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

November 30, 2023

2023 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES # ACS23-RER-13

MEMORANDUM FOR Donna M. Daily
Chief, American Community Survey Office

From: David Waddington *Stephanie Galvin, acting Division chief 11/30/23*
Chief, Social, Economic, and Housing Statistics Division

Prepared by: Clayton Gumber *CG*
Brian Mendez-Smith *BMS*
Robert Munk *RM*
Social, Economic, and Housing Statistics Division

Lindsay Longsine *LL*
Michael Risley *MR*
Decennial Statistical Studies Division

Subject: 2022 American Community Survey Content Test Evaluation Report: Labor Force

Attached is the 2022 American Community Survey (ACS) Content Test report for Labor Force. This report presents the methods and results of the test for a revised version of the Labor Force question(s).

If you have any questions about this report, please contact Clayton Gumber at (301) 763-4131.

cc:

Lynda Laughlin	SEHSD	Sharon Stern		Broderick Oliver
Robert Munk		Lauren Contard	DSSD	Samantha Spiers
Hyon Shin		Peter Massarone		acso.re.workgroup.list

This page is intentionally blank

American Community Survey Research and Evaluation Program

November 30, 2023

2022 American Community Survey Content Test Evaluation Report: Labor Force

FINAL REPORT

Clayton Gumber
Brian Mendez-Smith
Robert Munk
Social, Economic, and Housing Statistics Division
Lindsay Longsine
Michael Risley
Decennial Statistical Studies Division

This page is intentionally blank

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Overview of the 2022 Amercian Community Survey Content Test

The U.S. Census Bureau conducted the 2022 American Community Survey (ACS) Content Test, from September through December of 2022. The 2022 ACS Content Test tested the wording, format, and placement of proposed new ACS questions and proposed revisions of current ACS questions for potential inclusion in the ACS data collection instruments. The tested questions came from 10 topics. This report presents the results of this field test for Labor Force.

In preparation for the 2022 Content Test, the Census Bureau, in consultation with the Office of Management and Budget (OMB) and the Interagency Council on Statistical Policy Subcommittee on the ACS, determined which proposals solicited from over 25 federal agencies would be tested in 2022. Approved proposals for new content or changes to existing content were tested according to the ACS content change process, which includes cognitive testing and field testing.

The 2022 ACS Content Test consisted of a nationally representative sample of 120,000 housing unit addresses, excluding Puerto Rico, Alaska, and Hawaii. The sample, which was independent of production ACS, was divided evenly among three treatments, a Control treatment and two test treatments.

Like production ACS, the data collection for the 2022 ACS Content Test was conducted in two phases: a self-response phase, which lasted up to nine weeks, followed by a nonresponse followup phase, conducted via Computer-Assisted Personal Interviewing (CAPI). The CAPI operation lasted about one month. For households where we received a response in the original Content Test interview, a Content Follow-Up telephone reinterview was conducted to measure response error.

Overview of the Tested Labor Force Content

The primary goal of the Labor Force Statistics Branch for the 2022 ACS Content Test was to test a modified reference period that better aligns with existing sources of administrative data. Currently, the ACS reference period for all labor force content is the “past 12 months”, while the proposed reference period would encompass the preceeding calendar year, for example “2021”. By incorporating administrative data into the ACS, we can potentially reduce respondent burden and improve imputation quality. A decision regarding changes to the reference period has been delayed until analyses using administrative data can be run to compare the Control and test treatments.

The secondary goal of the Labor Force Statistics Branch for the 2022 ACS Content Test was to test wording changes to the Labor Force series of questions. Due to the interconnected nature of the labor force content, as well as other ACS content, changes to the question wording required more finesse than simply replacing the reference period. Additionally, cognitive testing revealed opportunities to improve question wording. There were three versions of the Labor Force series of questions used in the Content Test: production for the Control treatment, a Version 1 test treatment, and a Version 2 test treatment. A detailed description of the changes made to each version of the labor force questions can be found at 1.3.3, but a brief description follows.

Both Version 1 and Version 2 modified the questions to enhance clarity and were identical in the internet and CAPI instruments. The paper instrument, however, differed in that Version 1 added question instructions in the form of a bullet point list, while Version 2 added question instructions beneath the relevant question.

Two comparisons emerge from these content changes. First, a comparison of the Version 1 and Control treatments provides insight into whether the question and instruction modifications enhanced question clarity. Second, a comparison of the Version 1 and Version 2 test treatments' mail instruments provides insight into whether the bullet point instructions were superior to instructions beneath the questions.

Benchmarks

To assess whether the proposed changes to the ACS content would significantly affect employment estimates or distributions, each labor force item was compared to measures from two other surveys administered by the Census Bureau: the Annual Social and Economic Supplement of the the Current Population Survey (CPS ASEC) and the Survey of Income and Program Participation (SIPP). Estimates from these surveys were used as nominal benchmarks to the ACS Content Test due to their similarity in survey universe and measures. Items benchmarked include the proportion of full-time, year-round workers, the civilian employment to population ratio, distribution of annual weeks worked, and distribution of average annual hours worked per week.

Overall, results indicated marginal differences between surveys once changes to the ACS content were incorporated. Comparisons of the proportion of full-time, year-round workers showed that the ACS content test approximated the results of the CPS ASEC estimates, but both CPS ASEC and ACS content test were nominally higher than the SIPP proportion of full-time, year-round workers results. The analyses of employment to population ratio indicated that the test group estimates approximated both the CPS estimates and the SIPP estimates. Together these findings provide evidence of satisfactory external validity of the civilian employment to population ratio and proportion of full-time, year-round workers.

Item Missing Rate

Version 1 Treatment versus Control Treatment

The item missing rates were significantly lower overall for Version 1 than the Control treatment for the Worked Every Week, Weeks Worked, and Hours Worked questions (see Tables 21, 23, and 25). These differences were the result of lower missing rates in the Version 1 test treatment for the internet and mail modes. A noteworthy exception is the item missing rates for the Weeks Worked question, which had a significantly higher missing rate in the CAPI mode of the Version 1 test treatment. This finding suggests that the clarifications and additional instructions in the Version 1 test treatment decreased item missing rates. We, however, cannot rule out that the changes to the reference period were the cause of lower item missing rates.

Version 1 Test Treatment versus Version 2 Test Treatment

The Version 1 test treatment had significantly higher item missing rates on the mail mode than the Version 2 test treatment for the Worked Every Week question and the Hours Worked Question (see Tables 22 and 26). This finding suggests that the bulleted instructions led to higher missing rates on the mail mode than instructions beneath the question.

Response Distribution

Weeks Worked Distribution

The proportion of full-time, year-round workers in the Version 1 test treatment was significantly higher than Control (see Table 30). Version 1 also had a significantly smaller share of workers working 13 weeks or less than Control both overall and for the CAPI mode (see Table 32). To better assess whether these changes constitute a change in quality, we will run analyses with administrative data to assess the quality of the Version 1 and Control treatments.

Hours Worked

Among full-time, year-round workers, the Version 1 test treatment had significantly more workers who worked 40 hours and significantly less workers working 41-49 hours or 60 or more hours compared to Control (see Table 36). The difference may be a result of recall bias—workers may be more likely to report working 40 hours per week as a result of the longer recall period.

Increased Heaping in the Version 1 Treatment

For part-time or part-year workers, the Version 1 test treatment for the Weeks Worked question had significantly more heaping on numbers ending with 0 or 5 for the internet mode compared to the Control treatment (see Table 33). Furthermore, for part-time or part-year workers, the Version 1 treatment for the Hours Worked question had significantly more heaping on numbers ending with 0 or 5 than Control for the mail mode (see Table 41). These findings suggest heaping is more common in the Version 1 test treatment, which may be the result of a longer recall period.

Response Reliability

Response reliability was measured by comparing gross difference rates (GDR). The GDR assesses how often respondents changed their answers in the Content Follow-up Interview. We found that the Version 1 test treatment generally had lower GDR rates than Control across the labor force content, indicating more consistent responses. Specifically, the Version 1 test treatment had lower GDR rates for (1) the share of people 16 and over who worked over 5 years ago or never worked, (2) for those who reported working 13 weeks or less overall and for the Internet mode, (3) for those who reported working 50 weeks or more for the CAPI mode, and (4) for the overall number of hours worked each week.

Recommendations

Keep All Clarification of the Labor Force Version 1 and Version 2 Test Treatments

We found that both the Version 1 and Version 2 test treatments had lower item missing response rates than the Control treatment. In addition, the Version 1 test treatment had lower GDR rates for several response categories and no response categories having higher GDR rates than Control. While we cannot rule out that the lower item missing and GDR rates are a result of the changed reference period, we believe the added clarifications suggested by cognitive test participants and sponsors are the likely primary driver. Because of this, we recommend that all clarifying changes made in Version 1 be kept.

Use the Version 2 Test Treatment's Mail Questions

We found that the Version 2 treatment outperformed Version 1 in the mail mode, suggesting that placing the instructions beneath the questions was better for respondents than the bulleted list. Specifically, the Version 2 test treatment had lower item missing response rates in the mail mode, and no statistical differences were observed in the response distributions between the two treatments.

Delay Changing the Reference Period

At this time the administrative records matching process for the ACS is not ready to be used in production. Additionally, no comparisons of the Control and test treatments responses has been made to assess whether the changed reference period affected response quality. We would, therefore, like to reconsider the reference period change after administrative data has been used to examine whether the changes in the response distribution of the Version 1 test treatment corresponds to a change in response quality. We anticipate starting these analyses in Fall 2023 when administrative data that are appropriate for such comparisons become available.

1 BACKGROUND

The U.S. Census Bureau conducted the 2022 American Community Survey (ACS) Content Test from September to December of 2022. The 2022 ACS Content Test tested the wording, format, and placement of proposed new ACS questions and proposed revisions of current ACS questions for potential inclusion in the ACS data collection instruments. The questions came from these ten ACS topics, three of which, Sewer, Electric Vehicles, and Solar Panels are new:

- Household Roster
- Sewer
- Electric Vehicles
- Solar Panels
- Supplemental Nutrition Assistance Program (SNAP)
- Educational Attainment
- Health Insurance Coverage
- Disability
- Labor Force
- Income

This report presents the results of the field test for Labor Force.

1.1 Proposals for New and Revised ACS Questions

In June 2018, the Census Bureau solicited proposals for new or revised ACS content from over 25 federal agencies. For new questions, the proposals explained why these data were needed and why other data sources that provide similar information were not sufficient. Proposals for new content were reviewed to ensure that the requests met a statutory or regulatory need for data at small geographic levels or for small populations.

The Census Bureau, in consultation with the Office of Management and Budget (OMB) and the Interagency Council on Statistical Policy Subcommittee on the ACS, determined which proposals moved forward. Approved proposals for new content or changes to current content were tested via the ACS content change process. This process included cognitive testing and field testing. An interagency team consisting of Census Bureau staff and representatives from other federal agencies participated in development and testing activities.

In accordance with OMB's Standards and Guidelines for Statistical Surveys (OMB, 2006) and the Census Bureau's Statistical Quality Standards (U.S. Census Bureau, 2022a), the Census Bureau conducted cognitive interviewing to pretest survey questions prior to field testing or implementing the questions in production.

1.2 Cognitive Testing

For the 2022 ACS Content Test, the Census Bureau contracted with Research Triangle Institute (RTI) International to conduct three rounds of cognitive testing.[1] Cognitive interviews were conducted virtually, in English and Spanish.[2] In the first round of cognitive testing, each topic tested one or two versions of the question. Based on the results of the first round, wording modifications to the questions were made and one or two versions per topic were tested in the second round. The interagency team used the results of both rounds of cognitive testing to recommend question content for the field test. For more information on the cognitive testing procedures and results from rounds one and two, see RTI International (2022a).

The third round of cognitive testing was conducted in Puerto Rico and in Group Quarters (GQ), as the 2022 ACS Content Test did not include field testing in these areas. Cognitive interviews in Puerto Rico were conducted in Spanish; GQ cognitive interviews were conducted in English. For more information on the cognitive testing procedures and results from the third round, see RTI International (2022b).

Three topics included in the cognitive testing were not included in the field test: Homeowners Association or Condominium Fees, Home Heating Fuel, and Means of Transportation to Work. For the most part, the changes to these questions are expected to either impact a small population or result in a small change in the data that would not be detectable in the Content Test. The subject matter experts recommended that cognitive testing was sufficient for these questions and that field testing was not necessary; the Interagency Council on Statistical Policy Subcommittee on the ACS agreed with this recommendation. Content changes for these topics will be implemented in production ACS in 2024.

1.3 Field Testing Labor Force in the 2022 ACS Content Test

1.3.1 Justification for Inclusion of Labor Force in the Content Test

As the successor to the Decennial Census “long” form the ACS has included questions about respondents’ employment since its inception. Moreover, the ACS is periodically updated to collect the most relevant and up-to-date information on a person’s work experience. The current version of the labor force question series was last updated in 2019 after being tested in the 2016 ACS Content Test (Smith, Howard, and Risley, 2017).[3]

[1] For each test topic, subcommittees were formed to develop question wording and research requirements for cognitive testing. The subcommittees included representation from the Census Bureau and other federal agencies.

[2] Cognitive testing interviews were conducted virtually due to the COVID-19 pandemic. Interviews were attempted by videoconferencing first and were moved to phone interviews if there were technical problems with Skype or MS Teams.

[3] The Number of Weeks Worked question was adjusted to allow an open-ended response following the 2016 ACS Content Test. No changes were made to the reference period for the Number of Weeks Worked and Number of Hours Worked questions, which has been “during the past 12 months” since 1996.

In an effort to continue improving the ACS data quality and potentially reduce respondent burden, the labor force content was included as part of the 2022 ACS Content Test. In addition to wording changes that sought to improve question clarity, the reference period for all labor force items on the ACS Version 1 and Version 2 test treatment questionnaires was changed from "the past 12 months" to the prior calendar year. The goal of this test was to determine if the new ACS reference period corresponded better with existing aministrative records. By aligning the ACS reference period with administrative records, it eases the use of administrative records in data production, which could improve the quality of ACS estimates.

Administrative Records and the ACS

Administrative record linkage has been proposed as one solution to declining survey response rates and other issues facing the ACS (Jarmin, 2019). According to the "Agility in Action" report series, the U.S. Census Bureau could apply administrative data to the ACS to reduce respondent burden and eliminate redundant questions asked on the survey. However, the report notes that the quality of administrative records might vary in their ability to enhance or supplement the ACS. "In some cases, the quality of the records may be more accurate than the respondent's recollection (e.g., W2 information for wages)" while in other cases "we may not be able to decipher whether data from records are superior or inferior to response data" (Census Bureau, 2015). Regardless of how administrative records are ultimately incorporated into ACS, they would be more able to improve the ACS if they also coincided within the same reference period (Ortman, Pharris-Ciurej, and Clark 2018). Changing the ACS reference period to align with the calendar year will remove the existing temporal barrier that hinders administrative record linkage in ACS production.

There is precedence for a prior year reference period for the Number of Weeks Worked and Number of Hours Worked questions. The 2000 Census long form included the questions "LAST YEAR, 1999, did this person work at a job or business at any time?", "How many weeks did this person work in 1999?", and "During the weeks WORKED in 1999, how many hours did this person usually work each WEEK?". Income questions in the 2000 Census long form also referred to 1999. The proposed change in reference period, therefore, paves the way for improved ACS data and returns to a past convention that has already been tried and tested.[4]

Research also underscores the feasibility and potential benefits of using administrative records to increase the coverage and quality of employment data in the ACS. O'Hara, Bee, and Mitchell (2016) found that "Eighty-eight percent of PIK-linked ACS respondents aged 18 to 64 had one or

[4] One noteworthy difference between the 2000 Census long form questions and the ACS Labor Force series of questions is the ordering of the Industry and Occupation (I&O) series of questions. On the 2000 long form, the I&O questions appeared before the Labor Force series of questions. On the ACS, the I&O questions appear after the Labor Force series of questions and before the Income questions. Because the Class of Worker and I&O questions are in between two sets of questions being tested, Martinez et al(2023) analyzed whether the tested changes affected the data collected for those questions in a separate report.

more information return. The most common type of form is the W-2, reflecting wages and salaries. Nearly all PIK-linked ACS respondents aged 65 and older had at least one information return (98 percent), reflecting income received from SSA (92 percent) and pension distributions on Form 1099-R (62 percent)."[5] Further, "Among those whose ACS wages are imputed to be zero, 46,108 people (29 percent) do have a W-2, while among those imputed positive wages, 98,434 people (23 percent) have no W-2. Among those with self-reported ACS wages, more than 93 percent have one or more matching W-2s."

Administrative Records and the SIPP

Existing research on the Survey of Income and Program Participation (SIPP) has already established the correspondence between survey and administrative employment measures and demonstrated the benefits of using administrative records to increase the quality of survey employment data.[6] Chenevert et al. (2016) compare SIPP to administrative records from the Social Security Administration's Detailed Earnings Record, which contains all W-2 forms issued to and Form 1040 Schedule SE (self-employment) documents filed by SIPP sample members. They find a 92.1 percent rate of agreement between survey and administrative measures of employment from 2009 through 2012. Klee et al. (2019) argue that much of this disagreement when administrative data indicates employment stems from individuals who report no employment when administrative data indicate that this person was in the left tail of the earnings distribution. This finding is consistent with the hypothesis that respondents fail to report work that delivers little pay.

To the extent that this hypothesis is true, we would expect to find a similar effect in the ACS, especially because of its longer recall period (12 months in current ACS versus 4 months in 2008 SIPP). While we cannot verify whether ACS respondents receiving little pay in administrative data report no work to the survey given the current misalignment of ACS reference periods and tax years, changing the reference period would allow us to evaluate whether this trend is present in ACS data as well.

1.3.2 Cognitive Testing Development for Labor Force

In preparation for an eventual change of reference period, in 2016, the Census Bureau contracted with Westat to cognitively test the Labor Force and Income series of questions. Although we were primarily testing the ability of respondents to recall Labor Force and Income information from the prior year as compared to the past 12 months, the testing also revealed some areas of the questions that could be improved to provide greater clarity for respondents (Steiger, D., Robins, D., and Stapleton, M., 2017). Using the recommendations from the Westat

[5] The term PIK refers to a protected identification key, or an anonymous identifier used to link survey respondents with other forms of administrative data.

[6] The SIPP collects monthly income and weekly employment data that can be aggregated to create a calendar year reference period.

report, RTI International further cognitively tested versions of the questions specifically for this Content Test (RTI International, 2022a). The improvements made to the questions, as a result of cognitive testing, are outlined below.

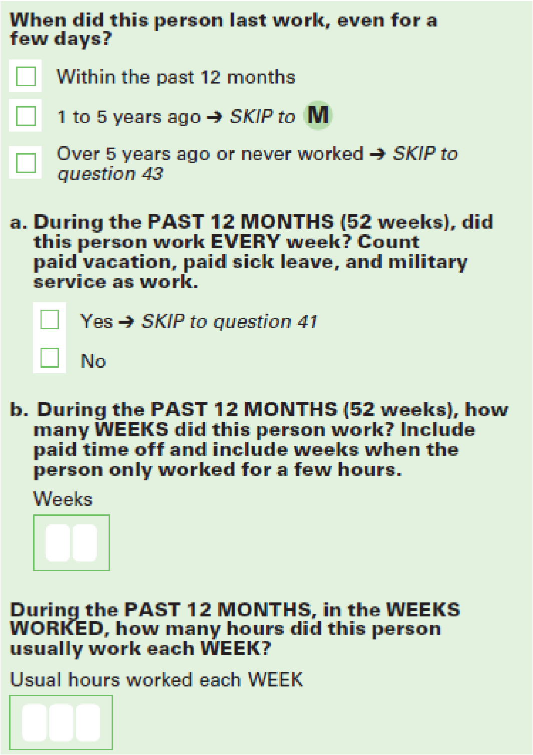
1.3.3 Question Content

Aside from the change of reference period to the questions in the Labor Force series, we also made the following changes:

- Added “for pay” to the question, “When did this person last work, even for a few days?” Some cognitive testing participants who worked in a volunteer capacity got confused by the question, so we added “for pay” for clarity.
- Added a new question, “In 2021, did this person work for pay, even for a few days?” to set up the proper universe for the new reference period to the Number of Weeks Worked and Number of Hours Worked questions.
- Added the instruction, “Include all jobs for pay” to the questions on weeks worked.
- Modified the instructions for Version 1 of the paper questionnaire. The paper instrument for version 1 gives the instructions in a bullet point format to test if this format makes the instructions clearer and easier to understand.
- Added “for at least one day” to the question about how many weeks were worked.
- Added the instructions “Include all jobs for pay and military service.” to the Hours Worked question.

Control, Version 1, and Version 2 of each question are shown as they appeared on the paper questionnaire. Automated versions of the questionnaire had the same content formatted accordingly for each mode. Version 1 and Version 2 differed only on paper; the internet and CAPI questions were identical. Version 1’s paper instrument used a note and bullet points to draw readers’ attention to important instructions. Version 2’s paper instrument kept all instructions under the questions; this version is more similar to the way the tested questions were asked on the internet and in CAPI.

Figure 1. Control Version of Labor Force Questions (Paper)



When did this person last work, even for a few days?

□ Within the past 12 months

□ 1 to 5 years ago ➔ *SKIP to* **M**

□ Over 5 years ago or never worked ➔ *SKIP to question 43*

a. During the PAST 12 MONTHS (52 weeks), did this person work EVERY week? Count paid vacation, paid sick leave, and military service as work.

□ Yes ➔ *SKIP to question 41*

□ No

b. During the PAST 12 MONTHS (52 weeks), how many WEEKS did this person work? Include paid time off and include weeks when the person only worked for a few hours.

Weeks

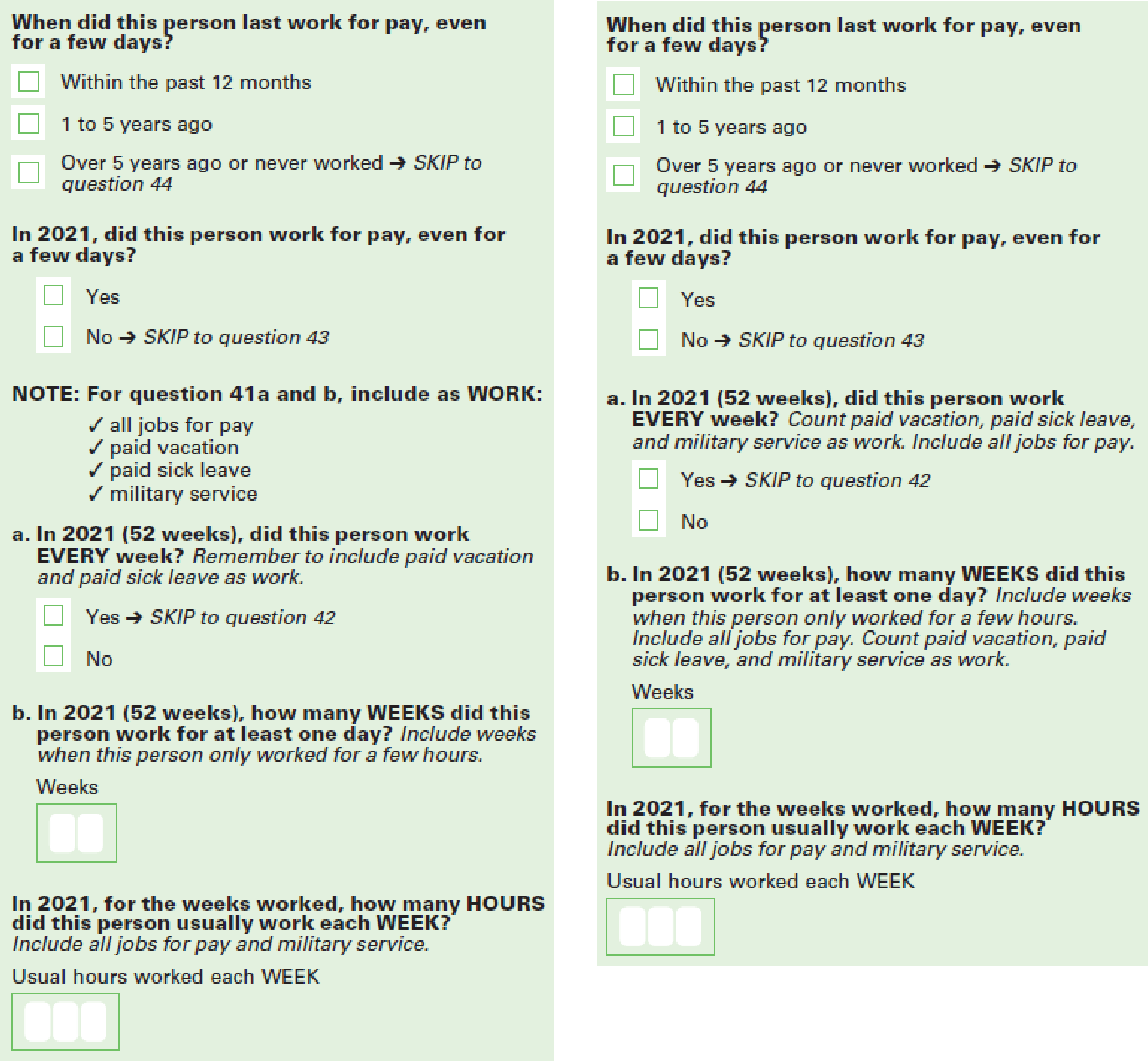
□□

During the PAST 12 MONTHS, in the WEEKS WORKED, how many hours did this person usually work each WEEK?

Usual hours worked each WEEK

□□□

Figure 2. Version 1 (Left) and Version 2 (Right) of Labor Force Questions (Paper)



When did this person last work for pay, even for a few days?

□ Within the past 12 months

□ 1 to 5 years ago

□ Over 5 years ago or never worked → *SKIP to question 44*

In 2021, did this person work for pay, even for a few days?

□ Yes

□ No → *SKIP to question 43*

NOTE: For question 41a and b, include as WORK:

✓ all jobs for pay
✓ paid vacation
✓ paid sick leave
✓ military service

a. In 2021 (52 weeks), did this person work EVERY week? *Remember to include paid vacation and paid sick leave as work.*

□ Yes → *SKIP to question 42*

□ No

b. In 2021 (52 weeks), how many WEEKS did this person work for at least one day? *Include weeks when this person only worked for a few hours.*

Weeks

□□

In 2021, for the weeks worked, how many HOURS did this person usually work each WEEK?
Include all jobs for pay and military service.

Usual hours worked each WEEK

□□□

When did this person last work for pay, even for a few days?

□ Within the past 12 months

□ 1 to 5 years ago

□ Over 5 years ago or never worked → *SKIP to question 44*

In 2021, did this person work for pay, even for a few days?

□ Yes

□ No → *SKIP to question 43*

a. In 2021 (52 weeks), did this person work EVERY week? *Count paid vacation, paid sick leave, and military service as work. Include all jobs for pay.*

□ Yes → *SKIP to question 42*

□ No

b. In 2021 (52 weeks), how many WEEKS did this person work for at least one day? *Include weeks when this person only worked for a few hours. Include all jobs for pay. Count paid vacation, paid sick leave, and military service as work.*

Weeks

□□

In 2021, for the weeks worked, how many HOURS did this person usually work each WEEK?
Include all jobs for pay and military service.

Usual hours worked each WEEK

□□□

1.3.4 Research Questions

The questions examined for this research are presented below.

RQ1. How does the proportion of full-time, year-round workers (age 16+) for the combined test treatments compare with CPS ASEC estimates?

RQ2. How does the annual, civilian employment-to-population ratio (age 16+) for the combined test treatments compare with CPS ASEC estimates?

RQ3. How does the distribution of weeks worked (age 16+) for the combined test treatments compare with CPS ASEC estimates?

RQ4. How does the distribution of hours worked (age 16+) for the combinted test treatments compare with CPS ASEC estimates?

RQ5. How does the proportion of full-time, year-round workers (age 16+) for the combined test treatments compare with SIPP estimates?

RQ6. How does the annual, civilian employment-to-population ratio (age 16+) for the combined test treatments compare with SIPP estimates?

RQ7. How does the distribution of weeks worked (age 16+) for the combined test treatments compare with SIPP estimates?

RQ8. How does the distribution of hours worked (age 16+) for the combined test treatments compare with SIPP estimates?

RQ9. For the When Last Worked question, is there a difference in item missing data rates between Control and Version 1? Version 1 and Version 2 mail responses?

RQ10. For the Work for Pay in 2021 question, is there a difference in item missing data rates between Version 1 and Version 2 for mail responses?

RQ11. For the Number of Weeks Worked questions, is there a difference in item missing data rates between Control and Version 1? Version 1 and Version 2 for mail responses?

RQ12. For the Hours Worked each Week question, is there a difference in item missing data rates between Control and Version 1? Version 1 and Version 2 for mail responses?

RQ13. For the When Last Worked question, is there a difference in response distributions between the Control and Version 1? Version 1 and Version 2 mail responses?

RQ14. Is there a difference in the proportion of people working in 2021 between the two test treatments (Version1 vs Version 2)?

RQ15. Does the proportion of full-time, year-round workers differ between Control and Version 1? Version 1 and Version 2 mail responses?

RQ16. Among part-time or part-year workers, does the distribution of Weeks Worked responses by mode differ between Control and Version 1? Version 1 vs. Version 2 mail responses?

RQ17. Does the distribution of Hours Worked responses by mode and full-time, year-round worker status differ between Control and Version 1? Version 1 vs. Version 2 mail responses?

RQ18. Is the Gross Difference Rate (GDR) of When Last Worked different between the Control and Version 1?

RQ19. Is the GDR of Number of Weeks Worked different between Control and Version 1?

RQ20. Is the GDR of Number of Hours Worked different between Control and Version 1?

In the Research and Analysis Plan for the 2022 Content Test, there were 28 research questions for the Labor Force topic (U.S. Census Bureau, 2022c). Research questions 21 and 22 involved comparing the results to administrative records from the 2022 Longitudinal Employer-Household Dynamics data. We have since learned that this data will not be available in time for this report and thus those analyses will be conducted at a later time. Research questions 23 through 28 involved analyzing the effect of changing the Labor Force questions on the Class of Worker, Industry, and Occupation questions on the ACS. Those analyses are in a separate report (Martinez et al., 2023).

2 METHODOLOGY

2.1 Sample Design

The 2022 ACS Content Test consisted of a national sample of roughly 120,000 housing unit addresses, excluding Puerto Rico, Alaska, and Hawaii (due to cost constraints, only stateside housing units were included). The sample was independent of the ACS production sample; however, the sample design for the Content Test was largely based on the ACS production sample design, with some modifications to meet the test objectives. The ACS production sample design is described in Chapter 4 of the ACS and Puerto Rico Community Survey (PRCS) Design and Methodology report (U.S. Census Bureau, 2022b).

The sample design modifications included stratifying addresses into high and low self-response areas, oversampling addresses from the low self-response areas to ensure equal response from both strata, and selecting an initial sample of addresses, followed by a nearest neighbor method for selecting the remaining addresses for sample. The high and low self-response strata were defined based on ACS self-response rates from the 2018 and 2019 panels at the tract level.

In the sample selection process, we selected an initial sample of 40,000 addresses, then selected the two nearest neighbors for each initially selected address. If possible, we selected nearest neighbors that were in both the same content test sampling stratum as well as the same state, county, and sub-county area as the initially selected address. In total, three samples were selected, one for the Control treatment and two for the two test treatments. These three treatments are shown in Table 1.

The Control treatment contained production questions and questions from the three new topics: Solar Panels, Electric Vehicles, and Sewer. The Test treatment contained a test version question for all topics except Household Roster. Two of the new topics, Solar Panels and Sewer, only had one version of the test question; therefore, the same question was asked in the Control and test treatments. The other new topic, Electric Vehicles, had two versions; one was asked in the Control and Roster Test treatments and the other in the Test treatment.

The primary purpose of the Roster Test treatment was to test the household roster test question separately since changes in the amount and types of people included in the household could impact the results of person-level topics. Therefore, the analyses for Test Version 2 of the Health Insurance Coverage, Labor Force, and Income questions could have been impacted by these changes. However, it was determined that the additional information gained from testing an additional version of the topics in the Roster Test treatment was worth the risk.[7]

[7] We examined differences in key household and person characteristics among the Control and Roster Test treatments to explore any indication of bias in the Health Insurance Coverage, Labor Force, and Income analyses. See Spiers et al. (2023) for more information.

Table 1. Questions by Treatment

Topic	Control Treatment	Test Treatment	Roster Test Treatment
Household Roster	Production	Production	Test Version
Solar Panels	Test Version	Test Version	Test Version
Electric Vehicles	Test Version 1	Test Version 2	Test Version 1
Sewer	Test Version	Test Version	Test Version
Educational Attainment	Production	Test Version	Production
Health Insurance Coverage	Production	Test Version 1	Test Version 2
Disability	Production	Test Version	Production
SNAP	Production	Test Version	Test Version†
Labor Force	Production	Test Version 1	Test Version 2
Income	Production	Test Version 1	Test Version 2

† The SNAP Test Version was in both test treatments to align with Labor Force and Income that also had a reference period change to the previous calendar year.

2.2 Data Collection

The 2022 ACS Content Test occurred in parallel with data collection activities for the September 2022 ACS production panel. Data collection for production ACS data consists of two main phases: an approximately two-month self-response data collection phase and a one-month follow-up phase.

During the self-response phase, addresses in sample are asked to self-respond by internet or mail. The Census Bureau sends addresses in sample up to five mailings to encourage self-response. This operation is followed by a one-month Computer-Assisted Personal Interviewing (CAPI) operation, where Census Bureau field representatives attempt to complete a survey for a sub-sample of the remaining nonresponding addresses.

The following data collection protocols for the 2022 ACS Content Test remained the same as production ACS:

- Data were collected using the self-response modes of internet (in English and Spanish) and paper questionnaires for the first and second month of data collection.
- In the third month of data collection, a sub-sample of nonresponding addresses were selected for CAPI.

- During CAPI, Census Bureau field representatives conducted interviews in person and over the phone.
- Self-response via internet or paper was accepted throughout the three-month data collection period.

The following data collection protocols for the 2022 ACS Content Test differed from production ACS:

- There were no paper versions of the 2022 ACS Content Test questionnaires in Spanish.[8]
- If respondents called Telephone Questionnaire Assistance (TQA) and opted to complete the survey over the phone, the interviewers conducted the survey using the production ACS questionnaire.[9] Since the TQA interviews did not include test questions, they were excluded from the analysis of the 2022 ACS Content Test.
- The 2022 ACS Content Test did not include the Telephone Failed-Edit Follow-Up (FEFU) operation. In production, this operation follows up on households that provided incomplete information on the form or reported more than five people on the roster of a paper questionnaire.[10]
- The 2022 ACS Content Test used a telephone reinterview component to measure response reliability or response bias (depending upon the ACS topic). This telephone reinterview operation is discussed in Section 2.3 below.

For detailed information about ACS data collection procedures, consult the ACS and PRCS Design and Methodology Report (U.S. Census Bureau, 2022b).

2.3 Content Follow-Up Operation

To measure response reliability or response bias, a Content Follow-Up (CFU) reinterview was attempted with every household with an original Content Test interview that met the CFU eligibility requirements. Among the requirements were that the household must be occupied, and the household must have a valid telephone number. See the CFU requirements document for the complete list of eligibility requirements (Spiers, 2021a).

[8] In 2019, 412 Spanish questionnaires were mailed back out of all mailable cases. Based upon this rate, we projected that only 8 Spanish questionnaires would be mailed back in the 2022 Content Test, which would not be cost-effective.

[9] The interviewer did not know which treatment the caller was in and therefore administered the production questionnaire. In 2019, less than one percent (0.6%) of cases responded by TQA and had no other response in a different mode. Based upon this rate, we projected about 744 TQA-only responses would be excluded from the 2022 ACS Content Test analysis.

[10] The information obtained from the FEFU improves accuracy in a production environment but confounds the evaluation of respondent behavior in the Content Test environment. For paper questionnaires, where the household size is six or more (up to 12), we only collected name, age, and sex of these additional persons, but not detailed information as we do in the FEFU operation for ACS production.

2.3.1 Content Test Follow-Up Protocol

As in previous ACS Content Tests, a case was sent to the CFU operation no sooner than two weeks (14 calendar days) after the original interview and had to be completed within three weeks after being sent to the CFU. This timing attempted to balance two competing needs: (1) to minimize the possibility of real changes in answers due to a change in life circumstances between the two interviews; (2) to minimize the possibility of the respondent repeating their previous answer based on their recollection of the original interview response, rather than considering the most appropriate answer.

All CFU reinterviews were conducted by telephone. At the first contact with a household, interviewers asked to speak with the original respondent. If that person was not available, interviewers scheduled a callback at a time when the original respondent was expected to be available. If this respondent could not be reached at the time of the second contact, the interviewer requested to speak with any other eligible household member (a household member who is 15 years or older). CFU reinterviews for the Content Test were conducted in either English or Spanish.

The CFU data collection instrument included the questions being tested for the 2022 ACS Content Test and some production ACS questions for context. It also included questions on public assistance from the 2022 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) to measure response bias in the income from the public assistance question.

The CFU collected an independent household roster by re-asking the Household Roster questions along with Relationship, Sex, Age, and Date of Birth. The remaining CFU questions were only asked of the original household roster members. Only the Control and Roster Test panels collected an independent household roster. The Test panel used the original household roster to ask housing and detailed person questions.[11]

2.3.2 Content Test Follow-Up for Labor Force

We asked the respondent the same questions in the Content Test Follow Up (CFU) interview as they were asked in the original interview. This helped us to determine the response reliability of the Labor Force series of questions. A survey question has good response reliability if respondents tend to answer the question consistently.

Reliability was assessed by the proportion of persons with a difference between the original response and the CFU response. The write-in value of the original response was compared to the CFU response. The comparison provided insight into how responses change. However, if

[11] The Test panel did not need to collect an independent household roster. The independent roster was needed to calculate the response reliability metrics for the Household Roster topic, which only used data from the Control and Roster Test treatments.

responses changed only slightly, there may be limited impact on estimates in official products. Consequently, we also assessed reliability by comparing whether the original response and the CFU response fell into the same bin of weeks worked and hours worked that are used in ACS data products.

For the Number of Weeks Worked question, reliability was assessed by comparing the proportion of persons in different weeks worked categories between the original response and the CFU response, with the categories being 1 to 13 weeks, 14 to 26 weeks, 27 to 39 weeks, 40 to 47 weeks, 48 to 49 weeks, and 50 to 52 weeks.

For the Number of Hours Worked question, reliability was assessed by comparing the proportion of persons in different hours worked categories between the original response and the CFU response, with the categories being 1 to 14 hours, 15 to 34 hours, and 35 or more hours.

2.4 Analysis Metrics

The sample addresses for the Control and test treatments were selected in a manner so that their response propensities and response distributions (on particular characteristics) would be the same. Similar distributions allow us to conclude that any difference in the metrics used to analyze Labor Force is attributable to differences in the wording and format. We tested these unit-level assumptions in both the original interview and the CFU interview. See Section 2.4.1 for details. The metrics that we used to evaluate Labor Force are presented in Section 2.4.2.

For the 2022 ACS Content Test, typical production ACS edits were not made because the primary concern of this test was how changes to existing questions and differences between versions of new questions affected the unaltered responses provided directly by respondents. For this reason, responses were not imputed either. A few edits were applied to the non-topic data, such as calculating a person’s age based on his or her date of birth, but such edits were minimal. [12]

All estimates from the ACS Content Test were weighted. The final content test weights took into account the initial probability of selection (the base weight) and CAPI sub-sampling. The weights used in the CFU analysis also included an adjustment for CFU non-response.[13]

Comparisons between the Control and test versions of Labor Force were conducted using a two-tailed t-test at the α=0.1 level of significance. The Content Test sample size was chosen to

[12] This only refers to edits made to the data sets before analysis. During the analysis phase, additional edits, such as collapsing categories, were made based on the needs of the individual question.

[13] The Content Test weight creation process does not include all the steps followed in the ACS, including the noninterview adjustment for the original interview and calibration to housing unit and population controls (see U.S. Census Bureau, 2022b, Chapter 11). For more information on the 2022 Content Test weighting procedure, see Risley and Oliver (2022) and Keathley (2022).

provide enough statistical power (0.80) to detect a difference in the gross difference rates (measuring differences in adds and deletes from the household roster) of at least two percentage points between the Control and Roster Test groups for the Household Roster question.[14] In statistical tests involving multiple comparisons, we controlled for the overall Type I error rate by adjusting the resulting p-values using the Hochberg method (Hochberg, 1988).[15]

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, 2022b, Chapter 12). We calculated the variance for each rate and difference using the formula below. The standard error of an estimate (X_0) is the square root of the variance:

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

where:

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

2.4.1 Unit-Level Analysis

The unit response rate is important, as it provides an indication of the quality of the survey data. As part of our analysis, we examined unit-level (i.e., address-level) responses for the Control and test treatments in the original interviews and CFU reinterviews. These results are provided in a separate report (Spiers et al., 2023).[16]

2.4.2 Topic-Level Analysis

To evaluate the changes to Labor Force questions, we calculated a variety of metrics, presented in Sections 2.4.2.1 through 2.4.2.6.

2.4.2.1 Benchmarks

To roughly gauge the accuracy of the responses to the Labor Force questions, we compared employment statistics derived from the 2022 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and the 2022 Survey of Income and Program Participation (SIPP) to combined estimates from the test treatments (Version 1 and Version 2). The benchmarks are only nominal because important differences in the survey methodoliges prevent direct statistical comparisons.

[14] See Section 2.4.2.4 for the definition of Gross Difference Rate.

[15] Use the MULTTEST Procedure in SAS®.

[16] As part of the 2022 ACS Content Test, we analyzed respondent burden. The results of this analysis is contained in Virgile et al. (2023).

2.4.2.2 Item Missing Data Rates

To measure nonresponse to the Labor Force questions, we calculated item missing data rates, the proportion of eligible persons for which a required response is missing. A high item missing data rate can be indicative of a question that lacks clarity, is sensitive, or is simply too difficult to answer.

We calculated the rates for the following questions: When Last Worked, Worked in 2021, Number of Weeks Worked, and Number of Hours Worked. For comparisons of the Version 1 and Version 2 treatments we only calculated item missing data rates for mail responses, since the only difference between the two versions was on the paper questionnaire. For comparisons with the Control Treatment, we calculated item missing data rates from all response modes.

We compared item missing data rates via two-tailed t-tests.

2.4.2.3 Response Distributions

To assess how changes to the Labor Force questions affected the resulting estimates, we compared the response distributions of the Control and the test versions of the Labor Force questions. We calculated the response distributions as the proportion of valid responses in a category to all valid responses. For Version 1 vs. Version 2 comparisons, we only analyzed mail responses.

For When Last Worked, we calculated and compared the response distributions of the categories: within the past 12 months, 1 to 5 years ago, and over 5 years ago or never worked.

We also calculated and compared the proportions of:

- People working in 2021.
- Full-time, year-round workers.
- Part-time or part-year workers.
- Hours worked for full-time, year-round workers.

Comparisons were made using a Rao-Scott chi-square test that checks for a significant difference between two sample distributions (Rao & Scott, 1987). If the chi-square test indicated a significant difference between the distributions, we tested for significant differences in the individual category proportions using two-tailed t-tests.

2.4.2.4 Response Reliability

Survey responses are subject to error. Response error occurs for a variety of reasons, such as flaws in the survey design, misunderstanding of the questions, misreporting by respondents, and interviewer effects. For the 2022 ACS Content Test, response error was measured through response reliability or response bias, not both. This was done to reduce respondent burden and breakoffs during the CFU operation. For Labor Force, we measured response error using response reliability.

A survey question has good response reliability if respondents tend to answer the question consistently. For the 2022 ACS Content Test, we measured response reliability for a given question by comparing the responses to this question in the original interview to the responses to this same question in the CFU reinterview.

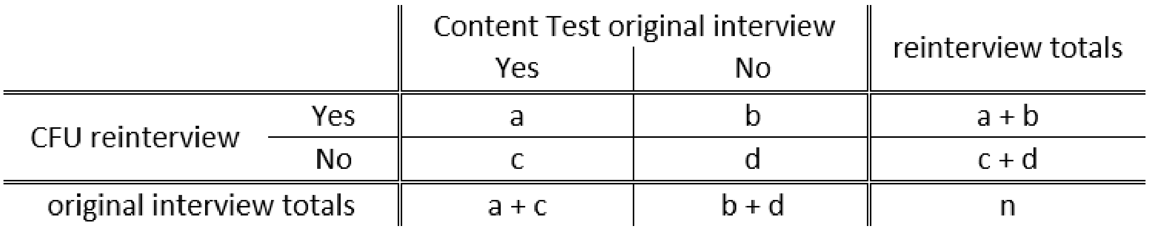
Re-asking the same question of the same respondent allows us to measure simple response variance, using the following measures:

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

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

In Table 2, “Yes” indicates that the unit is in the category of interest, according to the response from either the original interview or the CFU reinterview. “No” indicates that the unit is not reported to be in the category.

Table 2. Original Interview and CFU Reinterview Counts for Calculating GDR, IOI, and NDR



		Content Test original interview		reinterview totals
		Yes	No	
CFU reinterview	Yes	a	b	a + b
	No	c	d	c + d
original interview totals		a + c	b + d	n

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

a = units in category for both interview and reinterview
b = units not in category for original interview, but in category for reinterview
c = units in category for original interview, but not in category for reinterview
d = units in category for neither interview nor reinterview
n = total units in the universe = a + b + c + d

These counts were weighted to make them more representative of the population.

We calculated the GDR for this response category as:

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

The IOI and IOI$_L$ metrics are described below but were not calculated for the Labor Force questions.

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.

The SV is the part of total variance resulting from the differences between 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 total variance is due to inconsistency in responses, as measured by SRV. A preliminary definition of the IOI is:

$$IOI = \left(\frac{SRV}{SRV + SV}\right) \times 100$$

We can estimate SRV using the GDR, but also need to estimate the denominator (i.e., total variance) in this expression. Based on previous studies, the estimate we use for total variance is:

$$SRV + SV = \frac{p_1 q_2 + p_2 q_1}{2}$$

where:

$$p_1 = \frac{a + c}{n} = \text{original interview proportion in category}$$

$$q_1 = 1 - p_1 = \frac{b + d}{n} = \text{original interview proportion not in category}$$

$$p_2 = \frac{a + b}{n} = \text{CFU proportion in category}$$

$$q_2 = 1 - p_2 = \frac{c + d}{n} = \text{CFU proportion not in category}$$

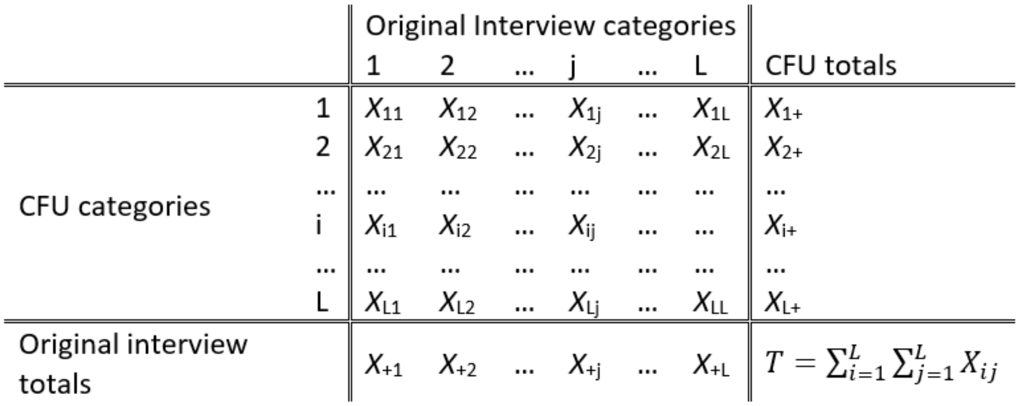
In comparing relative reliability (or response error) between treatments, if the response categories are essentially the same, then we looked at the differences in the GDR and IOI for each response category. We tested the significance of these differences, using two-tailed t-tests.

If the response categories did not match up exactly between the compared treatments, we either collapsed response categories to form equivalent categories for comparison, or we conducted comparisons for the response categories where it made sense.

So far, we have only discussed response reliability with respect to single response categories. If a question has three or more response categories (or “comparison categories” in cases where it is necessary to collapse some response categories for comparison), we also measured the overall response reliability of a question using the L-fold index of inconsistency, $IOI_{L.}$ We looked at the difference in IOI_L between treatments and tested for significance as with the single category measures.

Suppose a question has L response categories. Let X_{ij} be the weighted count of sample units (households or persons) for which we have CFU responses in category *i* and original interview responses in category *j*. Here, both *i* and *j* range from 1 to L. Table 3 shows a cross-tabulation of the original interview and CFU results for a generic analysis topic. Note that if L = 2, then Table 3 is equivalent to Table 2.

Table 3. Cross-Tab of Original Interview and CFU Results: Questions with Response Categories



		Original Interview categories						
		1	2	…	j	…	L	CFU totals
CFU categories	1	X_{11}	X_{12}	…	X_{1j}	…	X_{1L}	X_{1+}
	2	X_{21}	X_{22}	…	X_{2j}	…	X_{2L}	X_{2+}
	…	…	…	…	…	…	…	…
	i	X_{i1}	X_{i2}	…	X_{ij}	…	…	X_{i+}
	…	…	…	…	…	…	…	…
	L	X_{L1}	X_{L2}	…	X_{Lj}	…	X_{LL}	X_{L+}
Original interview totals		X_{+1}	X_{+2}	…	X_{+j}	…	X_{+L}	$T = \sum_{i=1}^{L} \sum_{j=1}^{L} X_{ij}$

Now define the following proportions:

$$p_{ij} = \frac{X_{ij}}{T}$$

$$p_{+j} = \frac{X_{+j}}{T}$$

$$p_{i+} = \frac{X_{i+}}{T}$$

The IOI_L is calculated as

$$\mathrm{IOI_L} = \frac{1 - \sum_{i=1}^{L} p_{ii}}{1 - \sum_{i=1}^{L} (p_{i+} p_{+i})} \times 100$$

It can be shown that the IOI_L is a weighted sum of the L category IOI values (Biemer, 2011), but this formula is easier for calculation.

The IOI metrics can be biased if the parallel measures assumption is violated, i.e., if the errors in the original interview and CFU reinterview are positively or negatively correlated (Biemer, 2011). We checked this assumption by testing if the net difference rate (NDR) is significantly different from zero. The NDR is the difference between the original interview proportion of positive responses (“Yes” or in the category of interest) and the CFU proportion of positive responses. The NDR is calculated as follows:

$$NDR = (p_1 - p_2) \times 100 = \left(\frac{c - b}{n}\right) \times 100$$

If the NDR is significantly positive or negative, the assumption of “parallel measures” necessary for the SRV and IOI to be valid is not satisfied (Biemer, 2011). In these situations, we use the following adjustment of the IOI, developed by Flanagan (2001):

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

3 DECISION CRITERIA

Before field testing Labor Force questions, a team of subject matter experts identified and prioritized which of the research questions presented in Section 1.3.4, would determine which version of Labor Force questions would be recommended for inclusion in the ACS. The decision criteria for Labor Force are presented in Table 4 and Table 5.

Table 4. Decision Criteria for Labor Force: Changing the Reference Period while also Modifying Question and Instructional Wording

Priority	Research Questions	Decision Criteria
1	1, 2, 3, 4, 5, 6, 7, 8	Benchmark Comparisons: We expected the distribution of full-time, year-round workers, civilian employment-to-population ratio, weeks worked, and hours worked for both Version 1 and Version 2 to be reasonably consistent with benchmarks from other data sources.
2	9, 10, 11, 12	Item Missing Data Rates: We hoped to see no difference or a decrease in item nonresponse for When Last Worked, Number of Weeks Worked, and Number of Hours Worked when Version 1 was compared to Control.
3	18, 19, 20	Response Reliability: We hoped to see no difference or an increase in response reliability for When Last Worked, Number of Weeks Worked, and Number of Hours Worked when Version 1 was compared to Control.
4	13, 14, 15, 16, 17	Response Distributions: We hoped to see the same response proportion of When Last Worked, Worked Full-time, Year-round, Part-time Workers, and Hours Worked when Version 1 was compared to Control.
5	21, 22*	Other metrics: We expect to see closer similarities in Quarters Worked when the test versions are compared to LEHD administrative data, as opposed to Control vs. LEHD administrative data.

*This was originally part of the decision criteria, however LEHD data was not available in time for this report.

Table 5. Decision Criteria for Labor Force: Paper Questionnaire Design— Version 1 vs. Version 2 mail responses

Priority	Research Questions	Decision Criteria
1	13, 14, 15, 16, 17	Response Distributions: We hoped to see the same response proportion of When Last Worked, Worked Full-time, Year-round, Part-time Workers, and Hours Worked when Version 1 was compared to Version 2.
2	9, 10, 11, 12	Item Missing Data Rates: We hoped to see no difference or a decrease in item nonresponse for When Last Worked, Number of Weeks Worked, and Number of Hours Worked when Version 1 was compared to Version 2.

4 ASSUMPTIONS AND LIMITATIONS

4.1 Assumptions

- The sample addresses for the Control and test treatments were selected in a manner so that their response propensities and response distributions would be the same. This assumption of homogeneity allowed us to conclude that any difference between treatments is attributable to differences in wording and format. See Section 5 for more details.

- There was no difference between treatments in mail delivery timing or subsequent response time. The treatments had the same sample size and used the same postal sort and mailout procedures. Previous research indicated that postal procedures alone could cause a difference in response rates at a given point in time between experimental treatments of different sizes, with response for the smaller treatments lagging (Heimel, 2016).

- We assume that the frequency of real changes in answers due to a change in life circumstances between the original interview and CFU were similar between treatments.

4.2 Limitations

- GQs were not included in the sample for the 2022 ACS Content Test. The results of the Content Test may not extend to GQ populations.

- Housing units from Alaska, Hawaii, and Puerto Rico were not included in the sample for the 2022 ACS Content Test. The results of the Content Test may not extend to the housing unit population in these areas.

- The paper questionnaire was only available in English and was not available in Spanish like in production. The Content Test results related to the English paper questionnaire may not extend to Spanish paper questionnaire.

- For paper questionnaires, where the household size was six or more (up to 12), we only collected name, age, and sex of these additional persons. Detailed information for these persons in ACS production are collected in the FEFU operation. We did not include the FEFU operation because the information collected from it improves accuracy and could confound respondent behavior in the Content Test environment.
- We did not have response data for some partial internet responses (179 cases) due to a server issue. These cases were excluded from the analyses.
- TQA responses were excluded from the analysis of the 2022 ACS Content Test response data because survey responses completed via the TQA operation were only conducted using the ACS production data collection instrument.

- CAPI interviewers were assigned 2022 ACS Content Test cases as well as regular production cases. The potential risk of this approach is the introduction of a cross-contamination or carry-over effect among Control and test treatments and production due to the same interviewer administering multiple versions of the same question item (despite their training to read questions verbatim).

- Due to budget constraints, the CAPI workload could not exceed 28,000 housing units. This workload was less than what was subsampled originally because we over-sampled addresses in low response areas. Limiting the CAPI workload caused an increase in the variances for the analysis metrics used.

- The CFU reinterviews were conducted by phone only, whereas the original interviews were completed online, by mail, by phone in CAPI, and in person in CAPI. Hence, some of the differences observed between the original interviews and the CFU interviews may be the result of mode effect.

- Not all households who provided a response in the original interview were eligible for the CFU reinterview (see Section 2.3 for more information). As a result, 2.5 percent (standard error 0.2) of households from the original Control interviews, 2.5 percent (standard error 0.2) of households from the original Test interviews, and 3.0 percent (standard error 0.2) of households from the original Roster Test interviews were not eligible for the CFU reinterview. These rates were not significantly different between treatments (chi-square p-value 0.11).

- We reinterviewed the same person who responded in the original interview when possible, but accepted interviewing a different person from the same household after two unsuccessful attempts at reaching the original person. Therefore, differences in results between the original interview and CFU reinterview for these cases could partly be from different people answering the questions. We interviewed a different household member in CFU for 7.3 percent (standard error 0.4) of CFU Control cases, 9.4 percent (standard error 0.5) of CFU Test cases, and 8.5 percent (standard error 0.5) of CFU Roster Test cases. These rates were significantly different between treatments (chi-square p-value 0.01) with the rate of CFU Test cases (t-test p-value <0.01) and CFU Roster Test cases (t-test p-value 0.04) being significantly higher than the rate of CFU Control cases.

- We examined potential differences between CFU respondents and nonrespondent within some socioeconomic and demographic characteristics because there were differences in the 2016 CFU reinterview (Spiers, 2021b). For all treatments combined, there were significant differences between CFU respondents and nonrespondents for *household size, tenure, age, race, Hispanic origin, language of original interview response,* and *high and low response areas*. These differences are similar to the ones found in the 2016 CFU (Spiers, 2021b).

- The 2022 ACS Content Test did not include the production weighting adjustments for unit nonresponse or population controls which are designed to minimize nonresponse and under-coverage bias. As a result, any estimates derived from the Content Test data did not provide the same level of inference as the production ACS and cannot be compared to production estimates.

5 RESULTS

This section of the report presents the results of various metrics used to evaluate the Labor Force questions. The comparisons presented assume homogeneity of the response distributions for the three treatments, prior to the field test. We tested this assumption via unit-level (i.e., address level) analyses.

Original Interview

The overall unit response rates were not significantly different between the treatments, nor were the response rate portions by mode. When looking at response rates within high and low response areas, a couple of modal comparisons were significant, but these results did not appear in the overall comparisons.

Additionally, when examining demographic and socioeconomic distributions, none of the response distributions were significantly different between treatments. When looking at distributions among self-responses and CAPI responses, only the distribution for race among CAPI responses for the Control and Test treatments (treatment with Version 1 question) was significantly different. This distribution difference showed up in the Other Race Only category.

We are confident there were no statistically significant differences that would impact original interview comparisons between treatments for the Labor Force questions.

CFU Reinterview

For the Labor Force topic, we only compared Control vs Version 1 when analyzing response reliability. The overall unit response rates were not significantly different between the Control and Test (Version 1) treatments, nor were the response rate portions by mode.

When examining demographic and socioeconomic distributions, none of the overall response distributions were significantly different between the Control and Test treatments. When looking at distributions among self-responses and CAPI responses, only the distribution for tenure among self-responses for the Control and Test treatments was significantly different. This distribution difference showed up in the Owned Free and Clear category.

We are confident there were no statistically significant differences that would impact response error analyses between treatments for the Labor Force questions.

For more information about the unit-level analyses, see (Spiers et al., 2023).

5.1 Benchmark Results for Labor Force

Benchmarking the labor force content from the ACS Content Test to the CPS ASEC and the SIPP provides evidence of whether questions accurately measure a particular concept. However, due to methodological and design differences between the ACS , CPS, and SIPP we can only compare benchmark results nominally. Results of each survey comparisons are presented below.

Note: the universe for the benchmarks comparisons is all persons 16 and older, but the universe for all other research questions is 15 and older. The ACS asks the labor force questions to all respondents 15 and older, however, CPS asks labor questions for 16 and older. For comparions made between ACS, CPS, and SIPP (research questions 1 through 8), we defined the same universe to match CPS. For all other questions, we defined the universe to those who were asked the ACS labor force questions (15 and older).

RQ1. How does the proportion of full-time, year-round workers (age 16+) for the combined test treatments compare with CPS ASEC estimates?

Table 6 provides the percentage of full-time, year-round workers for the combined test treatments (Version 1 and Version 2 combined) and the estimate from the 2022 CPS ASEC.

Table 6. Full-time, Year-round Workers – 2022 ACS Content Test vs 2022 CPS ASEC

Rate	Content Test	CPS ASEC
Full time, year-round workers	69.5 (0.4)	70.4 (0.3)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC| DRB No. CBDRB-FY23-ACSO003-B0065
Note: Standard errors are in parentheses.

The Content Test estimate of the proportion of full-time, year-round workers approximates the CPS ASEC estimate.

RQ2. How does the annual, civilian employment-to-population ratio (age 16+) for the combined test treatments compare with CPS ASEC estimates?

Table 7 provides the annual, civilian employment-to-population ratio for the test treatments and the estimate from the 2022 CPS ASEC.

Table 7. Civilian Employment-to-Population Ratio – 2022 ACS Content Test vs CPS ASEC

Rate	Content Test	CPS ASEC
Civilian employment-to-population ratio	62.0 (0.4)	62.9 (0.2)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Standard errors are in parentheses.

The Content Test estimate and CPS ASEC estimate of the civilian employment-to-population ratio are nominally similar.

RQ3. How does the distribution of weeks worked (age 16+) for the combined test treatments compare with CPS ASEC estimates?

Table 8 provides the distribution of weeks worked for the test treatments and the distribution from the 2022 CPS ASEC.

Table 8. Distribution of Weeks Worked – 2022 ACS Content Test vs CPS ASEC

Weeks Worked	Content Test	CPS ASEC
13 weeks of less	5.7 (0.1)	4.7 (0.2)
14 to 26 weeks	4.9 (0.2)	5.7 (0.1)
27 to 39 weeks	3.5 (0.1)	4.1 (0.1)
40 to 47 weeks	4.4 (0.2)	4.6 (0.1)
48 to 49 weeks	1.4 (0.1)	1.7 (0.1)
50 or more weeks	80.1 (0.3)	79.2 (0.2)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Standard errors are in parentheses.

The Content Test and CPS ASEC distributions of the number of weeks worked are nominally similar. We are not concerned with the changes tested to the weeks worked question for the ACS Content Test.

Table 9 collapses the distribution of weeks worked from the 2022 ACS Content Test and the 2022 CPS ASEC into two categories: ends in 0 or 5, and any other positive number. While both the ACS and CPS ASEC allow respondents to provide an integer response for the number of weeks worked per year, the differing reference periods between surveys allows for a rough analysis of recall bias answer in the ACS Content Test. Specifically, the categories used in this analysis are intended to show the effects of response heaping, where responses to a question may be an estimation rather than specific number. More broadly, response heaping could indicate potential issues with recall over long periods of time (such as 12 months, or the past calendar year).

Table 9. Distribution of Weeks Worked – 2022 ACS Content Test vs CPS ASEC

Weeks Worked	Content Test	CPS ASEC
End with 0 or 5	9.9 (0.2)	7.9 (0.2)
Other positive number	90.1 (0.2)	92.1 (0.2)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Standard errors are in parentheses.

The Content Test and CPS ASEC distributions of the number of weeks worked are nominally similar when collapsed into two categories.

RQ4. How does the distribution of hours worked (age 16+) for the combined test treatments compare with CPS ASEC estimates?

Table 10 provides the distribution of hours worked for the test treatments and the distribution from the 2022 CPS ASEC.

Table 10. Distribution of Hours Worked – 2022 ACS Content Test vs CPS ASEC

Hours Worked	Content Test	CPS ASEC
0 to 14 hours	6.4 (0.1)	3.5 (0.1)
15 to 34 hours	16.0 (0.3)	14.6 (0.2)
35 to 39 hours	5.8 (0.2)	6.2 (0.2)
40 hours	49.1 (0.4)	55.8 (0.3)
41 to 49 hours	7.2 (0.2)	6.1 (0.2)
50 to 59 hours	9.7 (0.2)	9.2 (0.2)
60 or more hours	5.7 (0.2)	4.6 (0.2)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Standard errors are in parentheses.

The Content Test and CPS ASEC distributions of the number of hours worked are nominally similar. We are not concerned with the changes tested to the hours worked question for the ACS Content Test.

Table 11 similarly collapses the distribution of hours worked in the 2022 ACS Content Test and the 2022 CPS ASEC into two categories: ends in 0 or 5 and any other positive number. Like the comparisons for the weeks worked question, these categories show how responses to the hours worked question may represent an estimation since they center on rounded numbers. Large differences between surveys could indicate an issue when respondents are asked to recall hours worked over long periods of time.

Table 11. Distribution of Hours Worked - 2022 ACS Content Test vs CPS ASEC

Hours Worked	Content Test	CPS ASEC
End with 0 or 5	86.0 (0.3)	87.6 (0.2)
Other positive number	14.0 (0.3)	12.4 (0.2)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Standard errors are in parentheses.

The Content Test and CPS ASEC distributions of the number of hours worked are nominally similar when collapsed into two categories.

RQ5. How does the proportion of full-time, year-round workers (age 16+) for the combined test treatments compare with SIPP estimates?

Table 12 provides the percentage of full-time, year-round workers for the test treatments and the estimate from the 2022 SIPP.

Table 12. Full-time, Year-round Workers – 2022 ACS Content Test vs 2022 SIPP

Rate	Content Test	SIPP
Full time, year-round workers	69.5 (0.4)	62.8 (0.4)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065

Note: Standard errors are in parentheses.

The Content Test estimate of the proportion of full-time, year-round workers is nominally higher than the SIPP estimate.

RQ6. How does the annual, civilian employment-to-population ratio (age 16+) for the test treatments compare with SIPP estimates?

Table 13 provides the annual, civilian employment-to-population ratio for the test treatments and the estimate from the 2022 SIPP.

Table 13. Civilian Employment-to-Population Ratio – 2022 ACS Content Test vs 2022 SIPP

Rate	Content Test	SIPP
Civilian employment-to-population ratio	62.0 (0.4)	63.4 (0.3)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065

Note: Standard errors are in parentheses.

The Content Test estimate and SIPP estimate of the civilian employment-to-population ratio are nominally similar.

RQ7. How does the distribution of weeks worked (age 16+) for the combined test treatments compare with SIPP estimates?

Table 14 provides the distribution of weeks worked for the test treatments and the distribution from the 2022 SIPP.

Table 14. Distribution of Weeks Worked – 2022 ACS Content Test vs 2022 SIPP

Weeks Worked	Content Test	SIPP
13 weeks of less	5.7 (0.1)	4.0 (0.1)
14 to 26 weeks	4.9 (0.2)	4.9 (0.2)
27 to 39 weeks	3.5 (0.1)	5.1 (0.2)
40 to 47 weeks	4.4 (0.2)	5.0 (0.2)
48 to 49 weeks	1.4 (0.1)	3.4 (0.2)
50 or more weeks	80.1 (0.3)	77.8 (0.4)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065xxxxxxx

Note: Standard errors are in parentheses.

The Content Test and SIPP distributions of the number of weeks worked are nominally similar. We are not concerned with the changes tested to the weeks worked question for the ACS Content Test.

Table 15 collapses the distribution of weeks worked into two categories (ends in 0 or 5 and any other positive number) and the distribution from the 2022 SIPP.

Table 15. Distribution of Weeks Worked – 2022 ACS Content Test vs 2022 SIPP

Weeks Worked	Content Test	SIPP
End with 0 or 5	9.9 (0.2)	5.0 (0.2)
Other positive number	90.1 (0.2)	95.0 (0.2)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065

Note: Standard errors are in parentheses.

The comparison is nominally similar between the ACS and SIPP. Given the differences in how weeks worked are collected—the SIPP collects weeks worked for up to seven jobs—and the SIPP is known to suffer from seam bias, we are not concerned with the changes from the content test.

RQ8. How does the distribution of hours worked (age 16+) for the combined test treatments compare with SIPP estimates?

Table 16 provides the distribution of hours worked for the test treatments and the distribution from the 2022 SIPP.

Table 16. Distribution of Hours Worked - 2022 ACS Content Test vs 2022 SIPP

Hours Worked	Content Test	SIPP
0 to 14 hours	6.4 (0.1)	6.2 (0.2)
15 to 34 hours	16.0 (0.3)	17.8 (0.3)
35 to 39 hours	5.8 (0.2)	8.1 (0.4)
40 hours	49.1 (0.4)	39.3 (0.3)
41 to 49 hours	7.2 (0.2)	10.5 (0.3)
50 to 59 hours	9.7 (0.2)	9.7 (0.2)
60 or more hours	5.7 (0.2)	8.1 (0.2)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Standard errors are in parentheses.

The comparison is nominally similar between the ACS and SIPP.

Table 17 collapses the distribution of hours worked into two categories (ends in 0 or 5 and any other positive number) and the distribution from the 2022 SIPP.

Table 17. Distribution of Hours Worked - 2022 ACS Content Test vs 2022 SIPP

Hours Worked	Content Test	SIPP
End with 0 or 5	86.0 (0.3)	67.6 (0.4)
Other positive number	14.0 (0.3)	32.9 (0.4)

Source: U.S. Census Bureau, 2022 American Community Survey Content Test and 2022 CPS ASEC | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Standard errors are in parentheses.

Due to methodological and design differences between the ACS and SIPP, we can only compare these rates nominally. The distributions are quite different, but because the SIPP collects start month and week for up to seven jobs heaping is less likely to be an issue.

5.2 Item Missing Data Rate Results for Labor Force

RQ9. For the When Last Worked question, is there a difference in item missing data rates between Control and Version 1? Version 1 and Version 2 mail responses?

Table 18 provides the item missing data rates of When Last Worked for the Control and Version 1 questions of the Content Test. We calculated the rates overall and by mode. We compared the item missing data rates using a two-sided t-test.

Table 18. Item Missing Data Rates for the When Last Worked Question - Control vs Version 1

Mode	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	2.3 (0.2)	2.3 (0.2)	<0.1 (0.3)	0.94
Internet	0.5 (0.1)	0.3 (0.1)	0.2 (0.2)	0.60
Mail	10.0 (0.8)	10.6 (0.7)	-0.6 (1.2)	0.94
CAPI	0.6 (0.2)	0.6 (0.2)	<0.1 (0.3)	0.94

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method.

There was no statistically significant difference in the item missing data rates between Control and Version 1 for the When Last Worked question.

Table 19 provides the item missing data rates When Last Worked for the Version 1 and Version 2 questions. We calculated the rates overall and for the mail mode.

Table 19. Item Missing Data Rates for the When Last Worked Question - Version 1 vs Version 2

Mode	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	2.4 (0.2)	2.3 (0.2)	0.1 (0.3)	0.93
Mail	9.9 (0.8)	10.0 (0.8)	-0.1 (1.2)	0.93

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method.

There was no statistically significant difference in the item missing data rates between Version 1 and Version 2 for the When Last Worked question. This is not unexpected because the wording for this question in both versions was identical.

RQ10. For the Work for Pay in 2021 question, is there a difference in item missing data rates between Version 1 and Version 2 for mail responses?

Table 20 provides the item missing data rates of the Worked for Pay question for Version 1 and Version 2. We calculated rates overall and for the mail mode.[17]

[17] The Worked for Pay question was not on the Control questionnaire.

Table 20. Item Missing Data Rates for the Worked for Pay Question - Version 1 vs Version 2

Mode	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	3.9 (0.2)	4.1 (0.2)	-0.2 (0.3)	0.42
Mail	3.9 (0.5)	4.7 (0.5)	-0.8 (0.7)	0.42

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method.

There was no statistically significant difference in the item missing data rates between Version 1 and Version 2 of the Worked for Pay question. This is not unexpected because the wording for this question in both versions was identical.

RQ11. For the Number of Weeks Worked questions, is there a difference in item missing data rates between Control and Version 1? Version 1 and Version 2 for mail responses?

For RQ11 we looked at the Worked Every Week and the Number of Weeks Worked questions.

Table 21 provides the item missing data rates of Control and Version 1 of the Worked Every Week question. We calculated the rates overall and by mode.

Table 21. Item Missing Data Rates for the Worked Every Week Question - Control vs Version 1

Mode	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	1.4 (0.2)	4.3 (0.2)	-2.9 (0.3)	<0.01*
Internet	1.0 (0.2)	4.9 (0.2)	-3.9 (0.3)	<0.01*
Mail	3.7 (0.6)	6.8 (0.7)	-3.1 (0.9)	<0.01*
CAPI	1.6 (0.3)	1.3 (0.3)	0.3 (0.4)	0.38

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method.

The Control treatment had higher item missing data rates than Version 1 for the Worked Every Week question in the internet mode, the mail mode, and overall.

Table 22 provides the item missing data rates of Version 1 and Version 2 of the Worked Every Week question. We calculated the rates overall and for the mail mode.

Table 22. Item Missing Data Rates for the Worked Every Week Question - Version 1 vs Version 2

Mode	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	1.3 (0.2)	1.4 (0.2)	-0.1 (0.2)	0.72
Mail	1.8 (0.4)	3.7 (0.6)	-1.8 (0.7)	0.01*

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method.

Version 1 had higher item missing data rates than Version 2 for the Worked Every Week question in the mail mode.

Table 23 provides the item missing data rates of Control and Version 1 of the Number of Weeks Worked question. We calculated the rates overall and by mode.

Table 23. Item Missing Data Rates for the Weeks Worked Question - Control vs Version 1

Mode	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	14.6 (0.9)	23.3 (1.0)	-8.7 (1.4)	<0.01*
Internet	12.1 (1.1)	25.3 (1.1)	-13.1 (1.6)	<0.01*
Mail	19.3 (2.2)	29.4 (2.3)	-10.1 (3.1)	<0.01*
CAPI	19.7 (1.7)	12.8 (1.5)	7.0 (2.4)	<0.01*

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method.

The Control treatment had higher item missing data rates than Version 1 for the Number of Weeks Worked question in the internet and mail modes and overall. Version 1 had higher item missing data rates in the CAPI mode.

Table 24 provides the item missing data rates of the Number of Weeks Worked question for Version 1 and Version 2. We calculated the rates overall and for the mail mode.

Table 24. Item Missing Data Rates for the Weeks Worked Question - Version 1 vs Version 2

Mode	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	14.2 (0.8)	14.6 (0.9)	-0.4 (1.2)	0.74
Mail	14.7 (1.8)	19.3 (2.2)	-4.6 (2.8)	0.20

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method.

There was no significant difference in the item missing data rates between the two versions of the Number of Weeks Worked question.

RQ12. For the Number of Hours Worked question, is there a difference in item missing data rates between Control and Version 1? Version 1 and Version 2 for mail responses?

Table 25 provides the item missing data rates of Control and Version 1 of the Number of Hours Worked question. We calculated the rates overall and by mode.

Table 25. Item Missing Data Rates for the Hours Worked Question - Control vs Version 1

Mode	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	5.2 (0.3)	7.7 (0.3)	-2.5 (0.4)	<0.01*
Internet	4.6 (0.3)	8.2 (0.4)	-3.6 (0.5)	<0.01*
Mail	7.8 (0.9)	10.2 (0.9)	-2.4 (1.2)	0.10
CAPI	5.6 (0.7)	5.0 (0.6)	0.5 (0.9)	0.53

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method.

The Control treatment had higher item missing data rates than Version 1 for the Number of Hours Worked question in the internet mode and overall.

Table 26 provides the item missing data rates of the Number of Hours Worked question for Version 1 and Version 2. We calculated the rates overall and for the mail mode.

Table 26. Item Missing Data Rates for the Hours Worked Question - Version 1 vs Version 2

Mode	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	4.9 (0.2)	5.2 (0.3)	-0.3 (0.4)	0.46
Mail	5.4 (0.6)	7.8 (0.9)	-2.5 (1.0)	0.03*

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method.

Version 1 had higher item missing data rates than Version 2 for the Number of Hours Worked question in the mail mode.

5.3 Response Distribution Results for Labor Force

RQ13. For the When Last Worked question, is there a difference in response distributions between the Control and Version 1? Version 1 and Version 2 mail responses?

Table 27 shows the distribution of When Last Worked for the Control and Version 1 questions, both overall and by mode.

Table 27. Distribution of When Last Worked – Control vs Version 1

When Last Worked	Version 1 Percent	Control Percent	Chi-square	P-value
Overall	--	--	--	--
In the past 12 months	20.2 (0.5)	19.8 (0.6)	0.5	0.79
1 – 5 years ago	19.3 (0.5)	19.1 (0.5)		
Over 5 years or never	60.5 (0.6)	61.2 (0.8)		
Internet	--	--	--	--
In the past 12 months	20.0 (0.7)	20.5 (0.6)	0.4	0.84
1 – 5 years ago	23.0 (0.7)	22.9 (0.6)		
Over 5 years or never	57.0 (0.8)	56.6 (0.9)		
Mail	--	--	--	--
In the past 12 months	28.8 (1.1)	25.5 (1.2)	4.2	0.12
1 – 5 years ago	16.6 (1.0)	16.7 (0.9)		
Over 5 years or never	54.5 (1.3)	57.8 (1.2)		
CAPI	--	--	--	--
In the past 12 months	12.5 (1.0)	13.0 (1.3)	0.2	0.92
1 – 5 years ago	12.7 (0.8)	12.2 (1.1)		
Over 5 years or never	74.9 (1.3)	74.8 (1.6)		

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a chi-square test at the α=0.1 level.

There was no statistically significant difference in the distributions of When Last Worked between the Control and Version 1 questions.

Table 28 shows the distribution of When Last Worked for the Version 1 and Version 2 questions, both overall and for the mail mode.

Table 28. Distribution of When Last Worked - Version 1 vs Version 2

When Last Worked	Version 2 Percent	Version 1 Percent	Chi-square	P-value
Overall	--	--	--	--
In the past 12 months	20.4 (0.6)	20.2 (0.5)	0.2	0.92
1 – 5 years ago	19.1 (0.5)	19.3 (0.5)		
Over 5 years or never	60.5 (0.6)	60.5 (0.6)		
Mail	--	--	--	--
In the past 12 months	30.8 (1.3)	28.8 (1.1)	2.5	0.29
1 – 5 years ago	15.3 (0.8)	16.6 (1.0)		
Over 5 years or never	53.9 (1.2)	54.5 (1.3)		

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a chi-square test at the α=0.1 level.

There was no statistically significant difference in the distributions of When Last Worked between the two versions.

RQ14. Is there a difference in the proportion of people working in 2021 between the Version 1 and Version 2 questions?

Table 29 shows the proportion of people working in 2021 for the Version 1 and Version 2 questions, both overall and for the mail mode.

Table 29. Proportion of People Working in 2021 – Version 1 vs Version 2

Mode	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	83.7 (0.4)	84.4 (0.3)	-0.8 (0.4)	0.17
Mail	64.4 (1.2)	65.3 (1.2)	-0.9 (1.7)	0.58

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method.

There was no statistically significant difference in the proportion of people working in 2021 between the Version 1 and Version 2 questions.

RQ15. Does the proportion of full-time, year-round workers differ between Control and Version 1? Version 1 and Version 2 mail responses?

Table 30 shows the proportion of full-time, year-round workers for the Control and Version 1 questions, both overall and by mode.

Table 30. Proportion of Full-time, Year-round Workers - Control vs Version 1

Mode	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	69.5 (0.5)	67.7 (0.5)	1.8 (0.8)	0.06*
Internet	70.2 (0.6)	68.9 (0.5)	1.3 (0.9)	0.24
Mail	56.6 (1.5)	58.5 (1.4)	-1.9 (2.0)	0.32
CAPI	74.0 (1.1)	68.5 (1.3)	5.5 (1.6)	<0.01*

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method.

Version 1 had a higher proportion of full-time, year-round workers than Control in the CAPI mode and overall.

Table 31 shows the proportion of full-time, year-round workers for the Version 1 and Version 2 questions, both overall and for the mail mode.

Table 31. Proportion of Full-time, Year-round Workers - Version 1 vs Version 2

Mode	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	69.2 (0.5)	69.5 (0.5)	-0.3 (0.7)	0.64
Mail	59.6 (1.5)	56.6 (1.5)	3.0 (2.0)	0.27

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method.

There was no statistically significant difference in the proportion of full-time, year-round workers between the two versions.

RQ16. Among part-time or part-year workers, does the distribution of weeks worked responses by mode differ between Control and Version 1? Version 1 vs. Version 2 mail responses?

The chi-square test distribution of weeks worked for Control vs Version 1 for full-time, year-round workers indicated a significant difference overall and for all modes, so we calculated t-tests.

Table 32 shows the distribution of weeks worked among part-time or part-year workers for the Control and Version 1 questions, both overall and by mode.

Table 32. Distribution of Weeks Worked for Part-time or Part-year Workers – Control vs Version 1

Weeks Worked	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	--	--	--	--
13 weeks or less	15.5 (0.6)	18.1 (0.8)	-2.6 (0.9)	0.02*
14 to 26 weeks	16.5 (0.8)	14.9 (0.7)	1.6 (0.9)	0.42
27 to 39 weeks	11.6 (0.5)	10.6 (0.5)	1.0 (0.7)	0.47
40 to 47 weeks	14.8 (0.7)	15.5 (0.8)	-0.8 (1.1)	0.47
48 to 49 weeks	4.7 (0.4)	5.3 (0.4)	-0.5 (0.6)	0.47
50 or more weeks	36.9 (1.0)	35.6 (0.8)	1.3 (1.3)	0.47
Internet	--	--	--	--
13 weeks or less	16.4 (0.8)	18.8 (0.9)	-2.4 (1.0)	0.10
14 to 26 weeks	16.1 (0.8)	14.7 (0.8)	1.4 (1.1)	0.74
27 to 39 weeks	11.4 (0.7)	9.3 (0.6)	2.1 (0.9)	0.10
40 to 47 weeks	15.4 (0.9)	16.0 (0.8)	-0.6 (1.2)	0.85
48 to 49 weeks	5.1 (0.5)	6.0 (0.6)	-0.9 (0.8)	0.74
50 or more weeks	35.6 (1.3)	35.3 (1.0)	0.3 (1.7)	0.85
Mail	--	--	--	--
13 weeks or less	16.1 (1.5)	16.3 (1.9)	-0.1 (2.5)	0.96
14 to 26 weeks	16.8 (2.0)	11.4 (1.4)	5.3 (2.6)	0.19
27 to 39 weeks	13.2 (1.7)	11.4 (1.3)	1.8 (2.0)	0.96
40 to 47 weeks	13.9 (1.7)	19.9 (1.9)	-6.0 (2.6)	0.12
48 to 49 weeks	6.6 (1.2)	5.4 (0.9)	1.2 (1.5)	0.96
50 or more weeks	33.4 (2.6)	35.6 (2.5)	-2.2 (3.3)	0.96
CAPI	--	--	--	--
13 weeks or less	12.3 (1.3)	17.3 (1.8)	-5.0 (2.1)	0.09*
14 to 26 weeks	17.6 (2.1)	17.3 (1.4)	0.3 (2.4)	0.90
27 to 39 weeks	10.9 (1.3)	13.8 (1.7)	-2.9 (2.2)	0.71
40 to 47 weeks	13.4 (1.7)	12.0 (1.6)	1.4 (2.4)	0.90
48 to 49 weeks	2.2 (0.8)	3.2 (0.9)	-1.0 (1.2)	0.90
50 or more weeks	43.7 (2.4)	36.4 (2.2)	7.3 (3.2)	0.11

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method.

The Control question had a higher proportion of weeks worked responses in the 13 weeks of less category than Version 1 in the CAPI mode and overall.

Table 33 collapses the distribution of weeks worked into two categories (ends in 0 or 5 and any other positive number) for the Control and Version 1 questions. This comparison tests whether the new reference period could impact how respondents reply to the weeks worked question and whether or not answers are heaped.

Table 33. Distribution of Weeks Worked for Part-time or Part-year Workers – Control vs Version 1

Weeks Worked	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	--	--	--	--
End with 0 or 5	26.5 (0.9)	25.4 (0.9)	1.1 (1.2)	0.36
Other positive number	73.5 (0.9)	74.6 (0.9)	-1.1 (1.2)	0.36
Internet	--	--	--	--
End with 0 or 5	29.1 (1.2)	26.0 (0.9)	3.1 (1.4)	0.03*
Other positive number	70.9 (1.2)	74.1 (0.9)	-3.1 (1.4)	0.03*
Mail	--	--	--	--
End with 0 or 5	27.8 (2.3)	27.6 (2.2)	0.2 (3.3)	0.96
Other positive number	72.2 (2.3)	72.4 (2.3)	-0.2 (3.3)	0.96
CAPI	--	--	--	--
End with 0 or 5	16.9 (1.8)	22.4 (2.4)	-5.5 (3.0)	0.07*
Other positive number	83.1 (1.8)	77.6 (2.4)	5.5 (3.0)	0.07*

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method.

In the internet mode, the Version 1 question had a higher proportion of weeks worked responses that ended with a 0 or 5 than Control. In the CAPI mode, the Control treatment had a higher proportion of weeks worked responses that ended with a 0 or 5 than Version 1. These results indicate that the longer recall period may be associated with more heaping in self-response modes. The CAPI mode provides suggestive evidence that when a field representive is present there is less heaping with the new reference period.

Table 34 shows the distribution of weeks worked among part-time or part-year workers for the Version 1 and Version 2 questions, both overall and for the mail mode.

Table 34. Distribution of Weeks Worked for Part-time or Part-year Workers - Version 1 vs Version 2

Weeks Worked	Version 2 Percent	Version 1 Percent	Chi-square	P-value
Overall	--	--	--	--
13 weeks or less	15.7 (0.7)	15.5 (0.6)	1.1	0.96
14 to 26 weeks	16.8 (0.8)	16.5 (0.8)		
27 to 39 weeks	12.1 (0.7)	11.6 (0.5)		
40 to 47 weeks	14.8 (0.8)	14.8 (0.7)		
48 to 49 weeks	5.0 (0.4)	4.7 (0.4)		
50 or more weeks	35.6 (0.9)	36.9 (1.0)		
Mail	--	--	--	--
13 weeks or less	16.9 (2.0)	16.1 (1.5)	3.5	0.62
14 to 26 weeks	12.9 (1.6)	16.8 (2.0)		
27 to 39 weeks	11.8 (1.6)	13.2 (1.7)		
40 to 47 weeks	17.0 (1.7)	13.9 (1.7)		
48 to 49 weeks	6.6 (1.6)	6.6 (1.2)		
50 or more weeks	34.8 (2.3)	33.4 (2.6)		

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065

Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a chi-square test at the α=0.1 level.

There was no statistically significant difference in the distributions of weeks worked between the two versions.

Table 35 collapses the distribution of weeks worked into two categories (ends in 0 or 5 and any other positive number) for the Version 1 and Version 2 questions.

Table 35. Distribution of Weeks Worked for Part-time or Part-year Workers – Version 1 vs Version 2

Weeks Worked	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	--	--	--	--
End with 0 or 5	27.7 (1.0)	26.5 (0.9)	1.2 (1.4)	0.39
Other positive number	72.3 (1.0)	73.5 (0.9)	-1.2 (1.4)	0.39
Mail	--	--	--	--
End with 0 or 5	31.3 (2.5)	27.8 (2.3)	3.5 (3.6)	0.33
Other positive number	68.7 (2.5)	72.2 (2.3)	-3.5 (3.6)	0.33

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065

Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method.

There was no statistically significant difference in the distributions of weeks worked between the Version 1 and Version 2 questions when collapsing into two categories.

RQ17. Does the distribution of hours worked responses by mode and full-time, year-round worker status differ between Control and Version 1? Version 1 vs. Version 2 mail responses?

Tables 36 through 39 show the distributions of hours worked for full-time, year-round workers. Tables 40 through 43 show the distributions of hours worked for those who were not full-time, year-round workers (defined as part-time or part-year workers).

The chi-square test distribution of hours worked for Control vs Version 1 for full-time, year-round workers indicated a significant difference overall and for all modes, so we calculated t-tests.

Table 36 shows the distribution of hours worked for full-time, year-round workers for the Control and Version 1 questions, both overall and by mode.

A full-time, year-round worker is defined as a person (age 15 or older) who worked at least 35 hours per week and 50 weeks per year in 2021.

Table 36. Distribution of Hours Worked for Full-time, Year-round Workers – Control vs Version 1

Full-time, year-round Hours Worked	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	--	--	--	--
35 to 39 hours	7.0 (0.3)	6.3 (0.4)	0.7 (0.5)	0.36
40 hours	63.9 (0.7)	60.9 (0.8)	3.0 (1.0)	0.01*
41 to 49 hours	9.5 (0.4)	11.6 (0.5)	-2.0 (0.6)	0.01*
50 to 59 hours	12.6 (0.4)	13.1 (0.5)	-0.5 (0.6)	0.47
60 or more hours	6.9 (0.3)	8.1 (0.4)	-1.2 (0.5)	0.08*
Internet	--	--	--	--
35 to 39 hours	6.9 (0.4)	6.6 (0.4)	0.4 (0.6)	0.53
40 hours	62.9 (0.7)	59.7 (1.0)	3.3 (1.2)	0.03*
41 to 49 hours	10.3 (0.4)	11.8 (0.6)	-1.5 (0.7)	0.10
50 to 59 hours	13.1 (0.5)	13.6 (0.6)	-0.5 (0.7)	0.53
60 or more hours	6.8 (0.4)	8.4 (0.5)	-1.7 (0.6)	0.02*
Mail	--	--	--	--
35 to 39 hours	9.0 (1.1)	8.7 (1.0)	0.3 (1.6)	0.86
40 hours	60.0 (1.9)	56.4 (2.1)	3.6 (2.6)	0.70
41 to 49 hours	10.2 (1.1)	12.6 (1.4)	-2.4 (1.8)	0.70
50 to 59 hours	12.8 (1.5)	15.1 (1.8)	-2.3 (2.4)	0.86
60 or more hours	7.9 (1.2)	7.1 (1.0)	0.8 (1.6)	0.86
CAPI	--	--	--	--
35 to 39 hours	6.5 (0.7)	4.8 (0.6)	1.7 (1.0)	0.41
40 hours	68.1 (1.7)	66.4 (1.7)	1.7 (2.5)	0.73
41 to 49 hours	7.1 (0.8)	10.4 (1.1)	-3.4 (1.3)	0.06*
50 to 59 hours	11.2 (0.9)	10.8 (1.1)	0.5 (1.4)	0.73
60 or more hours	7.1 (0.7)	7.6 (0.9)	-0.5 (1.2)	0.73

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method

In the internet mode, Version 1 had a higher proportion of responses of 40 hours and a lower proportion of responses of 60 or more hours than Control. In the CAPI mode, the Control treatment had a higher proportion of responses of 41 to 49 hours. Overall, the Version 1 question had a higher proportion of responses of 40 hours and a lower proportion of responses for both 41 to 49 hours and 60 or more hours than Control.

Table 37 collapses the distribution of hours worked into two categories (ends in 0 or 5 and any other positive number) for full-time, year-round workers in the Control and Version 1 questions.

Table 37. Distribution of Hours Worked for Full-time, Year-round Workers – Control vs Version 1

Full-time, year-round Hours Worked	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	--	--	--	--
End with 0 or 5	92.3 (0.4)	91.5 (0.4)	0.9 (0.5)	0.08*
Other positive number	7.7 (0.4)	8.5 (0.4)	-0.9 (0.5)	0.08*
Internet	--	--	--	--
End with 0 or 5	92.6 (0.5)	91.7 (0.5)	0.8 (0.7)	0.21
Other positive number	7.4 (0.4)	8.3 (0.5)	-0.8 (0.7)	0.21
Mail	--	--	--	--
End with 0 or 5	88.5 (1.4)	85.5 (1.9)	2.9 (2.2)	0.19
Other positive number	11.5 (1.4)	14.5 (1.9)	-2.9 (2.2)	0.19
CAPI	--	--	--	--
End with 0 or 5	93.1 (0.8)	93.0 (0.8)	0.1 (1.1)	0.91
Other positive number	6.9 (0.8)	7.0 (0.8)	-0.1 (1.1)	0.91

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method

Overall, the Version 1 question had a higher proportion of responses that ended with a 0 or 5. These results indicate possible heaping among survey responses, with responses ending in 0 or 5 likely being estimations about the number of hours worked rather than a specific figure. Consequestly, the Version 1 questions may have less accurate response data than those in the Control version due to the new reference period asking about the previous calendar year rather than the past 12 months.

Table 38 shows the distribution of hours worked for the Version 1 and Version 2 questions, both overall and for the mail mode.

Table 38. Distribution of Hours Worked for Full-time, Year-round Workers - Version 1 vs Version 2

Full-time, year-round Hours Worked	Version 2 Percent	Version 1 Percent	Chi-square	P-value
Overall	--	--	--	--
35 to 39 hours	6.6 (0.4)	7.0 (0.3)	7.8	0.10
40 hours	62.5 (0.7)	63.9 (0.7)		
41 to 49 hours	9.5 (0.4)	9.5 (0.4)		
50 to 59 hours	13.4 (0.4)	12.6 (0.4)		
60 or more hours	8.1 (0.4)	6.9 (0.3)		
Mail	--	--	--	--
35 to 39 hours	6.9 (0.9)	9.0 (1.1)	2.2	0.70
40 hours	61.1 (2.2)	60.0 (1.9)		
41 to 49 hours	9.5 (1.3)	10.2 (1.1)		
50 to 59 hours	13.4 (1.5)	12.8 (1.5)		
60 or more hours	9.0 (1.5)	7.9 (1.2)		

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a chi-square test at the α=0.1 level.

There was no statistically significant difference in the distributions of hours worked between the Version 1 and Version 2 questions for full-time, year-round workers.

Table 39 collapses the distribution of weeks worked into two categories (ends in 0 or 5 and any other positive number) for the Version 1 and Version 2 questions.

Table 39. Distribution of Hours Worked for Full-time, Year-round Workers – Version 1 vs Version 2

Full-time, year-round Hours Worked	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	--	--	--	--
End with 0 or 5	92.0 (0.4)	92.3 (0.4)	-0.3 (0.6)	0.55
Other positive number	8.0 (0.4)	7.7 (0.4)	0.4 (0.6)	0.55
Mail	--	--	--	--
End with 0 or 5	90.7 (1.1)	88.5 (1.4)	2.3 (1.8)	0.20
Other positive number	9.3 (1.1)	11.5 (1.4)	-2.3 (1.8)	0.20

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method.

There was no statistically significant difference in the distributions of hours worked between the Version 1 and Version 2 questions for full-time, year-round workers when collapsing into two categories.

Table 40 shows the distribution of hours worked for the Control and Version 1 questions.

Table 40. Distribution of Hours Worked for Part-time or Part-year Workers - Control vs Version 1

NOT Full-time, year-round Hours Worked	Version 1 Percent	Control Percent	Chi-square	P-value
Overall	--	--	--	--
0 to 14 hours	17.9 (0.7)	16.6 (0.7)	3.9	0.69
15 to 34 hours	52.6 (0.9)	52.2 (0.9)		
35 to 39 hours	4.0 (0.4)	3.7 (0.4)		
40 hours	18.6 (0.6)	20.0 (0.8)		
41 to 49 hours	2.6 (0.4)	2.8 (0.4)		
50 to 59 hours	2.3 (0.3)	2.6 (0.3)		
60 or more hours	1.9 (0.2)	2.1 (0.2)		
Internet	--	--	--	--
0 to 14 hours	18.8 (0.9)	18.5 (0.8)	1.8	0.93
15 to 34 hours	52.9 (1.1)	51.4 (1.0)		
35 to 39 hours	3.5 (0.5)	3.6 (0.4)		
40 hours	18.3 (0.8)	19.5 (0.8)		
41 to 49 hours	2.6 (0.4)	2.6 (0.4)		
50 to 59 hours	2.6 (0.4)	2.5 (0.3)		
60 or more hours	1.5 (0.2)	1.8 (0.3)		
Mail	--	--	--	--
0 to 14 hours	21.1 (1.8)	16.1 (1.8)	8.7	0.19
15 to 34 hours	49.6 (2.6)	56.6 (2.6)		
35 to 39 hours	4.6 (0.9)	5.0 (1.1)		
40 hours	16.1 (1.7)	15.9 (1.7)		
41 to 49 hours	4.2 (1.2)	2.0 (0.6)		
50 to 59 hours	1.8 (0.6)	2.6 (0.8)		
60 or more hours	2.7 (1.2)	1.8 (0.5)		
CAPI	--	--	--	--
0 to 14 hours	12.7 (1.6)	11.1 (1.7)	7.1	0.31
15 to 34 hours	53.8 (2.6)	52.0 (2.2)		
35 to 39 hours	5.4 (1.0)	3.1 (0.8)		
40 hours	21.3 (2.0)	23.5 (2.2)		
41 to 49 hours	1.8 (0.7)	3.7 (1.1)		
50 to 59 hours	1.9 (0.7)	3.1 (0.8)		
60 or more hours	3.0 (0.9)	3.4 (0.8)		

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065 Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a chi-square test at the α=0.1 level.

There was no statistically significant difference in the distributions of hours worked between the Control and Version 1 questions for part-time or part-year workers.

Table 41 collapses the distribution of hours worked into two categories (ends in 0 or 5 and any other positive number) for the Control and Version 1 questions.

Table 41. Distribution of Hours Worked for Part-time or Part-year Workers – Control vs Version 1

NOT Full-time, year-round Hours Worked	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	--	--	--	--
End with 0 or 5	71.1 (1.0)	70.0 (0.8)	1.1 (1.3)	0.42
Other positive number	28.9 (1.0)	30.0 (0.8)	-1.1 (1.3)	0.42
Internet	--	--	--	--
End with 0 or 5	70.9 (1.1)	70.3 (0.9)	0.6 (1.5)	0.69
Other positive number	29.1 (1.1)	29.7 (0.9)	-0.6 (1.4)	0.69
Mail	--	--	--	--
End with 0 or 5	67.5 (2.1)	62.4 (2.1)	5.1 (3.0)	0.09*
Other positive number	32.5 (2.1)	37.6 (2.1)	-5.1 (3.0)	0.09*
CAPI	--	--	--	--
End with 0 or 5	74.5 (2.4)	73.5 (2.3)	1.0 (3.6)	0.78
Other positive number	25.5 (2.4)	26.5 (2.3)	-1.0 (3.6)	0.78

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method

In the mail mode, Version 1 had a higher percentage of responses that ended with a 0 or 5 and a lower percentage of responses that ended in any other positive number than Control.

The chi-square test distribution of hours worked for part-time or part-year workers for Version 1 vs Version 2 indicated a significant difference in the mail mode, so we calculated t-tests.

Table 42 shows the distribution of hours worked for the Version 1 and Version 2 questions, both overall and for the mail mode.

Table 42. Distribution of Hours Worked for Part-time or Part-year Workers – Version 1 vs Version 2

NOT Full-time, year-round Hours Worked	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	--	--	--	--
0 to 14 hours	18.1 (0.7)	17.9 (0.7)	0.1 (1.0)	0.90
15 to 34 hours	54.4 (1.1)	52.6 (0.9)	1.8 (1.4)	0.90
35 to 39 hours	3.7 (0.4)	4.0 (0.4)	-0.3 (0.6)	0.90
40 hours	17.7 (0.8)	18.6 (0.6)	-0.9 (1.0)	0.90
41 to 49 hours	1.9 (0.3)	2.6 (0.4)	-0.7 (0.4)	0.77
50 to 59 hours	2.5 (0.5)	2.3 (0.3)	0.2 (0.6)	0.90
60 or more hours	1.7 (0.2)	1.9 (0.2)	-0.2 (0.4)	0.90
Mail	--	--	--	--
0 to 14 hours	19.5 (2.1)	21.1 (1.8)	-1.5 (2.8)	0.80
15 to 34 hours	56.2 (2.6)	49.6 (2.6)	6.6 (4.0)	0.57
35 to 39 hours	2.8 (0.9)	4.6 (0.9)	-1.8 (1.3)	0.67
40 hours	15.4 (1.9)	16.1 (1.7)	-0.7 (2.7)	0.80
41 to 49 hours	1.2 (0.4)	4.2 (1.2)	-3.0 (1.4)	0.19
50 to 59 hours	3.5 (1.1)	1.8 (0.6)	1.6 (1.2)	0.67
60 or more hours	1.4 (0.4)	2.7 (1.2)	-1.3 (1.2)	0.80

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method

There was no statistically significant difference in the distributions of hours worked between the Version 1 and Version 2 questions for part-time or part-year workers.

Table 43 collapses the distribution of hours worked into two categories (ends in 0 or 5 and any other positive number) for the Version 1 and Version 2 questions.

Table 43. Distribution of Hours Worked for Part-time or Part-year Workers – Version 1 vs Version 2

NOT Full-time, year-round Hours Worked	Version 2 Percent	Version 1 Percent	Difference	Adjusted P-value
Overall	--	--	--	--
End with 0 or 5	71.3 (0.7)	71.1 (1.0)	0.1 (1.2)	0.91
Other positive number	28.8 (0.7)	28.9 (1.0)	-0.1 (1.2)	0.91
Mail	--	--	--	--
End with 0 or 5	65.3 (2.8)	67.5 (2.1)	-2.2 (3.5)	0.53
Other positive number	34.7 (2.8)	32.5 (2.1)	2.2 (3.5)	0.53

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method

There was no statistically significant difference in the distributions of hours worked between the two versions for part-time or part-year workers when collapsing into two categories

5.4 Response Reliability Results for Labor Force

For all the response reliability research questions we compared Control versus Version 1 only.

RQ18. Is the Gross Difference Rate (GDR) of the When Last Worked question different between the Control and Version 1?

Table 44 provides the gross difference rates of When Last Worked overall and mode.

Table 44. Gross Difference Rates of When Last Worked - Control vs Version 1

Mode	Version 1 GDR	Control GDR	Difference	Adjusted P-value
Overall	--	--	--	--
In the past 12 months	5.3 (0.4)	6.0 (0.6)	-0.7 (0.6)	0.27
1 – 5 years ago	10.9 (0.6)	12.1 (0.8)	-1.3 (0.9)	0.27
Over 5 years or never	9.8 (0.6)	12.6 (0.8)	-2.9 (1.0)	0.01*
Internet	--	--	--	--
In the past 12 months	5.3 (0.5)	5.8 (0.7)	-0.5 (0.8)	0.66
1 – 5 years ago	11.5 (0.7)	12.0 (1.0)	-0.5 (1.2)	0.66
Over 5 years or never	9.5 (0.7)	12.0 (1.1)	-2.6 (1.3)	0.13
Mail	--	--	--	--
In the past 12 months	4.7 (1.1)	4.9 (1.0)	-0.1 (1.4)	0.92
1 – 5 years ago	10.7 (1.5)	11.0 (1.3)	-0.3 (2.2)	0.92
Over 5 years or never	10.4 (1.4)	11.7 (1.3)	-1.3 (2.0)	0.92
CAPI	--	--	--	--
In the past 12 months	5.7 (1.2)	7.4 (1.4)	-1.7 (1.8)	0.35
1 – 5 years ago	9.6 (1.5)	13.6 (2.0)	-4.0 (2.7)	0.29
Over 5 years or never	9.9 (1.5)	14.9 (2.1)	-5.0 (2.6)	0.16

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method

The Control treatment had a higher GDR than Version 1 for the category ‘Over 5 years or never’ overall. There were no other statistically significant differences in the GDRs between Control and Version 1 for the When Last Worked question.

RQ19. Is the GDR of the Number of Weeks Worked question different between Control and Version 1?

First, we looked for exact matches between the original interview and the CFU reinterview. Examining differences in the match rates by survey mode highlights whether responses could vary because of how questions were administered. Table 45 provides the percentages of times when the response from the original interview and the CFU reinterview did not match (or the mismatch rate).

Table 45. Mismatch Rate for Weeks Worked - Control vs Version 1

Mode	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	20.3 (1.0)	22.6 (0.9)	-2.3 (1.4)	0.30
Internet	19.1 (1.1)	20.6 (1.0)	-1.6 (1.5)	0.57
Mail	30.4 (2.7)	32.0 (2.4)	-1.6 (3.6)	0.67
CAPI	18.5 (2.0)	25.2 (2.4)	-6.6 (3.3)	0.18

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method

There were no statistically significant differences in the mismatch rates between the Control and Version 1 questions.

In addition to the overall mismatch rate, we evaluated gross difference rates (or GDR) to determine if differences in response reliability were attributed to changes in only certain kinds of responses. To this end, we looked at when the response to the CFU interview and the response to the original interview were not in the same group, where the groups are defined as:

Group 1 = 13 weeks or less
Group 2 = 14 to 26 weeks
Group 3 = 27 to 39 weeks
Group 4 = 40 to 47 weeks
Group 5 = 48 to 49 weeks
Group 6 = 50 or more weeks

Table 46 shows the GDRs for these six groups.

Table 46. Gross Difference Rates of Weeks Worked - Control vs Version 1

Weeks Worked	Version 1 GDR	Control GDR	Difference	Adjusted P-value
Overall	--	--	--	--
13 weeks or less	2.1 (0.3)	4.0 (0.4)	-1.9 (0.5)	<0.01*
14 to 26 weeks	4.3 (0.4)	5.2 (0.5)	-0.9 (0.7)	0.88
27 to 39 weeks	4.4 (0.4)	4.3 (0.4)	<0.1 (0.5)	0.95
40 to 47 weeks	5.2 (0.4)	5.7 (0.5)	-0.5 (0.7)	0.95
48 to 49 weeks	1.8 (0.2)	2.1 (0.3)	-0.2 (0.4)	0.95
50 or more weeks	9.3 (0.6)	11.5 (0.7)	-2.3 (1.0)	0.12
Internet	--	--	--	--
13 weeks or less	1.9 (0.3)	3.9 (0.5)	-2.0 (0.6)	<0.01*
14 to 26 weeks	3.9 (0.5)	4.4 (0.5)	-0.5 (0.7)	0.87
27 to 39 weeks	3.9 (0.4)	3.8 (0.4)	0.1 (0.6)	0.87
40 to 47 weeks	5.0 (0.5)	5.3 (0.6)	-0.3 (0.7)	0.87
48 to 49 weeks	1.6 (0.3)	2.2 (0.3)	-0.6 (0.4)	0.77
50 or more weeks	8.5 (0.7)	9.7 (0.7)	-1.2 (1.0)	0.87
Mail	--	--	--	--
13 weeks or less	3.3 (0.7)	5.8 (1.3)	-2.5 (1.5)	0.56
14 to 26 weeks	4.7 (0.9)	5.2 (1.4)	-0.5 (1.9)	0.89
27 to 39 weeks	5.9 (1.4)	4.9 (1.1)	1.0 (1.6)	0.89
40 to 47 weeks	7.7 (1.4)	7.4 (1.6)	0.3 (2.1)	0.89
48 to 49 weeks	3.6 (1.0)	2.2 (0.7)	1.4 (1.3)	0.89
50 or more weeks	15.1 (2.0)	17.3 (2.0)	-2.2 (1.3)	0.89
CAPI	--	--	--	--
13 weeks or less	2.2 (0.8)	3.5 (0.8)	-1.2 (1.2)	0.95
14 to 26 weeks	5.5 (1.2)	7.6 (1.6)	-2.1 (2.1)	0.95
27 to 39 weeks	5.2 (1.2)	5.9 (1.2)	-0.7 (1.8)	0.95
40 to 47 weeks	4.3 (1.1)	6.3 (1.4)	-2.0 (1.8)	0.95
48 to 49 weeks	1.4 (0.6)	1.5 (0.9)	-0.1 (1.1)	0.95
50 or more weeks	8.3 (1.4)	14.9 (2.1)	-6.6 (2.6)	0.07*

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method

The Control treatment had a higher GDR (higher number of group mismatches) than Version 1 for Group 1 in the internet mode and overall. The Control treatment also had a higher GDR for Group 6 in the CAPI mode for the Number of Weeks Worked question.

RQ20. Is the GDR of the Number of Hours Worked question different between Control and Version 1?

For the Number of Hours Worked question, we compared both exact matches and group matches.

Table 47 shows the percentages of times when the response from the original interview and the CFU reinterview did not match.

Table 47. Mismatch Rate for Hours Worked - Control vs Version 1

Mode	Version 1 Percent	Control Percent	Difference	Adjusted P-value
Overall	40.1 (1.2)	43.5 (1.0)	-3.4 (1.5)	0.08*
Internet	38.6 (1.4)	41.5 (1.2)	-3.0 (1.6)	0.22
Mail	42.8 (3.2)	49.1 (3.3)	-6.3 (4.0)	0.22
CAPI	43.9 (2.9)	47.3 (2.7)	-3.4 (3.7)	0.35

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. An asterisk (*) indicates a statistically significant result. P-values have been adjusted for multiple comparisons using the Hochberg method

The Control treatment had a higher percentage of mismatches than Version 1 overall for the Number of Hours Worked question.

For hours worked, the groups were defined as:

Group 1 = 0 to 14 hours
Group 2 = 15 to 34 hours
Group 3 = 35 to 39 hours
Group 4 = 40 hours
Group 5 = 41 to 49 hours
Group 6 = 50 to 59 hours
Group 7 = 60 or more hours

Table 48 shows the GDRs for these seven groups.

Table 48. Gross Difference Rate of Hours Worked - Control vs Version 1

Hours Worked	Version 1 GDR	Control GDR	Difference	Adjusted P-value
Overall	--	--	--	--
0 to 14 hours	3.1 (0.4)	3.8 (0.4)	-0.6 (0.5)	0.96
15 to 34 hours	7.6 (0.7)	7.7 (0.5)	<0.1 (0.8)	0.96
35 to 39 hours	5.0 (0.3)	5.6 (0.5)	-0.6 (0.6)	0.96
40 hours	19.4 (0.8)	19.7 (0.8)	-0.3 (1.2)	0.96
41 to 49 hours	8.5 (0.6)	9.7 (0.6)	-1.2 (0.9)	0.96
50 to 59 hours	9.5 (0.5)	9.0 (0.7)	0.5 (0.8)	0.96
60 or more hours	5.2 (0.5)	5.2 (0.4)	-0.1 (0.6)	0.96
Internet	--	--	--	--
0 to 14 hours	3.3 (0.5)	3.7 (0.4)	-0.4 (0.6)	0.85
15 to 34 hours	7.2 (0.7)	7.0 (0.5)	0.2 (0.9)	0.85
35 to 39 hours	4.3 (0.3)	5.7 (0.6)	-1.4 (0.7)	0.24
40 hours	17.6 (1.0)	19.2 (1.1)	-1.6 (1.5)	0.85
41 to 49 hours	8.9 (0.8)	9.2 (0.7)	-0.4 (1.1)	0.85
50 to 59 hours	9.3 (0.5)	9.0 (0.7)	0.3 (0.8)	0.85
60 or more hours	4.7 (0.5)	5.2 (0.5)	-0.5 (0.8)	0.85
Mail	--	--	--	--
0 to 14 hours	4.8 (1.2)	4.3 (1.0)	0.6 (1.5)	0.72
15 to 34 hours	7.0 (1.3)	10.0 (3.4)	-2.9 (3.5)	0.72
35 to 39 hours	7.1 (1.4)	5.4 (1.0)	1.7 (1.6)	0.72
40 hours	19.0 (2.5)	20.1 (2.7)	-1.2 (3.2)	0.72
41 to 49 hours	8.0 (1.3)	9.8 (1.9)	-1.7 (2.4)	0.72
50 to 59 hours	7.2 (1.4)	11.9 (3.7)	-4.7 (3.7)	0.72
60 or more hours	5.9 (1.4)	5.0 (1.3)	0.9 (1.8)	0.72
CAPI	--	--	--	--
0 to 14 hours	1.5 (0.6)	3.7 (0.9)	-2.2 (1.1)	0.30
15 to 34 hours	9.7 (1.8)	8.9 (1.3)	0.8 (2.4)	0.73
35 to 39 hours	6.5 (1.3)	5.4 (1.1)	1.0 (1.6)	0.73
40 hours	26.5 (2.3)	21.3 (2.2)	5.1 (2.9)	0.49
41 to 49 hours	7.7 (1.6)	11.1 (1.6)	-3.4 (2.3)	0.56
50 to 59 hours	11.6 (1.6)	7.8 (1.8)	3.8 (2.4)	0.56
60 or more hours	6.5 (1.4)	5.6 (1.0)	0.8 (1.8)	0.73

Source: U.S. Census Bureau, 2022 American Community Survey Content Test | DRB No. CBDRB-FY23-ACSO003-B0065
Note: Minor additive discrepancies are due to rounding. Standard errors are in parentheses. Significance was tested using a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Hochberg method

There were no statistically significant differences in the GDRs between the Control and Version 1 questions.

5.5 Other Metrics

Measures of respondent burden were also analyzed for the 2022 Content Test. Metrics of interest were completion times, help screen access rates, and breakoff rates. These metrics were looked at overall and by each Content Test topic. There were no statistically significant differences found for the Labor Force questions between any of the treatments. For more information see Virgile et al. (2023).

Additional analyses were conducted to compare item missing data rates and response distributions for the COW and I&O questions. These questions were not part of the Content Test, but we wanted to see if changing the Labor Force series of questions would impact COW and I&O (which come after Labor Force in the survey).

That analysis showed that both the Test and Roster treatments had higher item missing data rates than Control for COW and Industry and that Test also had higher item missing data rates than Control for Occupation. These differences appear to be driven by the internet mode. See Martinez et al. (2023) for more information.

Internet breakoff rates were higher for Test than Control for the Income questions (which come after COW and I&O) but not for the Labor Force questions. Also the item missing data rates for Labor Force were generally higher for Control than Test. This suggests something happened after the Labor Force questions in the internet mode, but breakoff rates were not looked at for COW and I&O, so we are unsure what happened. See Virgile et al. (2023) for more information on breakoffs.

6 CONCLUSIONS AND RECOMMENDATIONS

The primary purpose for testing the Labor Force questions was to test a modification of the reference period from "past 12 months" to the previous calendar year. As a reference period, the prior calendar year will better align the ACS data with administrative records data, which the Census Bureau plan to use as a resource for improving the ACS in the future. Currently, the Census Bureau is conducting research to determine the feasibility of using administrative sources as a replacement or supplement to the income questions currently fielded on the survey. If administrative data use in ACS production is implemented, it could provide far-reaching benefits for multiple ACS topics including income, SNAP, and employment. To be ready to successfully implement the administrative sources in a timely fashion, we must begin work to adjust the questions on income and employment. This test was a significant step in that process.

Cognitive testing of the questions with the new reference period revealed some areas that could be improved upon on the paper survey so respondents could more clearly understand and respond to the questions. This test explored whether those potential areas of improvement yielded better data in a large-scale field test.

Benchmarks

Results from benchmark comparisons broadly suggest the 2022 ACS Content Test was similar to the 2022 CPS ASEC and 2022 SIPP in terms of employment characteristics and variable distributions. However, differences in survey universe, methodology, and individual question items prevent the statistical analyses of these data beyond a nominal level.

Nominal benchmark comparisons to CPS ASEC for proportion of full-time, year-round workers and civilian employment to population ratio indicated that the combined test group estimates approximated the CPS estimates, which provides one data point suggesting that external validity is satisfactory.

While nominal benchmark comparisons between the 2022 ACS Content Test and the 2022 SIPP indicate minor differences between surveys, these findings also support the notion of external validity. For instance, the percentage of full-time, year-round employees was higher among the combined test treatment groups in the ACS Content Test compared with the SIPP. Yet the civilian employment-to-population ratio was relatively similar between surveys.

Item Missing Rate

Concerning non-response and item missingness, findings indicate the two test treatments (Version 1 and Version 2) significantly outperformed the Control treatment among select survey modes. Compared with the Control, Version 1 had lower item missing rates for the Worked Every Week and Weeks Worked questions among overall, internet, and mail modes.

The Version 1 item missing rates were also lower than the Control for the Hours Worked question among overall and internet modes. These results broadly suggest that the Version 1 treatment's revised instructions improved question clarity and decreased item non-response. However, one notable exception concerns the item missing rate for the Weeks Worked question for the CAPI mode, which was significantly higher in Version 1 than in the Control. As a result, it is possible the change in reference period may have also affected item missing rates among the Test Treatment questions.

Differences between Version 1 and Version 2 test treatments for the employment section were limited to the mail mode and the Worked Every Week and Hours Worked questions. In the mail mode, the Version 1 test treatment had significantly higher item missing rates on both questions in comparison to the Version 2 test treatment. Among the remaining questions and modes, all differences between Version 1 and Version 2 were non-significant. Altogether, these results suggest the full instructions used in Version 2's mail mode decreased item missingness more than the bullet point instructions used in Version 1's mail mode.

Response Distribution

The response distributions for the Weeks Worked and Hours Worked questions differed between the Version 1 and Control Treatments by full-time, year-round work status and survey mode. Overall, the proportion of full-time, year-round workers in the Version 1 Test Treatment was significantly higher than the Control. Version 1 also had significantly more full-time year-round workers than the Control reporting 40 hours of work and significantly less reporting 41-49 hours, or 60 or more hours. Among only part-time or part-year workers, the share reporting work of 13 weeks or less was significantly higher in Version 1 than in the Control for both overall and for the CAPI mode. Forthcoming research using administrative records will determine whether these differences in response distributions significantly affect the ACS data quality.

Examining differences between treatments' response patterns also demonstrated how heaping was more prevalent in Version 1 compared with the Control. Specifically, the Version 1 test treatment had more heaping on numbers ending in 0 or 5 for the Hours Worked and Weeks Worked questions compared with the Control. Though we can only speculate, it is possible this increase in heaping may be the result of a longer recall period used in the test treatment.

Response Reliability

Comparisons between the Version 1 and Control treatments indicate the former had better response reliability in terms of gross difference rates (GDR). Specifically, findings show respondents administered the Version 1 test treatment answered select labor force questions more consistently between their initial interview and follow-up telephone intereview. Consequently, the GDR for the Version 1 treatment was lower than the Control among the following: (1) the share of people 16 and over who worked over 5 years ago or never worked, (2) those who reported working 13 weeks or less category overall and for the internet mode, (3) those who reported working 50 or more weeks category for the CAPI mode, and (4) the overall number of hours worked each week. There were no other significant differences between the tested questions.

Recommendations

We recommend keeping all elements of the labor force test questions except for the new reference period based on the decision criteria and the results of our analyses. These changes include improvements to the question wording and instructions in the internet and CAPI instruments, as well as in the mail mode. A change in the ACS reference period is not preferred since it is intended to coincide with the implementation of administrative records matching (specifically, LEHD and SSA data) in ACS production. However, because these administrative records are not currently available for a comparison to the ACS content test estimates , we would like to delay any changes to the ACS reference period until a future time.

In addition to the reference period change the labor force questions were modified to clarify how to correctly answer the series of questions. These changes will improve the accuracy of labor force items and yield higher quality data. Our recommendation to keep all non-reference period changes is supported by the following evidence:

1. All item missing rates for mail and internet were either lower or not significantly different for the Version 1 or Version 2 questions versus the Control question. Since the internet and mail modes constitute the largest share of cases in the ACS, the improvements to internet and mail forms can have a meaningful impact on ACS missing data and imputation rates.
2. The Version 1 question had lower GDR for several response categories and no categories had a higher GDR than Control, suggesting that Version 1 has better test-retest reliability than Control.

Future Research Directions

Despite the potential benefits of integrating administrative records into the ACS, some obstacles remain. For instance, while survey and administrative employment concepts are generally consistent, administrative data do suffer from known coverage gaps. Contract work, “gig” work, and some self-employment do not generate W-2 reporting. When no W-2 is issued to an individual, we can only feasibly observe that individual’s employment in administrative data when that individual’s tax unit files a tax return and when that individual reports the earnings as self-employment income. Any inconsistency in survey and administrative employment concepts limits the potential benefit of replacing survey employment variables with administrative data.

The literature has also documented disagreement between an individual’s survey report of class of worker status and the type of employment record observed in administrative data, which offers information about that individual’s self-employment status (Abraham et al., 2021).

Nevertheless, Eggleston et al. (2022) demonstrate that the correlation between survey and administrative measures of self-employment is strong enough that incorporating administrative records into imputation models can improve the quality of imputed data.

Administrative data sources on employment, income, and public assistance benefits from the Internal Revenue Service (IRS), Social Security Administration (SSA), and state administrative offices could meet the agency needs for many types of income, transfer benefits, and employment data. The Census Bureau is conducting research to determine the feasibility of using administrative sources as a replacement or supplement to the income questions currently fielded on the survey. If administrative data use in ACS production is implemented, it could provide far-reaching benefits for multiple ACS topics including income, SNAP, and employment. To be ready to successfully implement the administrative sources in a timely fashion, we must begin work to adjust the questions on income, employment, and SNAP. This test was a significant step in that process.

7 ACKNOWLEDGMENTS

The 2022 ACS Content Test would not have been possible without the participation and assistance of many individuals from the Census Bureau and other agencies. We sincerely appreciate their contributions.

- Census Bureau staff in the American Community Survey Office, Application Development and Services Division, Center for Behavioral Science Methods, 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 OMB's Interagency Working Group for the ACS and the topical subcommittees formed by the Working Group for each topic tested in the 2022 ACS Content Test.
- Staff in OMB's Statistical and Science Policy Office.

8 REFERENCES

Abraham, K., Haltiwanger, J., Sandusky, K., & Spletzer, J. (2021). Measuring the gig economy: Current knowledge and open issues. Measuring and accounting for innovation in the twenty-first Century (pp. 257-298). National Bureau of Economic Research, Inc.Biemer, P. (2011). *Latent class analysis of survey error.* John Wiley & Sons, Inc.

Biemer, P. (2011). *Latent class analysis of survey error.* John Wiley & Sons, Inc.

Census Bureau. (2015). Agility in action: A snapshot of enhancements to the American Community Survey. Retrieved September 13, 2023 from https://www.census.gov/programs-surveys/acs/methodology/agility-in-action/agility-in-action.html

Chenevert, R., Klee, M., & Wilkin, K. (2016). *Do imputed earnings earn their keep? Evaluating SIPP earnings and nonresponse with administrative records*. U.S. Census Bureau. Retrieved January 27, 2022, from https://www.census.gov/library/working-papers/2016/demo/SEHSD-WP2016-18.html

Eggleston, J., Klee, M. A., & Munk, R. (2022). *Self-employment status: Imputations, implications, and improvements*. U.S. Census Bureau. Retrieved September 30, 2022 from https://www.census.gov/library/working-papers/2022/demo/SEHSD-WP2022-06.html

Flanagan, P.E. (2001). Measurement errors in survey response. (Unpublished doctoral dissertation). University of Maryland, Baltimore County.

Heimel, S. (2016). *Postal tracking research on the May 2015 ACS panel*. U.S. Census Bureau. 2016 American Community Survey Research and Evaluation Report Memorandum Series #ACS16-RER-01.

Hochberg, Y. (1988). A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika*, 75 (4), 800-802. Retrieved January 17, 2017, from https://doi.org/10.2307/2336325

Jarmin, R. (2019). Evolving measurement for an evolving economy: Thoughts on 21st century US economic statistics. *Journal of Economic Perspectives,* 33(a), 165-184. https://doi=10.1257/jep.33.1.165

Keathley, D. (2022). *Specifications for selecting and weighting the 2022 American Community Survey Content Test sample.* U.S. Census Bureau. DSSD 2022 American Community Survey Memorandum Series #ACS22-S-21.

Klee, M., Chenevert, R., & Wilkin, K. (November 2019) Revisiting the shape of earnings nonresponse. *Economics Letters*, 184. https://doi.org/10.1016/j.econlet.2019.108663

Martinez, A., Longsine, L., & Risley, M. (2023). *2022 American Community Survey Content Test evaluation report: Exploratory Analysis of I&O and COW*. Forthcoming.

Office of Management and Budget (2006). *Standards and guidelines for statistical surveys*. Retrieved February 24, 2022, from https://www.whitehouse.gov/wpcontent/ uploads/2021/04/standards_stat_surveys.pdf

Ortman, J., Pharris-Ciurej, N., & Clark, S. (2018). *Realizing the promise of administrative data for enhancing the American Community Survey*. U.S. Census Bureau. Retrieved January 27, 2022, from https://www.census.gov/programs-surveys/acs/operations-and-administration/agility-in-action/administrative-records-in-the-american-community-survey.html

O'Hara, A., Bee, C., & Mitchell, J. (2016). *Preliminary research for replacing or supplementing the income question on the American Community Survey with administrative records*. U.S. Census Bureau. Retrieved January 27, 2022, from https://www.census.gov/content/dam/Census/library/working-papers/2016/acs/2016_Ohara_01.pdf

Rao, J., & Scott A. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, 15 (1), 1987, 385–397. https://doi.org /10.1214/aos/ 1176350273

Risley, M., & Oliver, B. (2022). High-level sampling and weighting requirements for the 2022 American Community Survey Content Test. U.S. Census Bureau. DSSD 2022 American Community Survey Memorandum Series #ACS22-MP-01.

RTI International (2022a). *2022 ACS Content Test: Round 1 and Round 2 Cognitive Testing.* Retrieved July 19, 2022, from https://www.census.gov/content/dam/Census/library/working-papers/2022/acs/2022_Wilson_01.pdf

RTI International (2022b). *Cognitive Testing for the 2022 ACS Content Test: Round 3 Briefing* Retrieved August 29, 2022 from https://www.census.gov/library/working-papers/2022/acs/2022_Wilson_02.htm.

Smith, A., Howard, D., & Risley, M. (2017). *2016 American Community Survey Content Test evaluation report: Number of weeks worked*. U.S. Census Bureau. Retrieved January 27, 2022, from https://www.census.gov/library/working-papers/2017/acs/2017_Smith_01.html

Spiers, S. (2021a). *Requirements for the Content Follow-Up Reinterview Survey in the 2022 American Community Survey Content Test*. U.S. Census Bureau. American Community Survey Memorandum Series #ACS22-MP-02.

Spiers, S. (2021b). Coverage and nonresponse bias in the 2016 American Community Survey Content Follow-Up reinterview. *Proceedings of the 2021 Joint Statistical Meetings, Survey Research Methods Section* (pp. 1664-1677). American Statistical Association.

Spiers, S., Baumgardner, S., & Mojica, A. (2023). *2022 American Community Survey Content Test evaluation report: Unit-level response*. Retrieved November 30, 2023 from https://www.census.gov/library/working-papers/2023/acs/2023_Spiers_01.html

Steiger, D., Robins, C., & Stapleton, M. (2017). 0Y1 American Community Survey respondent burden testing: Weeks worked and income final briefing report. Westat.

U.S. Census Bureau. (2022a). *U.S. Census Bureau statistical quality standards*. Retrieved September 16, 2022, from https://www2.census.gov/about/policies/quality/quality-standards.pdf

U.S. Census Bureau. (2022b). *American Community Survey and Puerto Rico Community Survey Design and Methodology, Version 3.0.* Retrieved December 15, 2022, from https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html

U.S. Census Bureau. (2022c). *ACS Research & Evaluation Analysis Plan: 2022 ACS Content Test*. U.S. Census Bureau. Retrieved April 26, 2023 from Field Test Analysis - All Documents (census.gov)

Virgile, M., Weng, C. F., Mills, G., & Spiers, S. (2023). *2022 American Community Survey Content Test evaluation report: Respondent burden*. Retrieved November 30, 2023 from https://www.census.gov/library/working-papers/2023/acs/2023_Virgile_01.html.