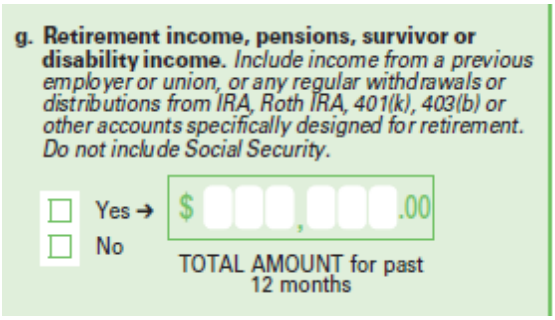

Figure 3. Internet/CATI/CAPI Test Version

Did <(Name)/you> receive any survivor or disability income DURING THE PAST 12 MONTHS? **What was the amount?** (Do not include Social Security)

Did <(Name)/you> receive a pension or any retirement income from a previous employer or union, or any regular withdrawals or distributions from retirement accounts such as a 401(k), 403(b), IRA, Roth IRA, or other accounts designed specifically for retirement DURING THE PAST 12 MONTHS? **What was the amount?** (Do not include Social Security)

1.4 Research Questions

The following research questions were formulated to guide the analyses of the Retirement, Survivor, and Disability Income question. The analyses assess 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 affect the resulting estimates.

The research questions were evaluated separately for each response mode and by self-response stratum (high/low). The questions are as follows:

1. *How do the proportions in each treatment compare with published CPS ASEC data?*
2. *Is the item missing data rate for combined retirement, survivor, and disability income recipiency lower for the test treatment than for the control treatment?*
3. *Is the item missing data rate for combined retirement, survivor, and disability income amount lower for the test treatment than for the control treatment?*
4. *Recipiency: Are there more eligible persons reported as receiving retirement, survivor, and disability income in the test treatment than in the control treatment?*
5. *Amount: Is the estimated overall aggregate income of combined retirement, survivor, and disability income for eligible persons higher for the test treatment than for the control treatment?*
6. *Amount: How do the distributions of persons among the income ranges defined in Table 1 compare between the control and test treatments? Are they significantly different? Are proportion estimates for any single range significantly different between the control and test treatments?*
7. *Is the absolute value of the net difference rate for combined retirement, survivor, and disability income recipiency smaller for the test treatment than for the control treatment?*
8. *For each of the ranges defined in Table 1, is the absolute value of the net difference rate for combined retirement, survivor, and disability income smaller for the test treatment than for the control treatment?*

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 Content Test sample universe did not include GQs, nor did it include housing units in Alaska, Hawaii, or

Puerto Rico.[5] The sample design for the 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 Content Test, the development of workload estimates for CATI and 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 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-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 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

[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).

both control and test cases. CAPI interviewers were assigned Content Test cases based on their geographic proximity to the cases and therefore could also conduct both control and test cases.

The ACS Content Test's data collection protocol differed from the production ACS in a few significant ways. The 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 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 Content Test had Spanish-language automated instruments; however, there were no paper versions of the Content Test questionnaires in Spanish.[10] Any case in the Content Test sample that completed a Spanish-language internet, CATI, or CAPI response was included in analysis. However, if a case sampled for the Content Test called TQA to complete an interview in Spanish or any other language, the production interview was conducted and the response was excluded from the 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 Content Test. CFU interviews for the Content Test were conducted in either Spanish or English. The practical need to limit the language response options for Content Test respondents is a limitation to the research, as some respondents self-selected out of the test.

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

[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 Content Test.

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

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 version of the questions in CFU. The only exception was for retirement, survivor, and disability income for which a different set of questions was asked in CFU. In the CFU operation, respondents were asked the series of questions about retirement income used in the 2016 CPS ASEC. For CFU retirement questions, see Appendix B.

The questions about survivor and disability income (asked separately from the retirement income questions) matched the ones used in the CATI ACS Content Test instrument for the test version. We consider the CPS ASEC questions to result in more accurate measures of both recipiency and amount for retirement income (Semega & Welniak, 2013). Therefore, by asking the series of CPS ASEC questions in the CFU for both treatments, we could estimate bias in each treatment by comparing each estimate derived from the Content Test original interview data with the corresponding estimate derived from the CFU data.

2.4 Analysis Metrics

This section describes the metrics used to assess the revised version of the 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 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 collapsing response categories or modes together or calculating a person's age based on his or her date of birth.

All estimates from the 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.

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

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. In select tables, we utilize the Holm-Bonferroni method to adjust for multiple comparisons (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.[13] 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 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.

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 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,

[13] 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.

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 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.* The topics of *race, Hispanic origin,* and *relationship* are also typically used for demographic analysis; however, those questions were modified as part of the 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.[14]

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 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.

To calculate item missing data rates for retirement, survivor, and disability recipiency, the universe is all persons aged 15 or older. Note that for mail and internet mode responses, if an amount was entered, we considered the response to the recipiency question to be "Yes" even if "No" was checked or if neither box was checked. Otherwise, a valid response was either "Yes"

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

or “No.” “Don’t Know” and “Refused” responses were considered missing. In addition, for the test treatment internet, CATI, and CAPI version, where survivor and disability income are asked about separately from retirement income, a “Yes” to the recipiency question for either type of income was counted as a “Yes” for aggregate recipiency.

To calculate item missing data rates for the amount, the universe is all persons aged 15 or older for whom there is a “Yes” response to the recipiency question (or an implied “Yes” for internet or mail). A valid amount response is a number greater than or equal to zero; “Don’t Know” and “Refused” are considered missing. For the test internet, CATI, and CAPI versions, a valid amount in either part of the question is counted as a valid amount for the combined sum.

We compare the *recipiency* and *amount* item missing data rates for the control treatment to the corresponding rates for the test treatment and determine whether a difference between versions is statistically significant using a one-tailed t-test. We expect the missing data rate to remain unchanged or to be higher for the control treatment than the test treatment.

2.4.3 Response Distributions

Comparing the response distributions between the control version of a question and the test version of a question allowed 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 and t-tests for single categories when the corresponding distributions were found to be statistically different.

We calculated the estimated proportion of persons receiving any retirement, survivor, and disability income (recipiency) as the weighted sum of those with a “Yes” response (including mail or internet responses converted to “Yes” because an amount is reported) divided by the weighted sum of all persons in the universe. The universe consists of all persons aged 15 or older. Note that for the test treatment, a “Yes” to either of the two parts in internet, CATI, or CAPI corresponds to a single “Yes” in the versions of the questions where survivor and disability are not asked about separately. For the income distributions, we tested all differences between control and test for statistical significance using two-tailed t-tests. A one-tailed test was used for recipiency and aggregate income.

Proportion estimates were calculated as:

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

Retirement income amounts were compared by categories as defined in Table 1.

Table 1. Retirement Income Ranges for Analysis

Range Category Number	Retirement Income Amount Range
1	$1 - $2,499
2	$2,500 - $19,999
3	$20,000 - $39,999
4	$40,000 – $64,999
5	$65,000 or more

2.4.4 Benchmarks

For the topic of Retirement, Survivor, and Disability Income, we compared data from both control and test treatments to information from the CPS ASEC. This comparison allows us to tell whether our results differ from another reliable source.

We compared combined retirement, survivor, and disability income recipiency (proportion of persons in the universe who received retirement, survivor, or disability income) in control and test versions with recipiency reported in the 2016 CPS ASEC. We also compared amount distributions (using specified dollar amount ranges as categories, see Table 1) from both treatments with the amount distribution from the 2016 CPS ASEC. These comparisons only determine whether the Content Test estimates are reasonably close to the benchmark estimates; we did not perform significance testing of differences.

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. Asking a question that is believed to produce "truth" from the same respondent (or housing unit) allows us to measure response bias.

For retirement income, the use of the CPS ASEC question in CFU is designed to elicit more accurate responses than in the original interview. Therefore, we consider the CFU responses for Retirement Income to be "true." Even though we believe the CPS ASEC series of retirement income questions will produce "true" results, this battery of questions is detailed and lengthy and therefore not appropriate for the ACS questionnaire.

We measured response bias by comparing valid responses to the CFU reinterview with valid responses to the original interview.[15] The Census Bureau has frequently used content reinterview surveys to measure response error for large demographic data collection efforts, including the

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

2010 ACS Content Test, and the 1990, 2000, and 2010 decennial censuses (Dusch & Meier, 2012).

We measured response bias by calculating the net difference rate (NDR), which is the difference between the original interview proportion of positive responses ("Yes", or in the category of interest) and the CFU proportion of positive responses. The NDR is calculated using the following table and formula.

Table 2. Interview and Reinterview Counts for Each Response Category Used for Calculating Net Difference Rates

	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.

$$NDR = \left(\frac{c - b}{n}\right) \times 100$$

Note the NDR can be negative, zero, or positive. If the NDR is significantly negative, this indicates that the Content Test original interview version of the question tends to result in an underestimate of the true proportion in a category. Conversely, if the NDR is significantly positive, the original interview question tends to result in an overestimate of the true proportion. If the NDR is zero (or not significantly different from zero), this is an indication that the original interview question results in an unbiased estimate of the true proportion. For this analysis, our interest is not in the direction of the NDR from zero, but in the magnitude of difference from zero (that is, which version is further away from the truth as assessed in CFU). For this reason, the response error research questions focus on the absolute value of the NDR in each treatment.

When comparing original interview results with CFU results for a given person, it is necessary to (a) determine CFU recipiency by combining the results of the survivor and disability income recipiency question with the results of the various CFU retirement income Recipiency questions; and (b) sum the survivor and disability income amount, pension amount(s), annuity amount, and

the retirement account withdrawal amounts. Note that "roll-over" amounts are not included in the CFU combined survivor, disability, and retirement income amounts.

We tested for significance of each difference using a one-tailed t-test.

2.4.6 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 DECISION CRITERIA

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. The following table identifies the research questions and associated metrics in priority order.

Table 3. Decision Criteria

Research Questions	Decision Criteria, in order of priority
5	The overall aggregate income amount for retirement, survivor, and disability income should be higher in the test version than in the control version.
4	The number of eligible persons receiving retirement, survivor, and disability income should be higher in the test version than in the control version.
2, 3	The item missing data rates for the test version should be lower than the control version.
7	The absolute value of the net difference rate for retirement, survivor, and disability income recipiency combined should be lower in the test version than in control.
8	For each of the ranges defined in Table 1, the absolute value of the net difference rate for retirement, survivor, and disability income combined should be lower in the test version than in control.
6	The income distribution (defined in Table 1) was tested between the control and test treatments for statistical difference. The income distributions should be statistically different between the control and test treatments.
1	The proportions of persons in the income ranges (defined in Table 1) were compared between CPS ASEC data and both the control and test treatments. These proportions should be reasonably close. No formal statistical test will be performed.

Note: Comparisons, with the exception of Research Question 1, are also made by mode of data collection and response stratum.

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.

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 relationship status, 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, by mail, or by 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 is statistically significant (p-value=0.06).

Although we reinterviewed the same person who responded in the original interview when possible, we interviewed a different member of the household in the CFU for 7.5 percent (standard error of 0.4) of the CFU cases for the control treatment and 8.4 percent (standard error of 0.5) of the CFU cases for the test treatment.[16] The difference between the test and control treatments is not statistically significant (p-value=0.26). This means that differences in results between the original interview and the CFU for these cases could be due in part to having different people answering the questions. However, those changes were not statistically significant between the control and test treatments and should not impact the conclusions drawn from the reinterview.

The 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 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 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 Content Test cases to be mailed production ACS questionnaires instead of 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.

[16] This is based on comparing the first name of the respondent between the original interview and the CFU interview. Due to a data issue, we were not able to use the full name to compare.

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.

For both treatments in the CFU, we added a CPS ASEC-like series of questions about retirement income. We asked this series of questions within the battery of standard ACS questions. One of the standard income questions, Social Security Income, asks if the person received social security or railroad retirement (RR) income. The CPS ASEC question starts by asking if the person received any pension income, and if so asks which type of pension(s) they received. The seventh possible type of pension listed is RR income. Therefore, it is possible for a person to report the same RR income twice. It was not feasible to change the CFU instrument or data processing to account for this scenario, therefore, we developed an edit to check for possible double reporting of RR income and corrected for it in the analysis.

5 RESEARCH QUESTIONS AND RESULTS

This section presents the results from the analyses of the 2016 ACS Content Test data for the Retirement, Survivor, and Disability Income 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, followed by a brief summary of the results and corresponding data.

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 Content Test 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 4 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 4. 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 5 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 (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.

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

Table 5. 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 6 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. These 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 6. 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

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

response distribution calculations can be found in Table 7. 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 in Table 7 is statistically significant.

Table 7. 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 was done at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method.

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 8 and Table 9 below.

Table 8. Comparison of Average Household Size

Topic	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 9. 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. Significance was tested based on a two-tailed t-test at the α=0.1 level.

The 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 detectable 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

This section addresses research questions #2-3:

2. *Is the item missing data rate for combined retirement, survivor, and disability income recipiency lower for the test treatment than for the control treatment? Analyze by mode and response stratum.*

3. *Is the item missing data rate for combined retirement, survivor, and disability income amount lower for the test treatment than for the control treatment? Analyze by mode and response stratum.*

Table 10 shows overall item missing data rates for recipiency were not statistically lower in the test treatment than the control treatment. Similarly, Table 11 shows overall item missing data rates for the retirement, survivor and disability income amount are not statistically lower in the test treatment than the control treatment. Further, Table 10 shows item missing data rates for recipiency across the different modes of interview were not statistically lower in the test than the control treatment. Table 11, however, shows the item missing data rate for amount was lower in the mail mode for the test treatment than for the control treatment.

Table A2 and Table A3 in Appendix A show item missing data rates for test and control treatments by response stratum for recipiency and amount respectively. The item missing data rate for recipiency in the low response stratum was statistically lower in the test treatment than control. Item missing data rates for amount were not statistically lower for the test than control treatments for either the high or low response strata.

Table 10. Retirement, Survivor, and Disability Income Recipiency Item Missing Data Rates for Test and Control Treatments, Overall and by Interview Mode

Mode	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	P-Value
Overall	35,248	11.4 (0.3)	35,341	11.5 (0.4)	-0.1 (0.4)	0.38
Internet	16,864	11.9 (0.5)	16,677	11.9 (0.4)	0.1 (0.6)	0.55
Mail	8,824	19.0 (0.6)	9,293	18.7 (0.7)	0.3 (0.9)	0.61
CATI	1,731	6.2 (0.9)	1,679	6.9 (1.3)	-0.7 (1.8)	0.34
CAPI	7,829	7.0 (0.6)	7,692	7.1 (0.7)	-0.1 (0.8)	0.46

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 one-tailed t-test (test < control) at the α=0.1 level.

Table 11. Retirement, Survivor, and Disability Income Amount Item Missing Data Rates for Test and Control Treatments, Overall and by Interview Mode

Item	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	P-Value
Overall	4,679	8.9 (0.7)	3,518	9.4 (1.0)	-0.5 (1.2)	0.35
Internet	2,286	5.0 (0.6)	1,564	4.0 (0.7)	1.1 (0.9)	0.88
Mail	1,640	3.5 (0.6)	1,361	5.5 (0.8)	-2.0 (1.0)	0.02*
CATI	348	22.1 (3.0)	288	18.3 (2.6)	3.8 (3.8)	0.84
CAPI	405	25.7 (3.5)	305	31.9 (4.9)	-6.2 (5.8)	0.14

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 one-tailed t-test (test < control) at the α=0.1 level.

5.3 Response Distributions

This section addresses research questions #4-6:

4. *Recipiency: Are there more eligible persons reported as receiving retirement, survivor, or disability income in the test treatment than in the control treatment? Analyze by mode and response stratum.*

5. *Amount: Is the estimated overall aggregate income of combined retirement, survivor, and disability income for eligible persons higher for the test treatment than for the control treatment? Analyze by mode and response stratum.*

6. *Amount: How do the distributions of persons among the income ranges defined in Table 1 compare between the control and test treatments? Are they significantly different? Are proportion estimates for any single range significantly different between the control and test treatments? Analyze by mode and response stratum.*

Table 12 shows the changes to the Retirement, Survivor, and Disability Income question significantly raised the percentage of eligible persons who reported receiving this source of income (14.5 percent). The control treatment estimate was 10.6 percent. Similarly, all modes of data collection with the exception of CATI showed significantly higher percentages of persons receiving retirement, survivor, and disability income in the test treatment. Recipiency rates for both the high response stratum and low response stratum were also significantly higher in the test treatment than in the control treatment. (See Table A4 in Appendix A.)

Note that a significant increase in the number of eligible persons receiving retirement, survivor, and disability income ranked high in the decision criteria shown in Table 3.

Table 12. Recipiency Rates for Control and Test Treatments

Mode	Test Sample Size	Test Percent	Control Sample Size	Control percent	Test minus Control	P-Value
Overall	30,446	14.5 (0.3)	30,358	10.6 (0.3)	3.9 (0.4)	<0.01*
Internet	14,620	16.3 (0.4)	14,349	11.2 (0.3)	5.2 (0.5)	<0.01*
Mail	6,901	25.1 (0.7)	7,270	19.9 (0.6)	5.3 (0.9)	<0.01*
CATI	1,581	22.1 (1.3)	1,568	20.7 (1.6)	1.4 (1.9)	0.23
CAPI	7,344	6.6 (0.5)	7,171	4.2 (0.4)	2.5 (0.6)	<0.01*

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 one-tailed t-test (test > control) at the α=0.1 level.

Table 13 shows aggregate retirement, survivor, and disability income was higher in the test treatment than in the control treatment. Note that this finding ranked highest among the decision criteria shown in Table 3. Aggregate retirement, survivor, and disability income was about 32 percent higher in the test treatment. Aggregate retirement, survivor, and disability income was significantly higher in all modes of data collection except CATI. The aggregate was similarly higher in the test treatment than control treatment for both the high response stratum and low response stratum. (See Table A5 in Appendix A).

Aggregate income from all sources of income asked in the ACS was compared between the control and test treatments. Interestingly, when respondents reported their overall income from all sources, the aggregate showed no statistically significant increase in the test treatment. However, aggregate wages and salary income was significantly lower in the test treatment compared to control as well as aggregate interest, dividend, and net rental income. This indicates respondents were likely including retirement, survivor, and disability income in these other income sources in the control treatment (See Table A11 in Appendix A).

Table 13. Aggregate Retirement, Survivor, and Disability Income for Control and Test Treatments

Mode	Test Aggregate	Control Aggregate	Test minus Control	P-Value
Overall	564,216 (22,228)	425,989 (16,450)	138,227 (24,628)	<0.01*
Internet	301,378 (14,303)	218,834 (9,725)	82,543 (17,288)	<0.01****
Mail	191,212 (13,435)	148,173 (10,781)	43,039 (15,431)	<0.01*
CATI	17,983 (2,031)	19,977 (2,847)	-1,994 (3,330)	0.73
CAPI	53,643 (6,928)	39,005 (7,829)	14,638 (8,959)	0.05****

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Aggregates and standard errors are shown in millions of dollars. 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 one-tailed t-test (test > control) at the α=0.1 level.

Tables 14 to 18 show the distribution of reported retirement, survivor, and disability income amount. Chi-squared tests were used to compare the distributions in each table, followed by two-tailed t-tests on each row in a table if the chi-square statistic was significant. When t-tests were conducted, the p-values were adjusted using the Holm-Bonferroni procedure for multiple comparisons.

Table 14 shows there was a significant difference in the distributions of retirement, survivor, and disability income between test and control treatment. The t-tests identified a higher percentage of persons reporting retirement, survivor, and disability income amounts of $1 to $2,499 in the test treatment and a lower percentage of persons reporting amounts of $40,000 to $64,999.

Since the lowest income category, less than $2,500, was higher in the test treatment, it could indicate that respondents are reporting something other than annual amounts, such as monthly or bi-weekly amounts. If these respondents are over 70 years of age, this is likely the case. In production, the income editing process will attempt to make this determination and adjust accordingly.

Similarly, the income distribution for the high response stratum was significantly different between the two treatments. The category "$40,000 to $64,999" was significantly lower in the test treatment than the control treatment. The income distribution for the low response stratum was not significantly different between the test and control treatments. (See Tables A6 and A7 in Appendix A).

Table 14. Retirement, Survivor, and Disability Income Distribution for Control and Test Treatments for All Modes of Data Collection

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	Adjusted P-Value
$1 to $2,499	735	15.9 (0.7)	473	13.1 (0.7)	2.8 (1.2)	0.06*
$2,500 to $19,999	2,015	46.3 (1.0)	1,496	45.5 (1.3)	0.8 (1.6)	1.00
$20,000 to $39,999	889	21.3 (0.9)	737	22.8 (1.0)	-1.5 (1.2)	0.60
$40,000 to $64,999	407	10.1 (0.5)	363	12.6 (0.8)	-2.5 (0.9)	0.03*
$65,000 or more	250	6.4 (0.5)	163	6.0 (0.6)	0.4 (0.7)	1.00
Total	4,296	100.0	3,232	100.0		

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 13.2$, p-value = 0.01. 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. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method.

Table 15 shows the income distribution of cases using the internet mode of data collection. Internet was the only mode of interview to have a significant chi-squared test result and showed the same categories to be significantly different as the overall distribution.

Table 15. Retirement, Survivor, and Disability Income Distribution for Control and Test Treatments for Internet Mode of Data Collection

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	Adjusted P-Value
$1 to $2,499	387	16.3 (1.0)	201	12.5 (1.0)	3.8 (1.6)	0.06*
$2,500 to $19,999	930	41.9 (1.4)	622	39.6 (1.6)	2.3 (2.4)	0.66
$20,000 to $39,999	463	22.2 (1.0)	368	24.5 (1.4)	-2.3 (1.7)	0.56
$40,000 to $64,999	229	11.5 (0.7)	208	15.8 (1.2)	-4.2 (1.3)	<0.01*
$65,000 or more	155	8.1 (0.7)	97	7.7 (0.7)	0.4 (1.0)	0.71
Total	2,164	100.0	1,496	100.0		

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 15.9$, p-value < 0.01. 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. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method.

Tables 16 through 18 below show no statistical difference in the income distribution between the test treatment and control treatment for mail, CATI, or CAPI.

Table 16. Retirement, Survivor, and Disability Income Distribution for Control and Test Treatments for Mail Mode of Data Collection

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent
$1 to $2,499	262	15.5 (1.2)	215	15.3 (1.3)
$2,500 to $19,999	752	46.7 (1.6)	620	47.7 (1.7)
$20,000 to $39,999	328	21.3 (1.4)	277	22.2 (1.3)
$40,000 to $64,999	148	10.6 (0.9)	120	10.6 (1.2)
$65,000 or more	82	6.0 (0.7)	53	4.3 (0.7)
Total	1,572	100.0	1,285	100.0

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 2.9$, p-value = 0.58. Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance testing was done at the α=0.1 level. The chi-square test showed that there was no significant difference in the distribution, therefore no t-tests were conducted for the individual category comparisons.

Table 17. Retirement, Survivor, and Disability Income Distribution for Control and Test Treatments for CATI Mode of Data Collection

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent
$1 to $2,499	42	11.2 (2.3)	28	9.1 (2.0)
$2,500 to $19,999	157	60.8 (4.2)	122	55.5 (4.4)
$20,000 to $39,999	49	18.9 (3.4)	55	23.7 (3.9)
$40,000 to $64,999	17	5.7 (2.0)	22	8.9 (2.2)
$65,000 or more	5	3.4 (1.6)	5	2.9 (1.4)
Total	270	100.0	232	100.0

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 2.4$, p-value = 0.67. Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance testing was done at the α=0.1 level. The chi-square test showed that there was no significant difference in the distribution, therefore no t-tests were conducted for the individual category comparisons.

Table 18. Retirement, Survivor, and Disability Income Distribution for Control and Test Treatments for CAPI Mode of Data Collection

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent
$1 to $2,499	44	16.8 (3.1)	29	10.2 (2.9)
$2,500 to $19,999	176	56.7 (3.9)	132	58.0 (6.2)
$20,000 to $39,999	49	19.0 (3.0)	37	17.6 (3.3)
$40,000 to $64,999	13	5.6 (1.9)	13	8.4 (3.0)
$65,000 or more	8	1.9 (1.0)	8	5.8 (2.9)
Total	290	100.0	219	100.0

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 6.0$, p-value = 0.20. Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance testing was done at the α=0.1 level. The chi-square test showed that there was no significant difference in the distribution, therefore no t-tests were conducted for the individual category comparisons.

5.4 Benchmarks

This section addresses research question #1 *How do the proportions in each treatment compare with published CPS ASEC data?*

Table 19 shows response distributions from both treatments with the response distribution from the 2016 CPS ASEC. These comparisons will only be used to determine whether the Content Test estimates are reasonably close to the benchmark estimates; we did not perform significance testing of differences. Formal statistical comparisons were not made because the Content Test data were not edited or imputed, adjusted for nonresponse, nor raked to known population totals.

Table 19. Response Distributions for the Current Population Survey (CPS) – Annual Social and Economic Supplement (ASEC) and Content Test Treatments

Category	CPS ASEC Percent	Test Percent (n=4,296)	Control Percent (n=3,232)
$1 to $2,499	14.0	15.9	13.1
$2,500 to $19,999	49.5	46.3	45.5
$20,000 to $39,999	20.3	21.3	22.8
$40,000 to $64,999	9.8	10.1	12.6
$65,000 or more	6.4	6.4	6.0
Total	100.0	100.0	100.0

Source: U.S. Census Bureau, 2016 American Community Survey Content Test and 2016 Current Population Survey Annual Social and Economic Supplement (CPS ASEC)

The overall distributions of retirement, survivor, and disability income for the test and control treatments are reasonably close to the CPS ASEC distribution.

5.5 Response Error

This section addresses research questions #7-8:

7. *Is the absolute value of the net difference rate for combined retirement, survivor, and disability income recipiency smaller for the test treatment than for the control treatment?*

8. *For each of the ranges defined in Table 1, is the absolute value of the net difference rate for combined retirement, survivor, and disability income smaller for the test treatment than for the control treatment?*

As discussed in Section 2.4.5, the CFU income question series comes from the CPS ASEC interview and differs from the income question series in the original ACS control and test interviews. The answers to the CFU questions are treated as the 'truth.' We use the absolute value of the NDR metrics to examine the magnitude of the difference between the original response and the CFU response.

Table 20 shows the absolute value of the net difference rates for combined retirement, survivor, and disability income recipiency was smaller for the test treatment than for the control treatment

overall and for internet and mail modes of data collection. This indicates the test treatment results are closer to the truth than the control treatment results for internet and mail modes of data collection. The same is true for the high response stratum. (See Table A8 in Appendix A).

Table 20. Net Difference Rates (NDR) for Control and Test Treatments for Retirement, Survivor, and Disability Income Recipiency: All Modes of Data Collection

Mode of Data Collection	Test Sample Size	Test NDR† Percent	Control Sample Size	Control NDR† Percent	Test minus Control	Adjusted P-Value
Overall	13,608	0.6 (0.3)	13,810	3.4 (0.4)	-2.9 (0.5)	<0.01*
Internet	7,537	0.8 (0.4)	7,468	4.3 (0.4)	-3.5 (0.5)	<0.01*
Mail	3,064	0.9 (0.7)	3,151	5.0 (0.8)	-4.9 (1.0)	<0.01*
CATI	682	1.5 (1.8)	710	6.3 (2.2)	-4.8 (3.0)	0.11
CAPI	2,325	0.5 (0.7)	2,481	1.4 (0.7)	-0.9 (1.0)	0.18

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
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 one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

Table 21 shows the absolute value of the net difference rate for combined retirement, survivor, and disability income amount for each income range was not lower in the test treatment than control treatment. The same results apply for each mode of interview (see Tables 22 through 25) as well as for both the high and low response strata (see Table A9 in Appendix A).

Table 21. Net Difference Rates (NDR) for Control and Test Treatments for Retirement, Survivor, and Disability Income Amount: All Modes of Data Collection Combined

Category	Test NDR† Percent (n=1,615)	Control NDR† Percent (n=1,321)	Test minus Control	Adjusted P-Value
$1 to $2,499	2.0 (0.9)	4.8 (1.0)	-2.8 (1.4)	0.11
$2,500 to $19,999	0.2 (1.2)	1.4 (1.3)	-1.2 (1.6)	0.23
$20,000 to $39,999	1.7 (0.9)	1.9 (1.4)	-0.2 (1.7)	0.93
$40,000 to $64,999	<0.1 (1.1)	<0.1 (1.2)	<0.1 (0.9)	0.93
$65,000 or more	0.1 (0.6)	1.6 (0.8)	-1.5 (1.0)	0.72

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

Table 22. Net Difference Rates (NDR) for Control and Test Treatments for Retirement, Survivor, and Disability Income Amount: Internet Mode of Data Collection

Category	Test NDR† Percent (n = 913)	Control NDR† Percent (n=733)	Test minus Control	Adjusted P-Value
$1 to $2,499	3.8 (1.0)	7.0 (1.5)	-3.2 (1.7)	0.15
$2,500 to $19,999	0.5 (1.3)	1.8 (2.1)	-1.3 (2.2)	0.72
$20,000 to $39,999	1.3 (1.4)	0.2 (1.9)	1.1 (2.1)	0.72
$40,000 to $64,999	0.8 (1.3)	2.1 (1.6)	-1.4 (1.9)	0.72
$65,000 or more	1.2 (0.8)	3.3 (1.0)	-2.1 (1.3)	0.19

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

Table 23. Net Difference Rates (NDR) for Control and Test Treatments for Retirement, Survivor, and Disability Income Amount: Mail Mode of Data Collection

Category	Test NDR† Percent (n=524)	Control NDR† Percent (n=426)	Test minus Control	Adjusted P-Value
$1 to $2,499	0.4 (1.8)	4.2 (1.9)	-3.8 (2.3)	0.26
$2,500 to $19,999	1.4 (2.5)	1.0 (2.5)	0.4 (2.9)	1.00
$20,000 to $39,999	3.1 (2.1)	4.8 (2.1)	-1.7 (3.2)	1.00
$40,000 to $64,999	2.9 (1.5)	2.9 (1.8)	<0.1 (2.9)	1.00
$65,000 or more	2.0 (1.1)	1.3 (1.0)	0.7 (1.4)	1.00

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

Table 24. Net Difference Rates (NDR) for Control and Test Treatments for Retirement, Survivor, and Disability Income Amount: CATI Mode of Data Collection

Category	Test NDR† Percent (n=98)	Control NDR† Percent (n=89)	Test minus Control	Adjusted P-Value
$1 to $2,499	4.2 (3.3)	0.4 (0.4)	3.8 (3.3)	1.00
$2,500 to $19,999	6.9 (5.1)	1.6 (3.4)	5.3 (6.2)	1.00
$20,000 to $39,999	4.6 (5.8)	3.6 (4.4)	1.0 (6.5)	1.00
$40,000 to $64,999	6.9 (5.3)	1.6 (3.7)	5.3 (6.2)	1.00
$65,000 or more	5.0 (3.4)	3.2 (2.6)	1.7 (4.3)	1.00

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

Table 25. Net Difference Rates (NDR) for Control and Test Treatments for Retirement, Survivor, and Disability Income Amount: CAPI Mode of Data Collection

Category	Test NDR† Percent (n=80)	Control NDR† Percent (n=73)	Test minus Control	Adjusted P-Value
$1 to $2,499	2.2 (3.1)	1.6 (1.3)	0.6 (3.5)	1.00
$2,500 to $19,999	1.2 (4.9)	1.6 (4.5)	-0.4 (4.5)	1.00
$20,000 to $39,999	0.7 (1.2)	6.2 (4.0)	-5.5 (4.0)	0.43
$40,000 to $64,999	5.3 (3.6)	3.5 (4.1)	1.8 (5.2)	1.00
$65,000 or more	1.2 (0.8)	5.9 (4.1)	-4.7 (4.2)	0.52

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

6 CONCLUSIONS AND RECOMMENDATIONS

Federal surveys, including the ACS, have lagged in addressing newer forms of retirement income, namely, defined contribution plans such as 401(k), 403(b), IRA, and Roth IRA leading to an underreporting of retirement income. Positive results of a retirement income question redesign on the Current Population Survey Annual Social and Economic Supplement led to the testing of a revised Retirement, Survivor, and Disability Income question on the 2016 ACS Content Test.

A revised Retirement, Survivor, and Disability Income question was tested in the 2016 ACS Content Test with the goals of improving income reporting, increasing response rates, reducing item missing data rates, reducing reporting error, and updating questions on retirement income and the income generated from retirement accounts in order to better measure retirement income

data. The current version of the question on the ACS makes no reference to specific types of retirement income, asking only about *"retirement, survivor, or disability pensions."*

For this test, the retirement income question was expanded to ask about *"retirement income, pensions, survivor, or disability income."* In addition, the instructions accompanying the question were expanded to note that income from *a previous employer or union, or any regular withdrawals or distributions from IRA, Roth IRA, 401(k), 403(b) or other accounts specifically designed for retirement"* should be included.

The study was guided by several research questions concerning total unit response rates, the socioeconomic and demographic characteristics of responding units, comparisons to CPS ASEC data, missing data rates, and differences in the reports of recipiency and income amount by treatment and other characteristics.

The results showed several positive changes:

- The overall distribution of retirement, survivor, and disability income for the test and control versions was similar to that of the CPS ASEC.
- The number of eligible respondents receiving retirement, survivor, and disability income was significantly higher in the test treatment than the control.
- The overall retirement, survivor, and disability aggregate income amount was significantly higher in the test treatment than the control.
- The income distribution for the test and control treatments were significantly different. There was a significantly higher share of responses in the income category, '$1 to $2,499' in the test treatment and a significantly lower share of responses in the category '$40,000 to $64,999' in the test treatment than in the control treatment. In every income category, there were more respondents in the test treatment than control.
- The absolute value of the difference between the original interview proportion of positive responses and the content follow-up proportion of responses, or net difference rate, for combined retirement, survivor, and disability income recipiency was significantly smaller for the test treatment than for the control treatment.

Overall, the results support implementation of the revised version of the Retirement, Survivor, and Disability Income question.

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.

The authors would like to thank the following individuals for their contributions to the analysis and review of the 2016 Retirement, Survivor, and Disability Income report: Sarah Heimel for statistical analysis, Jessica Semega for assistance with subject matter expertise regarding the CPS ASEC, Dorothy Barth and Jonathan Rothbaum for general guidance and suggestions.

8 REFERENCES

Butrica, B, Iams, H., Smith, K., and Toder, E. (2009). "The Disappearing Defined Benefit and Its Potential Impact on the Retirement Incomes of Baby Boomers." *Social Security Bulletin*, Vol. 69 no, 3, 1-27.

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

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

Munnell, Alicia H. and Anqi Chen (2014). "Do Census Data Understate Retirement Income?" Center for Retirement Research at Boston College, Number 14-19, December 2014.

Rao, J. N. K.; 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

Semega, J., & Welniak, E., Jr. (2013). "Evaluating the 2013 CPS ASEC Income Redesign Content Test." *Proceedings of the 2013 Joint Statistical Meetings.* Montreal, Canada: American Statistical Association.

Stapleton, M., & 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., & 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: Extra Tables

Table A1. 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.

Table A2. Item Missing Data Rates for Test and Control Treatments by Response Stratum for Retirement, Survivor, and Disability Income Recipiency

Response Stratum	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	P-Value
High and Low Combined	35,248	11.4 (0.3)	35,341	11.5 (0.4)	-0.1 (0.4)	0.38
High Response Stratum	14,503	11.4 (0.4)	14,505	11.2 (0.4)	0.1 (0.5)	0.61
Low Response Stratum	20,745	11.4 (0.4)	20,836	12.4 (0.3)	-1.0 (0.5)	0.02*

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 one-tailed t-test (test < control) at the α=0.1 level.

Table A3. Item Missing Data Rates for Test and Control Treatments by Response Stratum for Retirement, Survivor, and Disability Income Amounts

Response Stratum	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	P-Value
High and Low Combined	4,679	8.9 (0.7)	3,518	9.4 (1.0)	-0.5 (1.2)	0.35
High Response Stratum	2,287	8.0 (0.8)	1,690	8.6 (1.2)	-0.5 (1.5)	0.36
Low Response Stratum	2,392	13.5 (1.0)	1,828	13.3 (1.1)	0.2 (1.4)	0.56

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 one-tailed t-test (test < control) at the α=0.1 level.

Table A4. Recipiency Rates for Retirement, Survivor, and Disability Income Test and Control Treatments

Mode	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	P-Value
High and Low Combined	30,446	14.5 (0.3)	30,358	10.6 (0.3)	3.9 (0.4)	<0.01*
High Response Stratum	12,720	15.8 (0.4)	12,739	11.4 (0.3)	4.4 (0.5)	<0.01****
Low Response Stratum	17,726	10.3 (0.3)	17,619	8.1 (0.2)	2.2 (0.3)	<0.01*

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 one-tailed t-test (test > control) at the α=0.1 level.

Table A5. Aggregate Retirement, Survivor, and Disability Income for Test and Control Treatments

Category	Test Aggregate	Control Aggregate	Test minus Control	P-Value
High and Low Combined	564,216 (22,228)	425,989 (16,450)	138,227 (24,628)	<0.01*
High Response Stratum	493,369 (21,902)	366,529 (15,609)	126,840 (24,541)	<0.01****
Low Response Stratum	70,847 (2,914)	59,459 (3,067)	11,388 (4,255)	<0.01*

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Aggregates and standard errors are shown in millions of dollars. 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 one-tailed t-test (test > control) at the α=0.1 level.

Table A6. Retirement, Survivor, and Disability Income Distribution for Test and Control Treatments for All Modes of Data Collection for High Response Stratum

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	Adjusted P-Value
$1 to $2,499	326	15.4 (0.9)	202	12.5 (0.9)	2.9 (1.4)	0.13
$2,500 to $19,999	956	45.4 (1.1)	693	44.4 (1.6)	1.1 (1.9)	1.00
$20,000 to $39,999	470	21.8 (1.0)	370	23.2 (1.1)	-1.4 (1.4)	0.93
$40,000 to $64,999	238	10.7 (0.6)	216	13.5 (0.9)	-2.8 (1.1)	0.03*
$65,000 or more	155	6.7 (0.6)	100	6.4 (0.6)	0.3 (0.9)	1.00
Total	2,145	100.0	1,581	100.0		

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 11.2$, p-value=0.02
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. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method.

Table A7. Retirement, Survivor, and Disability Income Distribution for Test and Control Treatments for All Modes of Data Collection for Low Response Stratum

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent
$1 to $2,499	409	18.4 (1.0)	271	15.9 (1.0)
$2,500 to $19,999	1,059	51.1 (1.2)	803	51.1 (1.3)
$20,000 to $39,999	419	18.9 (0.8)	367	21.2 (1.0)
$40,000 to $64,999	169	7.2 (0.6)	147	8.3 (0.8)
$65,000 or more	95	4.4 (0.5)	63	3.6 (0.6)
Total	2,151	100.0	1,651	100.0

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 6.6$, p-value=0.16
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.

Table A8. Content Follow-Up Net Difference Rates (NDR) for Persons Receiving Retirement, Survivor, and Disability Income for Test and Control Treatments

Mode	Test Sample Size	Test NDR† Percent	Control Sample Size	Control NDR† Percent	Test minus Control	Adjusted P-Value
High and Low Combined	13,608	0.6 (0.3)	13,810	3.4 (0.4)	-2.9 (0.5)	<0.01*
High Response Stratum	6,124	0.4 (0.4)	6,142	3.9 (0.4)	-3.5 (0.5)	<0.01*
Low Response Stratum	7,484	1.2 (0.4)	7,668	1.9 (0.4)	-0.7 (0.6)	0.11

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
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 one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

Table A9. Retirement, Survivor, and Disability Income Content Follow-Up Net Difference Rates (NDR) for Test and Control Treatments for High Response Stratum

Category	Test NDR† Percent (n=866)	Control NDR† Percent (n=683)	Test minus Control	Adjusted P-Value
$1 to $2,499	2.2 (1.1)	5.5 (1.2)	-3.3 (1.7)	0.13
$2,500 to $19,999	0.3 (1.4)	2.2 (1.6)	-1.9 (2.0)	0.51
$20,000 to $39,999	<0.1 (1.4)	2.0 (1.4)	-1.9 (2.0)	0.93
$40,000 to $64,999	0.4 (1.2)	0.5 (1.4)	-0.1 (1.1)	0.93
$65,000 or more	0.5 (0.6)	0.3 (1.1)	0.2 (1.1)	0.28

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. P-values in boldface indicate a significant difference based on a one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

Table A10. Retirement, Survivor, and Disability Income Content Follow-Up Net Difference Rates (NDR) for Test and Control Treatments for Low Response Stratum

Category	Test NDR† Percent (n=749)	Control NDR† Percent (n=638)	Test minus Control	Adjusted P-Value
$1 to $2,499	1.1 (1.4)	1.5 (1.2)	-0.4 (1.7)	1.00
$2,500 to $19,999	0.2 (1.9)	3.0 (1.5)	-2.8 (2.4)	0.59
$20,000 to $39,999	<0.1 (1.4)	2.0 (1.4)	-1.9 (2.0)	0.67
$40,000 to $64,999	1.8 (1.3)	2.1 (1.1)	-0.3 (1.7)	1.00
$65,000 or more	0.5 (1.0)	0.3 (1.1)	0.2 (1.1)	1.00

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
† Represents the absolute value of the net difference rate.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. P-values in boldface indicate a significant difference based on a one-tailed t-test (test < control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The value in the "Test minus Control" column represents the difference in the absolute values of test and control NDRs.

Table A11. Aggregate Retirement, Survivor, and Disability Income by Source of Income for Test and Control Treatments

Source	Test Aggregate	Control Aggregate	Test minus Control	P-Value
Wages and Salary	5,272,840 (117,973)	5,583,487 (153,289)	-310,647 (163,190)	0.06*
Self-Employment	344,803 (34,664)	406,602 (48,652)	-61,798 (58,737)	0.29
Interest, Dividends, Net Rental Income	264,536 (26,375)	344,733 (29,930)	-80,197 (38,836)	0.04*
Social Security Income	470,158 (9,986)	461,271 (9,926)	8,887 (12,458)	0.48
Supplemental Security Income (SSI)	43,099 (2,482)	47,051 (2,160)	-3,952 (3,610)	0.27
Public Assistance Income	7,558 (1,429)	5,950 (1,046)	1,608 (1,862)	0.39
Retirement, Survivor, and Disability Income	564,216 (22,228)	425,989 (16,450)	138,227 (24,628)	<0.01*
Other Income	71,899 (5,901)	83,486 (7,845)	-11,588 (9,938)	0.24
Total Income (reported)	8,963,489 (1,056,704)	8,977,659 (612,199)	-14,170 (1,292,054)	0.92

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

Note: Aggregates and standard errors are shown in millions of dollars. 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 one-tailed t-test (test > control) at the α=0.1 level.

Appendix B: Content Test Follow Up Questions - Retirement Income from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC)

The following list summarizes the questions that were asked during follow up interviews. The wording of the questions in this list varies slightly from those used by interviewers. The wording was changed to correspond to the situation of either the respondent answering for himself or herself or the respondent answering in proxy for another household member. The questions also adhered to skip patterns, so all questions may not have been asked to each respondent.

1. **Did you receive any pension income from a previous employer or union (other than Social Security or VA benefits) DURING THE PAST 12 MONTHS?** Yes or No.

   Interviewer Instructions: PLEASE DO NOT INCLUDE ANNUITIES OR DISTRIBUTIONS OR WITHDRAWALS FROM IRAs, 401(k)s, OR SIMILAR ACCOUNTS!

2. **What was the source of your pension income?**

   Interviewer Instructions: If "Yes", specify other source of pension income. Enter "Other Pension" if the answer is "Don't Know" or "Refuse".

3. **How much did you receive DURING THE PAST 12 MONTHS? (This question is asked for each type of pension income that was reported.)**

4. **How much did you receive in annuities DURING THE PAST 12 MONTHS?**

5. **Did you have any retirement accounts such as a 401(k), 403(b), IRA, or other account designed specifically for retirement savings DURING THE PAST 12 MONTHS?** Yes or No.

6. **What type of retirement account did you have DURING THE PAST 12 MONTHS? Did you have a …**

   1. 401(k)
   2. 403(b)
   3. Roth IRA
   4. Regular IRA
   5. KEOGH plan ("KEE-OH")
   6. SEP plan (Simplified Employee Pension)
   7. another type of retirement account

7. **What was the source of your retirement account?**

8. **Did you withdraw any money or receive a distribution from your retirement account DURING THE PAST 12 MONTHS?** Yes or No.

9. **Did you re-invest or "roll over" any of the money into an IRA or some other kind of retirement plan DURING THE PAST 12 MONTHS? Yes or No.**

10. **How much do you plan to re-invest or “roll over” into an IRA or some other kind of retirement plan?**