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

September 7, 2017

2017 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES # ACS17-RER-08

MEMORANDUM FOR	Victoria Velkoff Chief, American Community Survey Office
From:	David Waddington Chief, Social, Economic, and Housing Statistics Division (SEHSD)
Prepared by:	Adam Smith Social, Economic, and Housing Statistics Division (SEHSD)
Subject:	2016 American Community Survey Content Test Evaluation Report: Number of Weeks Worked

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

If you have any questions about this report, please contact Rebecca Chenevert at 301-763-8538 or Adam Smith at 301-763-9340.

Attachment

cc:
Kathryn Cheza (ACSO)
Jennifer Ortman (ACSO)
David Raglin (ACSO)
Patrick Cantwell (DSSD)
Elizabeth Poehler (DSSD)
Michael Risley (DSSD)
Anthony Tersine (DSSD)
Rebecca Chenevert (SEHSD)
Jennifer Day (SEHSD)
Nicole Scanniello (SEHSD)

Intentionally Blank

2016 American Community Survey Content Test Evaluation Report: Number of Weeks Worked

FINAL REPORT

Adam Smith, David Howard
Social, Economic, and Housing Statistics Division

Michael Risley
Decennial Statistical Studies Division

Intentionally Blank

TABLE OF CONTENTS

List of Tables

List of Figures

EXECUTIVE SUMMARY

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.

Number of Weeks Worked

Employment questions in the ACS include a measure for the number of weeks that a respondent worked in the last year, known as *weeks worked*. The chief purpose of the *weeks worked* question is to establish a time-period foundation for studying earnings data collected in the ACS. An especially important function of the question is to identify people who worked year-round (50-52 weeks) in the past 12 months. This year-round identification makes it possible to classify workers as "full-time, year-round," which is a key measure used to analyze and present ACS earnings data.

The current version of *weeks worked* was introduced in 2008. This version consists of two parts – part A and part B. Part A asks respondents if they worked all 52 weeks in the past year. Those who respond 'no' to part A are then asked part B, which asks for the number of weeks the respondent worked in the past year using range categories as answer choices. Prior to 2008, *weeks worked* was a single-part question that asked respondents to write in the number of weeks he or she worked in the last year. The 2008 *weeks worked* question change was part of a larger change to the entire employment series of questions. These changes were implemented due to evidence that the employment section was undercounting employment and overcounting unemployment.

After switching *weeks worked* from a write-in response to a categorical response in 2008, the Census Bureau lost the ability to provide continuous measures for the number of weeks worked,

such as means, medians, and aggregates. In the years since this question change, many stakeholders have expressed the need for these types of estimates as well as additional specificity in regards to weeks worked, particularly for use with usual hours worked, income, and occupation. This became the motivation for including *weeks worked* in the 2016 ACS Content Test, to determine whether a write-in style for part B could allow for additional specificity without sacrificing data quality.

For the 2016 ACS Content Test, this question was revised with an open-ended response for part B while retaining the Yes/No response for part A of the question. Question wording for both parts were also revised to provide more context and clarity.

Research Results

The analysis of *weeks worked* data collected in the Content Test was guided by several research questions. These questions focused on results for item missing data rates, response distributions, derived estimates, response error, and earnings estimates. This analysis also used data from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and the Survey of Income and Program Participation (SIPP) for benchmark comparisons, as well as the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) administrative data for further earnings comparisons and comparisons using administrative data.

- Item missing data rates were not significantly lower for the test treatment, and in fact were one percentage point higher for part B.
- The percentage of year-round workers was higher in the test version by about 2 percentage points.
- The full-time, year-round rate in the test version was greater by about one percentage point.
- Both the test and control distributions of *weeks worked* appear comparable to the CPS ASEC distribution. Both the test and control distributions of weeks worked appeared to differ from SIPP estimates. The percentage of workers whose reported weeks worked fell within the range of values implied by the number of reported quarters in the LEHD was largely comparable across treatments. There was no significant difference in mean quarters worked between treatments.
- The test proportion of full-time, year-round workers (those working 35 or more hours for 50 or more weeks) appears comparable to the CPS ASEC distribution, while the control proportion is not. Both the test and control proportions of full-time, year-round workers appear to differ from SIPP estimates.
- Response reliability improved in the test version.
- Median ACS-reported earnings were not significantly different between the test and control treatments in any category.
- Mean LEHD administrative earnings were higher in the test treatment than the control treatment for "27 to 39 weeks worked" (by $4,500) and "0 to 13 weeks worked" (by $4,500). All other categories were not statistically different.
- Median LEHD administrative earnings were higher in the test treatment than the control treatment by $2,200 for "27 to 39 weeks worked," and lower by $3,500 in the "40 to 47 weeks worked." All other categories were not significantly different.

Conclusion

The results presented in this report support recommending the test version of *weeks worked* for implementation on the ACS. Empirical results suggest that changing part B of *weeks worked* to an open-ended response, along with question text changes to part A and part B, will allow for greater specificity without adversely affecting data quality, and in fact might improve data consistency.

All decision criteria were met, save for the test version's increase in the item missing data rate for part B (at 4.3 percent versus 3.3 percent). This was not a major concern because write-in responses generally result in a higher missing data rate. The estimate of full-time, year-round workers was higher in the test treatment, and earnings estimates between treatments were not significantly different overall. In addition, the test version did not adversely impact the distribution of weeks worked or response reliability. Administrative data comparisons also suggested that recommending the test version of *weeks worked* over the control version will not adversely affect data quality for the number of weeks worked.

Intentionally Blank

1 BACKGROUND

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 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 a revision to the *Number of Weeks Worked* topic, commonly known as *weeks worked.*

1.1 Justification for Inclusion of Number of Weeks Worked in the Content Test

From 2005 to 2007, *weeks worked* was asked as a single-part question with an open-ended response, where respondents wrote in a specific number of weeks worked. In 2008, *weeks worked* was split into two parts: part A that has a yes/no question for working the full year, and part B that offers six discrete ranges as answer categories to those who worked less than the full year. The 2008 question change was part of a larger change to the entire employment series of questions. These changes were implemented due to evidence that the employment section was undercounting employment, and overcounting unemployment.

After switching *weeks worked* from a write-in response to a categorical response in 2008, the Census Bureau lost the ability to provide continuous measures for the number of weeks worked, such as means, medians, and aggregates. Since this question change, many stakeholders have expressed the need for these types of estimates as well as additional specificity in regards to weeks worked, particularly for use with usual hours worked, income, and occupation.

Although the 2008 question change significantly increased the estimate of year-round workers (which was deemed an improvement for the measure), it remained unclear which change drove this increase. In addition, because both the *weeks worked* question and the remainder of the employment series were changed in 2008, the Census Bureau was unsure if the categorical response was necessary to maintain year-round levels. This became the motivation for including *weeks worked* in the 2016 ACS Content Test, to determine whether a write-in style for part B could allow for additional specificity without sacrificing data quality.

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.[1] For proposed new questions, the justification also described the need for the new data, whether federal law or regulation required the data for small areas or small population groups, if other data sources were currently available 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, 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

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

assess the revised versions of the questions in Spanish and identify any issues with questionnaire wording unique to Puerto Rico and GQ populations.[2] 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.[3]

For *weeks worked*, two versions of the question were submitted for the first round of cognitive testing. Both versions were based on the current ACS production version of *weeks worked*. These versions were as follows:

- Version 1
    - Part A: During the PAST 12 MONTHS (52 weeks), did this person work 50 or more weeks? Count paid time off as work. [Yes/No]
    - Part B: How many weeks DID this person work, even for a few hours, <u>including</u> paid vacation, paid sick leave, and military service? [____ weeks]
- Version 2
    - During the PAST 12 MONTHS (52 weeks), how many WEEKS did this person work? Count paid vacation, paid sick leave, and military service. [____ weeks]

As shown in the above list, version 1 used part A of the current production version of *weeks worked* without the skip instructions, and changed part B to have an open-ended response. The second version was solely part B of the production version with an open-ended response and a reference timeframe added to the question.

Based on results from the first round of testing, a single, revised version of the question was submitted for the second round of testing. The second round of testing used Version 1 from the first round, but switched the examples of "paid time off" from part B to part A, and placed an emphasis on "WEEKS" instead of "DID" for part B. This version was as follows:

- Part A: During the PAST 12 MONTHS (52 weeks), did this person work 50 or more weeks? Count paid vacation, paid sick leave, and military service as work. [Yes/No]
- Part B: How many WEEKS did this person work, even for a few hours, <u>including</u> paid time off? [____ weeks]

After the second round of testing, it was clear that *weeks worked* still suffered from reference period issues historic to the question. Although the current question specifically notes there are 52 weeks in the past 12 months, respondents continued to have difficulty understanding this concept (Stapleton & Steiger, 2015). Respondents with very transitory work schedules had the

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

[3] 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.

most difficulty. As a result, we chose to change part A of the test question to ask if a respondent worked "every week" to improve clarity.

The final version of *weeks worked* – the version to be used in the 2016 ACS Content Test – implemented the following changes to the question introduced in 2008:

- Asked if the respondent worked "EVERY week" (instead of "50 or more weeks);
- Brought back skip instructions;
- Repeated the reference timeframe in part B; and
- Moved guidance for paid time off and marginal work into an instruction portion of the item after the question itself.

Cognitive testing results indicate that these changes should reduce the following: reporting error for paid time off, reporting error related to the reference period, and respondent burden for those who worked every week in the last year. The final test and control versions of *weeks worked* used in paper forms for the 2016 ACS Content Test are shown in Figures 1 and 2 below.

1.3 Question Content

As previously stated, question revisions to *weeks worked* aim to capture more detailed data for stakeholders – specifically a continuous measure for the number of weeks worked for the past 12 months. To this end, both part A and part B of the *weeks worked* question were revised and tested during the 2016 ACS Content Test. Figures 1 and 2 show the control and test versions of each question as they appeared on the Content Test's mail questionnaire.[4]

Figure 1: Control Version of Weeks Worked Question

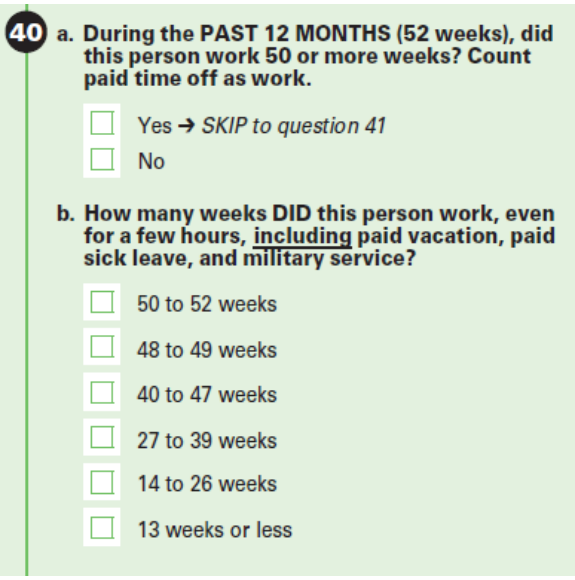
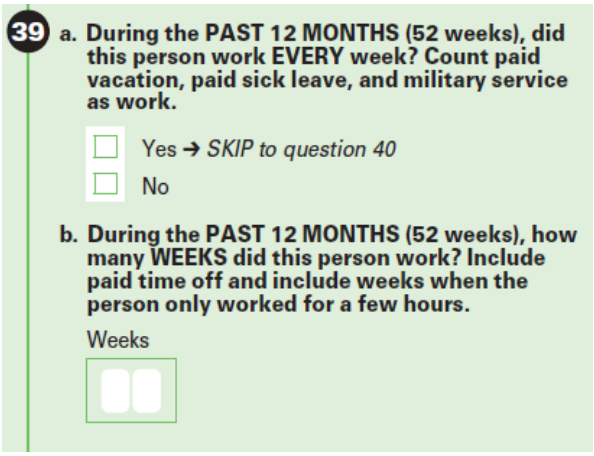


Figure 2: Test Version of Weeks Worked Question



[4] Automated versions of the questionnaires included the same text with formatting changes appropriate to each mode.

In part A, the test version rephrased the period of interest (from working "50 or more weeks" to "EVERY week"). It also provided additional information in the second sentence by changing the original instruction from "Count paid time off as work" to "Count paid vacation, paid sick leave, and military service as work." Part B of the question was changed to a write-in response (the primary motivation for this research), repeated the reference period ("the PAST 12 MONTHS"), and added new guidance for what to count as work.

1.4 Research Questions

The following research questions were formulated to guide our analyses of the *weeks worked* question. The analyses assess how the test version of the questions performed compared to the control version in the following ways: how often the respondents answered the question, how the responses affected the resulting estimates, the consistency and accuracy of the responses, how the responses compared to benchmarks, how associated administrative data compared between versions, and how earnings estimates compared between versions.

1. *Is the missing data rate, for each part of the weeks worked question, lower for the test treatment than for the control treatment?*

2. *After assigning test version responses to corresponding range categories from the control version, how do the test and control response distributions compare? If they are not the same, which range categories have different proportions?*

3. *Is the proportion of full-time, year-round workers (worked 35+ hours a week for 50+ weeks) greater in the test version than in the control version?*

4. *How does the distribution of weeks worked for each treatment compare with the distribution from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC)?*

5. *How does the proportion of full-time, year-round workers (worked 35+ hours a week for 50+ weeks) for each treatment compare with CPS ASEC estimates?*

6. *How does the distribution of weeks worked for each treatment compare with the distribution from the Survey of Income and Program Participation (SIPP)?*

7. *How does the proportion of full-time, year-round workers for each treatment compare with SIPP estimates?*

8. *Are the measures of response reliability (gross difference rate, index of inconsistency) better for the test treatment than for the control treatment?*

9. *How do test and control distributions compare to Longitudinal Employer-Household Dynamics (LEHD) data?*

10. *Is there a difference in the mean number of LEHD quarters worked between the test sample and the control sample?*

11. *Is there a difference in mean LEHD earnings between the test sample and the control sample?*

12. *How do median LEHD earnings compare between the test sample and the control sample?*

13. *How do ACS-reported median earnings compare between test and control responses?*

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 the Computer-Assisted Telephone Interviews (CATI) and Computer-Assisted Personal Interviews (CAPI) did not take into account the oversampling of low response areas. This oversampling resulted in a higher than expected workload for CATI and CAPI and therefore required more budget than was allocated. To address this issue, the CAPI sampling rate for the Content Test was adjusted to meet the budget constraint.

---

[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).

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 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.

[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 mail 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.

2.3 Content Follow-Up

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

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

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

2.4 Analysis Metrics

This section describes the methodology and metrics used to assess revisions to the *weeks worked* question. These metrics include unit response rates, item missing data rates, response distributions, derived estimates, comparisons to benchmarks, response error, earnings differences, and comparisons using administrative data. This section also describes the methodology used to calculate unit response rates and standard errors used 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

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

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

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

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

Only those aged 16 and older were considered in-universe for *weeks worked* analysis. This means that those aged 15 – even though they were instructed to answer the questionnaire – were excluded from the results. This age requirement is consistent with the Census Bureau's published employment estimates. In addition, unless otherwise noted, only those who answered "yes" for either "worked last week" or "worked within the past 12 months" are included in analyses. Records with a missing value for *weeks worked* were only included for Research Question 1. Respondents who reported working zero weeks in part B of the test version were included in the analyses because they would likely be edited to have a positive value for *weeks worked* in the production version of ACS.

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.

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

2.4.1 Unit Response Rates and Demographic Profile of Responding Households

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

For both control and test treatments, we calculated the overall unit response rate (all modes of data collection combined) and unit response rates by mode: internet, mail, CATI, and CAPI. We also calculated the total self-response rate by combining internet and mail modes together. Some 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.

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

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, 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.[15]

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

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

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 questions 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.

For *weeks worked*, missing data rates for both part A and part B of the question were of interest. The universe for item missing data analysis included those on path to answer the question series and those who answered it even though they should have skipped the question. For part B analysis of the control version, mail mode responses where multiple (two or more) categories were selected (checked) were counted as responses, even though the variable is set to 'missing' in normal ACS processing.

For both part A and part B, statistical significance for item missing data rates between versions was determined using a one-tailed t-test to measure whether the test's missing data rate was lower than the control's rate (see Research Question 1). Because we observed a higher missing data rate in part B of the test version, we also performed a two-tailed t-test to determine the statistical significance of this increase.

2.4.3 Response Distributions and Derived Estimates

Comparing response distributions between the control and test versions of *weeks worked* allowed us to assess whether and how the question change affects the resulting estimates. Comparisons were made using Rao-Scott chi-squared tests (Rao & Scott, 1987) for distributions, and t-tests for single categories when the corresponding distributions were found to be statistically different. Because the goal of changing *weeks worked* was to allow for more specificity without a loss in data quality, it was most desired for the distributions to be similar; however, with the question text changes made to part A for the test version, an increase in the full-year category was expected.

Proportion estimates were calculated as:

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

In order to compare continuous distributions from the test version with category-based distributions of the control version, responses from the test version were mapped to matching control-version categories. For example, anyone in the test version who wrote in a response

between 40 to 47 weeks worked was assigned to a “40 to 47 weeks” category. For both treatments, the “50 to 52 weeks” category was assigned by either a "yes" response to part A or identifying 50-52 weeks worked for part B.

Regarding derived estimates, *weeks worked* is a part of determining the full-time, year-round rate of workers. The full-time, year-round rate is the percent of workers who identified as year-round workers *and* reported working 35 hours or more as their usual work hours per week. This derived estimate is a key measure produced by the ACS employment series, and is also used in most ACS-based earnings tables published by the Census Bureau. Given its importance, we found it necessary to include the full-time, year-round rate in our analysis.

2.4.4 Benchmarks

The federal government reports employment and unemployment estimates from several major surveys and programs; however, estimates specific to the number of weeks worked are less common. For evaluating *weeks worked,* Content Test data were compared with the CPS ASEC and the SIPP.

The CPS is a long-running survey conducted jointly by the Census Bureau and the Bureau of Labor Statistics. The CPS ASEC is an annually administered supplement to the core CPS labor force topics. The key purpose of this supplement is to provide timely, detailed estimates of income and poverty and to measure changes in these estimates. It asks about work experience for the past calendar year, which includes the number of weeks worked. The 2016 CPS ASEC data, whose reference period is 2015, was the latest available at the time of writing this report.

The SIPP is a longitudinal survey that collects data related to various types of income, labor force participation, social program participation and eligibility, and general demographic characteristics in order to measure the effectiveness of existing federal, state, and local programs. As part of its labor force section, the SIPP collects information on all jobs worked using an “Event History Calendar,” which defines start and end dates for jobs throughout the year. Wave 1 of the SIPP’s 2014 Panel, whose reference period is 2013, was the latest available.

Because the 2016 ACS Content Test was not designed to produce national estimates, comparisons to these sources are for informational purposes only.[16] The results cannot be statistically compared, but similarities can be discussed based on whether or not the Content Test estimates fall within a benchmark’s confidence interval. The relevant comparisons for *weeks worked* were the overall response distributions from both treatments, as well as their full-time, year-round rates.

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

[16] The 2016 ACS Content Test cannot make national estimates in part because the sample did not include group quarters, Alaska, or Hawaii, and because no editing was done for consistency and imputation. The 2016 ACS Content Test also does not include the production-ACS weighting adjustments for seasonal variations in response patterns, nonresponse bias, and under-coverage bias.

two components of response error: response bias and simple response variance. Response bias is the degree to which respondents consistently answer a question incorrectly. Simple response variance is the degree to which respondents answer a question inconsistently. A question has good response reliability if respondents tend to answer the question consistently. Re-asking the same question of the same respondent (or housing unit) allows us to measure response variance.

We measured simple response variance by comparing valid responses to the CFU reinterview with valid responses to the corresponding original interview.[17] The Census Bureau has frequently used content reinterview surveys to measure simple response variance for large demographic data collection efforts, including the 2010 ACS Content Test, and the 1990, 2000, and 2010 decennial censuses (Dusch & Meier, 2012).

The following measures were used to evaluate consistency:

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

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

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

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

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

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

a = weighted count of units in the category of interest for both the original interview and reinterview

b = weighted count of units NOT in the category of interest for the original interview, but in the category for the reinterview

c = weighted count of units in the category of interest for the original interview, but NOT in the category for the reinterview

d = weighted count of units NOT in the category of interest for either the original interview or the reinterview

n = total units in the universe = a + b + c + d.

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

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

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

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

In order to define the IOI, we must first discuss the variance of a category proportion estimate. If we are interested in the true proportion of a total population that is in a certain category, we can use the proportion of a survey sample in that category as an estimate. Under certain reasonable assumptions, it can be shown that the total variance of this proportion estimate is the sum of two components, sampling variance (SV) and simple response variance (SRV). It can also be shown that an unbiased estimate of SRV is half of the GDR for the category (Flanagan, 1996).

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

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

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

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

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

The measures of response error assume that those characteristics in question did not change between the original interview and the CFU interview. To the extent that this assumption is incorrect, we assume that it is incorrect at similar rates between the control and test treatments. For instance, a person who was not in the labor force during the original interview might have started a job before the CFU interview and then accurately reported a different follow-up response to *weeks worked*.

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

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

$$\mathrm{IOI}_{\mathrm{testimate}} = \frac{\frac{n^2(b+c) - n(c-b)^2}{n-1}}{(a+c)(c+d) + (a+b)(b+d)} * 100$$

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

2.4.6 Other Analysis and Methodology Specific to Number of Weeks Worked

Weeks worked is a vital piece to editing, imputing, and publishing ACS earnings data, so it was also important to test for differences in earnings between the test and control groups. Any differences could signal that earners were interpreting the test and control versions differently. Overall, the test and control question wording was quite similar, so no differences were expected.

For the 2016 ACS Content Test, we measured earnings using two different sources: 1) earnings reported in the Content Test and 2) earnings reported in the Census Bureau's LEHD administrative data. LEHD combines employer and employee records from federal, state, and Census data. These linked employer-employee records ultimately cover over 95 percent of private sector jobs, as well as most public sector jobs (Goetz, Hyatt, McEntarfer, & Sandusky, 2015). The Person Identification Validation System (PVS) attempts to identify Personal Identification Keys (PIKs) of survey respondents in order to link survey records back to administrative records (Wagner & Layne, 2014).

Research questions 9 through 13 make use of the Unemployment Insurance (UI) component of the LEHD data. UI data indicate whether someone participated in the labor force in a given quarter, assuming his or her job was covered by unemployment insurance, as well as earnings amounts for the quarters that he or she worked.

2.4.7 Standard Error Calculations

We estimated the variances of the estimates using the Successive Differences Replication (SDR) method with replicate weights, the standard method used in the ACS (see U.S. Census Bureau,

2014, Chapter 12). We calculated the variance for each rate and difference using the formula below. The standard error of the estimate (X_0) is the square root of the variance:

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

where:

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

3 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. Table 2 identifies the research questions and associated metrics in priority order.

Table 2: Decision Criteria for Number of Weeks Worked

Research Questions	Decision Criteria, in order of priority
3, 5, 7, 9	(1) The estimate of full-time, year-round workers in the test version should remain consistent with that in the control version. If the full-time, year-round level is higher in the test version, compare with benchmarks to determine if the increase is consistent.
1	(2) The test version should not adversely impact item missing data rates. An increase in item missing data rates for part A would be more concerning than for part B, unless the increase in item missing data rates for part B is dramatic.
11, 12, 13	(3) The test version should not significantly impact earnings estimates.
8, 2, 4, 6, 10	(4) The test version should not adversely impact response reliability or the resulting distribution of weeks worked. If the distributions are different, compare with benchmarks to determine if the changes are consistent with other data sources.

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 conducted in English and Spanish only. Respondents who needed language assistance in another language were not able to participate in the test. Additionally, the 2016 ACS 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.[18] 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 2016 ACS Content Test does not include the production weighting adjustments for seasonal variations in ACS response patterns, nonresponse bias, and under-coverage bias. As a result, any estimates derived from the 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

[18] 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.

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. Additionally, the control treatment questionnaires were processed more quickly in keying than the test treatment questionnaires resulting in a longer delay for test mail cases to be eligible for CFU. On average, it took 18 days for control cases to become eligible for CFU; it took 20 days for test cases. The difference is statistically significant. This has the potential to impact the response reliability results.

The assumption of parallel measures for the GDR and IOI calculations was not met for the "48 to 49 Weeks" category. For this category, the GDR and IOI estimates are biased to some extent.

The analysis for weeks worked also had a few limitations when comparing to benchmark and administrative data. These limitations were based on the different reference periods between the 2016 ACS Content Test, CPS ASEC, SIPP, and LEHD. In general, the reference period for the 2016 ACS Content Test data is March of 2015 to March of 2016 because most responses were collected in March of 2016. The 2016 CPS ASEC data – the latest available at the time of writing this report – has a reference period of January to December of 2015. Wave 1 of SIPP's 2014 Panel – also the latest available – has a reference period of January to December of 2013. As for LEHD data, administrative records for quarter 2 of 2015 through the end of quarter 1 of 2016 (i.e., April 2015 through March 2016) were available. This time-frame overlaps with the reference period for most Content Test responses.

Table 3 summarizes the reference periods used in our analysis.

Table 3. Reference Periods Used for Benchmark Analysis

Source	Reference Period
2016 ACS Content Test	March 2015 - March 2016[1]
2016 CPS ASEC	January 2015 - December 2015
2014 SIPP - Wave 1	January 2013 - December 2013
LEHD	April 2015 - March 2016

[1]This range reflects the reference period for most Content Test respondents.

5 RESEARCH QUESTIONS AND RESULTS

This section presents the results from the analyses of the 2016 ACS Content Test for the *weeks worked* 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 corresponding data and a brief summary of its results.

5.1 Unit Response Rates and Demographic Profile of Responding Households

This section provides results for unit response rates for both control and test treatments for the original Content Test interview and for the CFU interview. It also provides 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 by 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, 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.[19] 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.[20] If a respondent is offended by the questionnaire or feels that the questions are too burdensome they may just throw the questionnaire away, and not respond by mail. This could be a possible explanation for the unit response rate being lower for test than control in the mail mode.

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

Table 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 of H_0:*test=control* 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

[19] Table A-1 (including all modes) can be found in Appendix A.
[20] 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.

and interviewer. This can influence response rates. We do not believe that the difference suggests any underlying CFU response issues that would negatively affect topic-specific response reliability analysis for comparing the two treatments.

Table 6. Content Follow-Up Interview Unit Response Rates, 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 of H_0:*test=control* at the α=0.1 level

5.1.3 Demographic and Socioeconomic Profile of Responding Households

One of the underlying assumptions of our analyses 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.*[21] The response distribution calculations are 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 below is statistically significant.

[21] 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.

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 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 are shown in Table 8 and Table 9.

Table 8. Comparison of Average Household Size

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

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Standard errors are shown in parentheses. Significance was tested based on a two-tailed t-test of H_0:*test=control* 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. Minor additive discrepancies are due to rounding.
Significance was tested based on a two-tailed t-test of H_0:*test=control* 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 were no detectable differences between control and test for *average household size* or *language of response*. There were 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 details item missing data rates for *weeks worked*. Item missing data rates measure the proportion of persons eligible for the question with missing data. To support the question change to *weeks worked*, the test should not adversely affect item missing data rates.

Research Question 1: *Is the missing data rate, for each part of the weeks worked question, lower for the test treatment than for the control treatment?*

Table 10 shows item missing data rates for part A and part B of *weeks worked*. For part A, the test treatment's missing data rate was not significantly lower than the control's missing data rate. Both missing data rates were around 2.5 percent. Likewise, for part B the test rate was not significantly lower than the control's rate.

Because we observed a higher missing data rate in part B of the test version, we also performed a two-tailed t-test to determine the statistical significance of this increase. The results showed that the test treatment was significantly higher by one percentage point (4.3 percent versus 3.3

percent).[22] This increase was reasonable given that write-ins generally result in a higher missing data rate.

Table 10. Weeks Worked Item Missing Data Rates

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test Minus Control	P-Value
Part A	21,120	2.5 (0.1)	21,524	2.4 (0.2)	<0.1 (0.2)	0.58
Part B	6,151	4.3 (0.4)	6,760	3.3 (0.4)	1.0 (0.6)	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. Significance was tested based on a one-tailed t-test of H_0*:test ≥ control* at the α=0.1 level

5.3 Response Distributions and Derived Estimates

Research Question 2: *After assigning test version responses to corresponding range categories from the control version, how do the test and control response distributions compare? If they are not the same, which range categories have different proportions?*

A chi-square test of independence suggested that the test and control distributions were different. Because of this, we looked for significant differences within each category of response.

As seen in Table 11, the percent of year-round workers was higher in the test group by about two percentage points (78.8 percent versus 77.0 percent).[23] This was likely due to changing part A of *weeks worked* to ask respondents if they worked "EVERY week" instead of asking "50 or more weeks." Asking respondents if they worked "EVERY week" may also have curbed response errors related to miscalculating the number of weeks in a year. Other changes to the distribution included a 0.8 percentage point increase for the "14 to 26 weeks" group and a decrease of 0.8 and 1.5 percentage points for the "48 to 49 weeks" and "27 to 39 weeks" groups respectively.

[22] A two-tailed t-test of H_0*:test ≤ control* resulted in a p-value of 0.08

[23] For control results, year-round workers (50 to 52 weeks) include those who responded either 'yes' to part A of *weeks worked* or checked "50 to 52 weeks" in part B of the question. For test results, year-round workers include those who responded either either 'yes' to part A of *weeks worked* or provided a write-in response of 50, 51, or 52 weeks in part B of the question.

Table 11. Weeks Worked Response Distributions

Category	Test Percent (n=19,233)	Control Percent (n=19,676)	Test Minus Control	Adjusted P-Value
50 to 52 weeks	78.8 (0.5)	77.0 (0.5)	1.9 (0.7)	0.02*
48 to 49 weeks	1.6 (0.1)	2.4 (0.2)	-0.8 (0.2)	<0.01*
40 to 47 weeks	5.3 (0.2)	5.5 (0.2)	-0.2 (0.3)	0.94
27 to 39 weeks	4.1 (0.2)	5.6 (0.2)	-1.5 (0.3)	<0.01*
14 to 26 weeks	5.0 (0.3)	4.2 (0.2)	0.8 (0.3)	0.03*
0 to 13 weeks	5.3 (0.2)	5.3 (0.2)	-0.1 (0.3)	0.94
Total	100.0	100.0		

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 45.4$, 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 of H_0:*test=control* at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method.

Research Question 3: *Is the proportion of full-time, year-round workers (worked 35+ hours for 50+ weeks) greater in the test version than in the control version?*

Table 12 shows the full-time, year-round rates from each treatment. Respondents are considered "full-time, year-round" if they report usually working at least 35 hours per week (a question that was consistent between treatments) for 50 or more weeks in the past year. The full-time, year-round rate in the test group was greater than the control group's rate by 1.2 percentage points (67.1 percent versus 65.9 percent). Again, this was likely due to changing part A of *weeks worked* to ask respondents if they worked "EVERY week" instead of "50 or more weeks."

Table 12. Full-Time Year-Round Rate

Category	Test Percent (n=18,851)	Control Percent (n=19,232)	Test minus Control	P-Value
Full-time, year-round	67.1 (0.6)	65.9 (0.6)	1.2 (0.8)	0.06*

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 of H_0:*test ≤ control* at the α=0.1 level.

5.4 Benchmarks

Weeks worked data from the 2016 ACS Content Test were compared with estimates from the 2016 CPS ASEC and the 2014 SIPP. As previously stated, because the 2016 ACS Content Test was not designed to produce national estimates, these comparisons are for informational purposes only. The results cannot be statistically compared, but similarities can be discussed based on whether or not the Content Test estimates fall within a benchmark's confidence interval.

Research Question 4: *How does the distribution of weeks worked for each treatment compare with the distribution from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC)?*

The CPS ASEC collects information on the number of weeks worked in the past calendar year. Table 13 shows the distribution of weeks worked from the test and control groups along with the distribution from the 2016 CPS ASEC. Overall, the test and control distributions appear comparable to the CPS ASEC distribution. For the test group, all categories except for "0 to 13" fell within CPS ASEC's confidence interval (at the 90 percent confidence level). For the control group, the "48 to 49 weeks" and "40 to 47 weeks" categories were within CPS ASEC estimates' confidence interval.

Table 13. Weeks Worked Distribution, 2016 ACS Content Test and 2016 CPS ASEC

Category	Test Percent (n=19,233)	Control Percent (n=19,676)	2016 CPS Percent (n=91,008)
50 to 52 weeks	78.8 (0.5)	77.0 (0.5)	79.2 (1.2)
48 to 49 weeks	1.6 (0.1)	2.4 (0.2)	1.9 (0.4)
40 to 47 weeks	5.3 (0.2)	5.5 (0.2)	5.1 (0.6)
27 to 39 weeks	4.1 (0.2)	5.6 (0.2)	4.1 (0.6)
14 to 26 weeks	5.0 (0.3)	4.2 (0.2)	5.5 (0.6)
0 to 13 weeks	5.3 (0.2)	5.3 (0.2)	4.3 (0.6)
Total	100.0	100.0	100.0

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, 2016 Current Population Survey - Annual Social and Economic Supplement
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. CPS population controls are updated periodically. As a result, published CPS estimates may not match those in this report.

Research Question 5: *How does the proportion of full-time, year-round workers (worked 35+ hours a week for 50+ weeks) for each treatment compare with CPS ASEC estimates?*

Table 14 shows the full-time, year-round rates from the test and control treatments along with the 2016 CPS ASEC rate. The test treatment's estimate was within the CPS ASEC's confidence interval, whereas the control's estimate was not. Based on recent full-time, year-round estimates for the ACS and CPS ASEC, the 2016 ACS Content Test rates were expected to be around three to four percentage points less than the CPS ASEC rate. Because of this, the test percent moving closer to the CPS ASEC percent was a favorable result.

Table 14. Full-Time Year-Round Rate, 2016 ACS Content Test and 2016 CPS ASEC

Category	Test Percent (n=18,851)	Control Percent (n=19,232)	2016 CPS Percent (n=91,008)
Full-time, year-round	67.1 (0.6)	65.9 (0.6)	68.6 (1.3)

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, 2016 Current Population Survey - Annual Social and Economic Supplement
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. CPS population controls are updated periodically. As a result, published CPS estimates may not match those in this report.

Research Question 6: *How does the distribution of weeks worked for each treatment compare with the distribution from the Survey of Income and Program Participation (SIPP)?*

SIPP collects information on all jobs held by respondents in the past calendar year, including beginning and end dates. This enables the calculation of the number of weeks worked for respondents. Table 15 shows the distribution of weeks worked from the test and control treatments along with the distribution from SIPP. No categories in either treatment fell within SIPP's confidence interval (at the 90 percent confidence level). Because there was a major instrument redesign for the 2014 SIPP panel, we did not have an expectation for these results. Although there were noticeable differences, the overall pattern of SIPP's distribution is very similar to the test and control distributions, with the vast majority in the 50 to 52 weeks category and fewer in the remaining categories.

Table 15. Weeks Worked Distribution, 2016 ACS Content Test and 2014 SIPP Wave 1

Category	Test Percent (n=19,233)	Control Percent (n=19,676)	2014 SIPP Percent (n=28,993)
50 to 52 weeks	78.8 (0.5)	77.0 (0.5)	83.3 (0.1)
48 to 49 weeks	1.6 (0.1)	2.4 (0.2)	1.4 (0.1)
40 to 47 weeks	5.3 (0.2)	5.5 (0.2)	2.8 (0.1)
27 to 39 weeks	4.1 (0.2)	5.6 (0.2)	4.3 (0.1)
14 to 26 weeks	5.0 (0.3)	4.2 (0.2)	4.3 (0.1)
0 to 13 weeks	5.3 (0.2)	5.3 (0.2)	3.8 (0.1)
Total	100.0	100.0	100.0

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, 2014 Survey of Income and Program Participation
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding.

Research Question 7: *How does the proportion of full-time, year-round workers for each treatment compare with SIPP estimates?*

Table 16 shows the full-time, year-round rates from the Content Test treatments along with the 2014 SIPP rate. Neither treatment was within the confidence interval of SIPP's full-time, year-round rate.

Table 16. Full-Time Year-Round Rate, 2016 ACS Content Test and 2014 SIPP Wave 1

Category	Test Percent (n=18,851)	Control Percent (n=19,232)	2014 SIPP Percent (n=35,088)
Full-time, year-round	67.1 (0.6)	65.9 (0.6)	61.6 (0.2)

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, 2014 Survey of Income and Program Participation
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding.

5.5 Response Error

The GDR, IOI, and IOI L-fold were used to evaluate consistency between responses in the original interviews and the CFU interviews. These results show if the test version of *weeks worked* had more inconsistent responses than the control version.

Research Question 8: *Are the measures of response reliability (gross difference rate, index of inconsistency) better for the test treatment than for the control treatment?*

The GDR for a specific response category is the percent of inconsistent answers between the original interview and the reinterview (CFU). Table 17 shows the GDR's for the 2016 ACS Content Test treatments. The "50 to 52 weeks" category was not significantly lower between the test and control versions, nor were the two groups covering 26 weeks or less. However, the three groups covering 27 to 49 weeks had a measurably lower GDR for the test version, suggesting that the test group gave differing answers to *weeks worked* less often than the control group. This indicates that responses to the test version of the question are more reliable than the control version.

Table 17. Weeks Worked Gross Difference Rates

Response Category	Test Sample Size	Test GDR Percent	Control Sample Size	Control GDR Percent	Test Minus Control	Adjusted P-value
50 to 52 weeks	15,077	11.0 (0.5)	14,801	11.9 (0.5)	-0.8 (0.7)	0.33
48 to 49 weeks	304	2.0 (0.2)	468	3.0 (0.3)	-1.0 (0.4)	0.02*
40 to 47 weeks	942	6.3 (0.4)	1,198	7.7 (0.4)	-1.4 (0.6)	0.03*
27 to 39 weeks	821	5.7 (0.5)	1,198	7.1 (0.4)	-1.4 (0.6)	0.03*
14 to 26 weeks	958	4.8 (0.3)	862	5.4 (0.4)	-0.6 (0.5)	0.33
13 weeks or less	1,131	3.4 (0.4)	1,149	3.4 (0.2)	<0.1 (0.5)	0.51

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 in boldface indicate a significant difference based on a one-tailed t-test of H_0:*test* ≥ *control* at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method.

The Index of Inconsistency (Table 18) and Index of Inconsistency L-fold (Table 19) are measures of how much of the total estimate's variance is due to inconsistency in responses. As in the GDR, the full-year category did not have a measurably smaller IOI between the test and control versions. This again suggests that the reporting of full-year work is accurate and

consistent across treatments. The only category with a measurably smaller IOI was the "14 to 26 weeks" category, where the test version had greater consistency for this category.

Also of note, in both treatments the middle four categories – the categories covering 14 to 49 weeks worked – had an IOI above 50 percent. This result indicates a high amount of variability within these categories, and was not surprising given this question's historic difficulty for partial-year workers.

The Index of Inconsistency L-fold was smaller for the test treatment than the control treatment, suggesting that overall the test version was more consistent than the control version.

Table 18. Weeks Worked Index of Inconsistency

Response Category	Test Sample Size	Test IOI Percent	Control Sample Size	Control IOI Percent	Test Minus Control	Adjusted P-value
50 to 52 weeks	15,077	34.4 (1.7)	14,801	34.4 (1.4)	<0.1 (2.1)	1.00
48 to 49 weeks	304	71.0 (4.8)	468	83.1 (4.2)	-12.2 (6.1)	0.12
40 to 47 weeks	942	65.8 (3.5)	1,198	72.5 (2.6)	-6.7 (4.3)	0.24
27 to 39 weeks	821	67.7 (3.5)	1,198	64.3 (3.3)	3.4 (4.7)	1.00
14 to 26 weeks	958	52.3 (3.6)	862	67.0 (3.2)	-14.6 (4.4)	<0.01*
13 weeks or less	1,131	39.5 (4.0)	1,149	37.5 (2.7)	2.0 (5.2)	1.00

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 of H_0*:test ≥ control* at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method.

Table 19. Weeks Worked Index of Inconsistency L-Fold

Test L-fold	Control L-fold	Test Minus Control	P-value
47.0 (1.5)	50.0 (1.4)	-3.0 (1.9)	0.05*

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 of H_0*:test ≥ control* at the α=0.1 level.

5.6 Administrative Records and Earnings

5.6.1 Analysis Using Administrative Data

Administrative data provide a unique opportunity for 2016 ACS Content Test analysis. For *weeks worked*, LEHD data were the most relevant source to augment comparisons with 2016 ACS Contest Test data. As described in Section 2.4.6, Content Test respondents were matched with their LEHD administrative records through the PVS matching process.

Table 20 shows the percent of Content Test respondents whose Personal Identification Key number were identified during the PVS matching process. This percent is labeled as "Percent matched." Further, the percent of those who were matched *and* had at least one quarter of earnings data between quarter 2 of 2015 and quarter 1 of 2016 is reported as "Percent of matched

with any earnings during quarter 2 of 2015 and quarter 1 of 2016." Those who were matched and had reported earnings during the 2015 quarter 2 through 2016 quarter 1 period were used for LEHD analysis.

Table 20. LEHD Match and Earnings Rates for Those Aged 16+

Category	Test Percent (n=19,233)	Control Percent (n=19,676)
Percent matched	94.6	94.6
Percent of matched with any earnings during quarter 2 of 2015 and quarter 1 of 2016	84.8	84.5

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, Longitudinal Employer-Household Dynamics: 2015 quarter 2 through 2016 quarter 1

Match rates for the 2016 ACS Content Test sample to the LEHD data were comparable at 94.6 percent for both treatments. In addition, both treatments' reported earnings rates were comparable at 84.8 percent for the test group and 84.5 percent for the control group. Our analysis also compared match and earnings rates across age, sex, and industry of employment, again with no concerning difference between treatments. To protect privacy, those results are not included in this report.

These results suggest that reporting and matching to administrative data was not related to the treatment, and therefore the analysis using these records should be unbiased.

Research Question 9 compares respondents' Content Test data with their administrative records. Research Questions 10, 11, and 12 compare LEHD outcomes between treatments; we replaced Content Test responses with administrative data, and then tested for differences between treatments using these updated records (testing only those who were matched to LEHD data).

Research Question 9: *How do test and control distributions compare to Longitudinal Employer-Household Dynamics (LEHD) data?*

Table 21 compares the consistency between 2016 ACS Content Test responses and LEHD records. For each category, which represent the number of reported quarters with work in the LEHD, the control and test percentages are the proportion of respondents whose 2016 ACS Content Test *weeks worked* response is within their category's possible range of weeks worked. For example, one quarter of LEHD data means that the person worked somewhere between 1 and 13 weeks in the year, and two quarters means that the person worked between 2 and 26 weeks in the year. The control and test percentages are then the proportion of respondents who gave an answer for *weeks worked* that falls within their associated range.

For both treatments, those with four quarters of LEHD data had the largest percent of similar Content Test responses.[24] This is likely due to the relative prominence of year-round work. Between treatments, the consistency rate for the "4 Quarters" group was the only statistically

[24] Based on two-tailed testing of H_0*:"4 Quarters"=Next highest value.* P-value for test = <0.01; p-value for control = <0.01.

significant difference; however, the difference between them (0.4 percentage points) was not quantitatively large.

Note that because Table 21 only includes matched cases with valid nonmissing records from both sources, all control version responses are within the range for the "4 Quarters" group. This is because of the categorical nature of the control version, and every category contains at least 4 weeks. This led to the 100 percent reporting rate in the Control column. Test version responses between 0 and 3 weeks worked were the only responses that did not fall within the "4 Quarters" range.

Table 21. Percent of Respondents Reporting in Their LEHD Range

Number of Quarters With LEHD Reported Work	Test Percent (n=15,439)	Control Percent (n=15,727)	Test Minus Control	P-Value
4 Quarters	99.6 (0.1)	100.0 (<0.1)	-0.4 (0.1)	<0.01*
3 Quarters	40.5 (2.3)	44.8 (2.0)	-4.3 (3.0)	0.15
2 Quarters	47.2 (2.2)	42.9 (2.2)	4.3 (3.1)	0.17
1 Quarter	36.2 (2.9)	37.7 (3.0)	-1.4 (4.1)	0.73

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, Longitudinal Employer-Household Dynamics: 2015 quarter 2 through 2016 quarter 1.
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 of H_0:*test=control* at the α=0.1 level. Because this table only includes matched cases with valid nonmissing records from both sources, all control version responses are within the range for the "4 Quarters" group.

Research Question 10: *Is there a difference in the mean number of LEHD quarters worked between the test sample and the control sample?*

Comparing the mean LEHD-reported quarters worked between treatments can signal if respondents are interpreting the *weeks worked* questions differently. Table 22 shows the mean number of quarters reported in LEHD for each *weeks worked* category. In this case, there was no statistically significant difference between treatments for mean quarters worked.

Table 22. Mean Number of LEHD Quarters Worked, by Weeks Worked Category

Category	Test (n=18,200)	Control (n=18,615)	Test Minus Control	P-Value
50 to 52 weeks	3.3 (0.1)	3.3 (0.1)	<0.1 (0.1)	0.92
48 to 49 weeks	3.0 (0.1)	3.1 (0.1)	-0.1 (0.2)	0.57
40 to 47 weeks	3.0 (0.1)	3.0 (0.1)	<0.1 (0.1)	0.82
27 to 39 weeks	2.7 (0.1)	2.8 (0.1)	-0.1 (0.1)	0.27
14 to 26 weeks	2.3 (0.1)	2.2 (0.1)	0.1 (0.1)	0.39
0 to 13 weeks	1.6 (0.1)	1.6 (0.1)	<0.1 (0.1)	0.37

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, Longitudinal Employer-Household Dynamics: 2015 quarter 2 through 2016 quarter 1.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a two-tailed t-test of H_0:*test=control* at the α=0.1 level.

Research Question 11: *Is there a difference in mean LEHD earnings between the test sample and the control sample?*

The LEHD also served as one of two sources for testing earnings amounts between treatments. Earnings reported in the Content Test served as the second source. Results for Content-Test-based earnings are presented under Research Question 13.

Table 23 shows mean LEHD earnings between the test and control groups. These estimates are conditional on positive earnings. Mean earnings were measurably higher in the test treatment for the "27 to 39 weeks" and "0 to 13 weeks" groups by about $4,500 and $2,600, respectively. Because means are sensitive to outliers, median earnings were computed for each group, seen under Research Question 12.

Table 23. Mean LEHD Earnings, by Weeks Worked Category

Category	Test ($) (n=15,439)	Control ($) (n=15,727)	Test Minus Control ($)	P-Value
50 to 52 weeks	62,048 (1,104)	61,897 (1,028)	151 (1,508)	0.92
48 to 49 weeks	38,448 (3,663)	37,588 (2,397)	860 (4,377)	0.84
40 to 47 weeks	33,716 (3,918)	40,305 (3,595)	-6,589 (5,318)	0.21
27 to 39 weeks	25,202 (1,815)	20,696 (1,344)	4,506 (2,258)	0.05*
14 to 26 weeks	13,731 (1,261)	12,941 (1,031)	790 (1,629)	0.63
0 to 13 weeks	8,915 (998)	6,305 (504)	2,610 (1,118)	0.02*

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, Longitudinal Employer-Household Dynamics: 2015 quarter 2 through 2016 quarter 1.
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 of H_0:*test=control* at the α=0.1 level.

Research Question 12: *How do median LEHD earnings compare between the test sample and the control sample?*

Median earnings were measurably different for those in the "40 to 47 weeks" and "27 to 39 weeks" groups. For the "40 to 47 weeks" group, median earnings were lower in the test version by $3,450. Table 24 shows this difference. Conversely, the test median was higher in the "27 to 39 weeks" group by about $2,220.

Despite these differences, mean and median LEHD earnings appear comparable between treatments overall. These results support the requirement that any change to *weeks worked* should not adversely affect earnings estimates.

Table 24. Median LEHD Earnings, by Weeks Worked Category

Category	Test ($) (n=15,439)	Control ($) (n=15,727)	Test Minus Control ($)	P-Value
50 to 52 weeks	43,863 (592)	44,399 (585)	-536 (833)	0.52
48 to 49 weeks	21,014 (2,092)	26,142 (2,813)	5,128 (3,506)	0.14
40 to 47 weeks	18,954 (922)	22,404 (1,304)	-3,450 (1,597)	0.03*
27 to 39 weeks	13,531 (954)	11,323 (510)	2,208 (1,082)	0.04*
14 to 26 weeks	7,495 (444)	7,352 (526)	143 (689)	0.84
0 to 13 weeks	2,635 (382)	3,173 (299)	-538 (485)	0.27

Source: U.S. Census Bureau, 2016 American Community Survey Content Test, Longitudinal Employer-Household Dynamics: 2015 quarter 2 through 2016 quarter 1.
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 of H_0:*test=control* at the α=0.1 level.

5.6.2 Analysis Using ACS-based Earnings

This section compares earnings amounts reported in the 2016 ACS Content Test.

Research Question 13: *How do ACS-reported median earnings compare between test and control responses?*

Table 25 shows the median ACS-reported earnings for the test and control treatments. Median earnings were not statistically different between the treatments for any category. This suggested that the test version did not affect median earnings estimates.

Table 25. Median ACS Earnings, by Weeks Worked Category

Category	Test ($) (n=16,409)	Control ($) (n=16,846)	Test minus Control ($)	P-Value
50 to 52 weeks	43,713 (1,160)	45,064 (811)	-1,352 (1,265)	0.29
48 to 49 weeks	27,246 (3,436)	27,067 (2,192)	178 (3,894)	0.96
40 to 47 weeks	21,789 (1,148)	22,459 (1,340)	-670 (1,827)	0.71
27 to 39 weeks	13,717 (1,389)	12,653 (785)	1,063 (1,478)	0.47
14 to 26 weeks	7,155 (510)	7,184 (372)	-29 (685)	0.97
0 to 13 weeks	2,295 (109)	2,128 (88)	167 (145)	0.25

Source: U.S. Census Bureau, 2016 American Community Survey Content Test.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding.
Significance was tested based on a two-tailed t-test of H_0:*test=control* at the α=0.1 level.

6 CONCLUSIONS AND RECOMMENDATIONS

For the 2016 ACS Content Test, the Census Bureau tested changes to the *Number of Weeks Worked* question. The purpose of this test was to study if more specificity in the number of weeks worked could be gained without a loss in data quality. Overall, empirical results suggest that changing part B of *weeks worked* to an open-ended response, along with question text changes to part A and part B, will allow for greater specificity without adversely affecting data quality, and in fact might improve data consistency.

The results presented in this report support recommending the test version of *weeks worked* for implementation on the ACS. All decision criteria outlined in Section 3 were met, save for the test version's increase in the item missing data rate for part B (at 4.3 percent versus 3.3 percent). This was not a major concern because write-in responses generally result in a higher missing data rate.

The estimate of full-time, year-round workers was higher in the test treatment, and earnings estimates between treatments were not significantly different overall. In addition, the test version did not adversely impact the distribution of weeks worked or response reliability. Administrative data comparisons also suggested that recommending the test version of *weeks worked* over the control version will not adversely affect data quality for the number of weeks worked.

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 also like to acknowledge the following individuals for their contributions to the *Number of Weeks Worked* report: **Rebecca Chenevert** (Social, Economic, and Housing Statistics Division (SEHSD)) for oversight and management during the planning, data collection, and analysis phases of the 2016 ACS Content Test; **Braedyn Kromer** (Food and Drug Administration (FDA), formerly Census Bureau) for oversight and management during the planning and data collection phases, as well as preparing question proposals and the Research Evaluation and Analysis Plan; **Alfred Gottschalck** (SEHSD) for *weeks worked* question proposals and cognitive testing oversight; **Nicole Scanniello** (SEHSD) for guidance and representation during all phases of the Content Test; **Andrew Roberts** (Population Division (POP)) for statistical review and analysis; **Agnes Kee** (American Community Survey Office (ACSO)), **Anthony Tersine** (Decennial Statistical Studies Division (DSSD)), **David Raglin** (ACSO), **Dorothy Barth** (DSSD), **Elizabeth Poehler** (DSSD), **Jennifer Cheeseman Day** (SEHSD), **Jennifer Ortman** (ACSO), and **Kathryn Cheza** (ACSO) for their guidance in preparing this report and for reviewing it; **Garrett Schmitt** and **Hugette Sun** from the Bureau of Labor Statistics (BLS); the Westat cognitive interviewing implementation and analysis team, including **Darby Steiger**, **Jennifer Anderson**, **Jasmine Folz**, **Mariel Leonard**, and **Martha Stapleton**; and the Census Bureau field representatives and telephone interviewers, without who's dedication the preparation of this report would be impossible.

8 REFERENCES

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

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

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

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

Goetz, C., Hyatt, H., McEntarfer, E., & Sandusky, K. (2015). The Promise and Potential of Linked Employer-Employee Data for Entrepreneurship Research. *NBER Working Paper No. 2163*.

Holder, K., & Raglin, D. (2007). *Evaluation Report Covering Employment Status.* Retrieved June 23, 2015, from http://www.census.gov/acs/www/Downloads/library/2007/2007_Holder_01.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

Kromer, B., & Howard, D. (2012). *Comparison of ACS and CPS Data on Employment Status.* Retrieved on June 25, 2015 from http://www.census.gov/people/laborforce/publications/ACS-CPS_Comparison_Report.pdf

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

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. Bureau of Labor Statistics. (2013). *Labor Force Statistics from the Current Population Survey -- Household Data Annual Averages (2013) -- 8. Employed and unemployed full- and part-time workers by age, sex, race, and Hispanic or Latino ethnicity*. Retrieved June 26, 2015 from U.S. Bureau of Labor Statistics Official Web Site: http://www.bls.gov/cps/aa2013/cpsaat08.htm

U.S. Census Bureau. (1966). *1960 Censuses of Population and Housing: Procedural History*. Washington, D.C.: U.S. Government Printing Office. Retrieved on June 24, 2015 from http://www2.census.gov/prod2/decennial/documents/12110770.zip

U.S. Census Bureau. (2014). *American Community Survey Design and Methodology (January 2014).* Retrieved on 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/

Wagner, D., & Layne, M. (2014). *The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software.* U. S. Census Bureau. Retrieved May 2017 from https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf

Appendix A: Unit Response Rates Supplemental Table

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

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

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

Note: 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 of H_0*:test=control* at the α=0.1 level. The weighted response rates account for initial sample design as well as CAPI subsampling.