American Community Survey Research and Evaluation Program

ACS Research & Evaluation Analysis Plan (REAP)

2022 ACS Content Test

United States™
Census
Bureau

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

1. INTRODUCTION

The U.S. Census Bureau will conduct the 2022 American Community Survey (ACS) Content Test, a field test of new and revised content. The goal of this test is to improve the content of the ACS data collection instruments. The 2022 ACS Content Test will test the wording, format, and placement of new questions and revised questions. The results of this test will help determine which new or revised questions will be implemented in the ACS. Although the 2022 ACS Content Test will employ a separate sample from production ACS, it will occur in parallel with data collection activities for the September 2022 ACS production panel.

This research and evaluation analysis plan covers all topics included in the 2022 ACS Content Test. Section 2 provides the background, the reasons for the proposed content changes, and the phases of the 2022 ACS Content Test. Section 3 describes the methodology common to all topics. Sections 4 through 13 are topic specific and provide the background information, the question wording being tested, the research questions, the methodology specific to the topic, and the criteria which will be used to determine whether to implement the new or modified content in the ACS data collection instruments.

The development of this analysis plan was a collaboration among different directorates within the Census Bureau and external agencies.

2. BACKGROUND

In June 2018, the Census Bureau solicited proposals for new or revised ACS content from over 25 federal agencies. For revisions to existing questions, the proposals contained a justification for each change and described previous testing of question wording, the expected impact of revisions to ACS estimates, and the estimated net impact on respondent burden.

For new questions, the proposals provided an explanation of the need for the new data, why other data sources that provide similar information are not sufficient, how policy or data needs would be addressed through the new questions, and an explanation of why the data were needed with the geographic precision and frequency provided by the ACS. 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 existing content are tested according to the ACS content change process, which includes cognitive testing and field testing.

Census Bureau staff along with representatives from other federal agencies, which form the OMB Interagency Committee for the ACS, participated in development and testing activities.

2.1 COGNITIVE TESTING

In accordance with OMB's Standards and Guidelines for Statistical Surveys (OMB, 2006) and the Census Bureau's Stastistical Quality Standards (U.S. Census Bureau, 2013), the Census Bureau conducts cognitive interviewing to pretest survey questions prior to field testing them or implementing them in production. For the 2022 ACS Content Test, the Census Bureau contracted with Research Triangle Institute (RTI) International to conduct three rounds of cognitive testing.[1]

The results of the first two rounds of cognitive testing informed decisions on specific revisions to the proposed content for the stateside 2022 ACS Content Test field test (RTI International, 2021). 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 teams used the results of both rounds of cognitive testing to recommend question content for the field test.

A third round of cognitive testing was conducted in Puerto Rico and in Group Quarters (GQ), as the 2022 ACS Content Test will 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 (2022).

Three topics included in the cognitive testing will not be included in the field test: Homeowners Association or Condominium Fee, Home Heating Fuel, and Means of Transportation to Work. For the most part, the changes 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 deemed that cognitive testing was sufficient for these questions and that field testing was not necessary before implementing these questions in production ACS.

---

[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.

2.2 CONTENT OVERVIEW

Table 1 lists the current ACS questions for which content revisions will be field tested. Table 2 describes the new questions being tested for potential addition to the ACS.

Table 1. Current ACS Questions with Revisions Being Tested

Question/Topic [Section of Paper Questionnaire]	Requested by	Brief Summary of Revisions Being Tested and Reason(s) for Revision
Household Roster [Front Cover]	Census Bureau	Provided a new set of rostering instructions (paper questionnaire) and a new set of rostering instructions and questions for internet and interviewer-administered modes. The roster instructions have not changed since the 1990s while household living arrangements have increased in complexity. The revisions attempt to capture more complex living situations and improve within household coverage, especially among young children and tenuously attached residents.
Educational Attainment [Detailed Person Questions]	Census Bureau	Modified question format and wording. A relatively high percentage of respondents are selecting the response category "No schooling completed." Ongoing research suggests that this includes adults who have completed some level of schooling. The revision attempts to reduce these cases by collapsing the response categories for educational attainment lower than grade 1 into a single response category.
Health Insurance Coverage [Detailed Person Questions]	Census Bureau (See note at the bottom of Table 1.)	Modified the question by reordering the categories, providing additional instructions about what types of plans should be included, updating the wording of employer provided insurance to include insurance through a professional association, updating the wording of direct purchase insurance to include plans purchased through a health insurance marketplace or healthcare.gov, updating the wording of Medicaid to include the Children's Health Insurance Program (CHIP), and updating the wording of VA health care to specify Veteran's health care. The revisions attempt to capture changes to the health insurance landscape that occurred with the passage of the Patient Protection and Affordable Care Act. Created a second version to be tested based on cognitive testing that, in addition to the other modifications, changes the format to check-all-that-apply and adds an uninsured response option.

Attachment C

Question/Topic [Section of Paper Questionnaire]	Requested by	Brief Summary of Revisions Being Tested and Reason(s) for Revision
Disability [Detailed Person Questions]	National Center for Health Statistics	Modified question wording and response categories by eliminating the word "serious" from four of the questions, changing the response categories from an absolute scale (yes/no) to a degree of difficulty scale with four response options, and adding a question on communication difficulty. The revisions attempt to reflect advancements in the measurement of disability by basing the disability questions on the Washington Group Short Set on Functioning (WG-SS), which is recommended by the United Nations for measuring disability and is currently collected as part of the National Health Interview Survey and the National Health and Nutrition Examination Survey.
Supplemental Nutrition Assistance Program (SNAP) [Housing Questions]	Census Bureau	Modified the reference period from "past 12 months" to the previous calendar year. The revision attempts to align ACS data to better match administrative records, which the Census Bureau plans to use as a data source in the future.
Labor Force [Detailed Person Questions]	Census Bureau	Modified the reference period from "past 12 months" to the previous calendar year. The revision attempts to align ACS data to better match administrative records, which the Census Bureau plans to use as a data source in the future. An additional question was added to properly set up the universe to ask the Labor Force questions for the previous calendar period.
		Cognitive testing of the questions with the new reference period revealed some areas that could be improved upon so respondents could more clearly understand and respond to the questions. The test questions include the revisions suggested as a result of cognitive testing.
		This series of questions involves several skip patterns and many instructions. Thus, navigating the questions is more difficult on paper than on an automated instrument (internet or interviewer-administered). To provide clarity, and hopefully reduce respondent burden in the future, the Census Bureau is testing two versions of the questions on the paper questionnaire.

Attachment C

Question/Topic [Section of Paper Questionnaire]	Requested by	Brief Summary of Revisions Being Tested and Reason(s) for Revision
Income [Detailed Person Questions]	Census Bureau	Modified the reference period from "past 12 months" to the previous calendar year. The revision attempts to align ACS data to better match administrative records, which the Census Bureau plans to use as a data source in the future. One version of the questions to be tested uses production wording and only changes the reference period.
		Cognitive testing of the questions with the new reference period revealed some areas that could be improved upon so respondents could more clearly understand the questions and respond to them accurately. A second version of the test questions includes the revisions suggested as a result of cognitive testing (changes in question wording and instructional wording) in addition to the change in reference period. Both versions are being tested.

Note: The Census Bureau requested a change to the Health Insurance Coverage question in response to feedback from partners at the Health and Human Services Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Veterans Affairs, National Center for Health Statistics, and Agency for Healthcare Research and Quality.

Table 2. Proposed New ACS Questions and the Reason for Potential Inclusion in the ACS

Question/Topic [Section of Questionnaire]	Requested by	Reason for Adding the Question
Solar Panels [Housing Questions]	Energy Information Administration	This new question asks if the housing unit uses solar panels that generate electricity. By adding this question, the Energy Information Administration (EIA) will be able to obtain data for operational solar panels on a housing unit level across the country. This information will help the EIA match energy consumption to energy production across the United States.
Electric Vehicles [Housing Questions]	Energy Information Administration	This new question asks if there are plug-in electric vehicles kept at the housing unit. By adding this question, EIA will be able to project future energy sources, infrastructure, and consumer needs for the growing popularity of electric vehicles. The ACS would be the only data source at the housing unit level to adequately make these projections.

U.S. Census Bureau

Question/Topic [Section of Questionnaire]	Requested by	Reason for Adding the Question
Sewer [Housing Questions]	Environmental Protection Agency and the Department of Agriculture	This new question asks if the housing unit is connected to a public sewer, septic tank, or other type of sewage system. By adding this question, the Environmental Protection Agency and Department of Agriculture will be able to obtain consistent data on the decentralized wastewater infrastructure status in rural and other communities. This is needed to protect public health, water quality, and to understand and meet the country's growing infrastructure needs. The ACS is the only available survey that can provide this level of data in a timely, consistent, and standardized manner.

2.3   DATA COLLECTION

The 2022 ACS Content Test will occur in parallel with data collection activities for the September 2022 ACS production panel. The 2022 ACS Content Test will use a nationally representative sample, independent of the ACS production sample, distributed evenly into three treatments – Control Treatment, Test Treatment, and Roster Test Treatment. More information about the sample is provided in Section 3.1. This independent sample will use mostly the same data collection protocols as production ACS, except where noted.

The data collection protocols that will be the same are:

- Data will be collected using the self-response modes of internet (in English and Spanish) and paper questionnaires (in English only) for the first and second month of data collection.
- In the third month of data collection, a sub-sample of nonresponding addresses will be selected for follow-up in the Computer-Assisted Personal Interviewing (CAPI) operation. The CAPI instrument is only in English and Spanish.
- During CAPI, Census Bureau field representatives will conduct interviews in person and over the phone.
- Self-response via internet or paper will be accepted throughout the three-month data collection period.

The data collection protocols that will be different are:

- There will be no paper versions of the 2022 ACS Content Test questionnaires in Spanish.[3]
- If respondents call Telephone Questionnaire Assistance (TQA) and opt to complete the survey over the phone, the interviewers will conduct the survey using the production ACS questionnaire.[4] Since the TQA interviews will not include test questions, they will be excluded from the analysis of the 2022 ACS Content Test.
- The 2022 ACS Content Test will not have 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.[5]
- The 2022 ACS Content Test will have a telephone reinterview component used to measure response reliability (response variance or response bias). This telephone reinterview operation is named Content Follow-Up (CFU). We describe the CFU operation in more detail in Section 2.4.

The nonresponse bias that results from excluding Spanish paper questionnaire interviews and TQA interviews from analysis is expected to be negligible, as indicated in footnotes 3 and 4. Additional information about ACS data collection procedures can be found in the ACS Design and Methodology Report (U.S. Census Bureau, 2014b).

2.4   CONTENT FOLLOW-UP

To measure response reliability, we will attempt a CFU reinterview with every household with an original Content Test interview that meets the eligibility requirements. Some of those requirements include the household must be occupied, and the household must have a valid telephone number. See the CFU requirements document for the complete list of eligiblity requirements (Spiers, 2021). This document also details which questions will be asked in the CFU reinterview and provides instructions for occasions where the original respondent is not available.

---

[3] In 2019, 412 Spanish questionnaires were mailed back out of all mailable cases. Based upon this rate, we project that only 8 Spanish questionnaires would be mailed back in the 2022 Content Test, which would be neither cost-effective nor of sufficient size for drawing meaningful conclusions for the subdomain.

[4] The interviewer will not know if a caller is in a test treatment and will therefore administer 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 project about 744 TQA-only responses will be excluded from the 2022 ACS Content Test analysis.

[5] 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 will only collect name, age, and sex of these additional persons, but not detailed information as we do in ACS production.

The CFU reinterview will be conducted two to five weeks after the original interview. As in previous ACS Contest Tests, a case will be sent to the CFU operation no sooner than two weeks (14 calendar days) after the original interview and must be completed within three weeks after being sent to the CFU. This timing attempts to balance two competing needs: to minimize the possibility of real changes in answers due to a change in life circumstances between the two interviews and to minimize the possibility that the respondent repeats their previous answer based on their recollection of the original interview response rather than considering the most appropriate answer.

All CFU reinterviews will be conducted by telephone. At the first contact with a household, interviewers will ask to speak with the original respondent. If that person is not available, interviewers will schedule a callback at a time when the original respondent is expected to be home. If at the second contact we cannot reach the original respondent, the interview may be conducted with any other eligible household member (a household member who is 15 years or older). CFU reinterviews for the Content Test will be conducted in either English or Spanish.

Conducting CFU reinterviews by telephone presents coverage and participation problems. The concern is whether the households that complete a CFU reinterview are representative of all households that completed an original interview.

A recent study of CFU respondents in the 2016 ACS Content Test examined differences between CFU respondents and nonrespondents (Spiers, 2021). The study found that there were some differences both in those eligible and ineligible for the 2016 CFU and in those who responded and did not respond to it. However, the study was unable to determine the magnitude of the bias in the variables examined (e.g., original interview response mode, high versus low response strata, age, sex, education, and marital status). Despite these shortcomings, the CFU reinterview is beneficial in that it provides an indication of response reliability in the population examined.

The CFU instrument will not include all ACS questions, but for context, it will include some production questions in addition to the ones being tested for revision or addition. It will also include 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 public assistance question. Questions will be asked in the reinterview regardless of whether the question was left blank in the original interview.

The questions that will be asked in the CFU reinterview include:

- Household Roster
- Relationship
- Sex
- Age and Date of Birth

- Building Type
- Sewer
- Total Vehicles
- Electric Vehicles
- Solar Panels
- Educational Attainment
- Health Insurance series of questions
- Disability
- Work for Pay
- Layoff
- Looking for Work
- Start Job if Offered
- Labor Force series of questions
- Class of Worker
- Industry and Occupation
- Income series of questions (self-employment, rental, retirement, and public assistance)

CFU respondents will be asked the same version of the question they answered in the original interview with the following exceptions:

- The CFU questions for Income will come from the 2022 Current Population Survey Annual Social and Economic Supplement (CPS ASEC).
- There will be no income questions asked of CFU respondents in the Control treatment.
- CFU respondents in the Test treatment will not be asked household roster or basic person questions.
- CFU respondents in the Roster Test treatment will not be asked the full set of CFU questions.[6]

For more information on the CFU operation, see Spiers (2021).

3. METHODOLOGY

3.1 EXPERIMENTAL DESIGN

The 2022 ACS Content Test will consist of a national sample of roughly 120,000 housing unit addresses, excluding Puerto Rico, Alaska, and Hawaii.[7] The sample design for the Content Test will largely be based on the ACS production sample design with some modifications to meet the

---

[6] The Roster Test treatment will exclude Building Type, Sewer, Electric Vehicles, Solar Panels, Educational Attainment, and Disability in the CFU reinterview.

[7] Due to cost constraints, we are only including stateside housing units.

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

The modifications include adding an additional level of stratification by stratifying addresses into high and low self-response areas, over-sampling addresses from the low self-response areas to ensure equal response from both strata, and sampling units as evenly split trios (i.e., a control and two test treatments). The high and low self-response strata will be defined based on ACS self-response rates from the 2018 and 2019 panels at the tract-level.

Sampled trios will be selected using the nearest neighbor method. The selection process begins with a geographically sorted list. The first address is selected by systematically sampling an address within the defined sampling strata. The selection process then proceeds forwards and backwards in the sorted list to select the nearest neighbors. It first checks to see if the next record in the sort is in the same /RESPSTR /FIPST/FCNTY/SBSTR/ STRM/ combination as the initially selected address.[8] If it is, then this record becomes the first nearest neighbor. If it isn't in the same combination, but the immediately preceding record in the sort is in the same combination as the initially selected record, then that record is selected as the first nearest neighbor.

If the /RESPSTR/FIPST/FCNTY/SBSTR/STRM/ values of the initially selected address don't match those of the immediately preceding and following records, then we repeat the process for just the /RESPSTR/FIPST/FCNTY/SBSTR/ values. Then, if needed, we check just the /RESPSTR /FIPST/ /FCNTY/ values. Then finally, if necessary, the same /RESPSTR/FIPST/ values. The second nearest neighbor is chosen similarly. Each address in this trio is is randomly assigned to one of the three treatments. Addresses in sample for the 2022 ACS Content Test will be split evenly across three treatments, one control treatment and two test treatments named Test Treatment and Roster Test Treatment (See Table 3).

The Test treatment will contain one version of the test question for all topics except Household Roster (which will use the control wording). The Test treatment is where we will test one version of the test question against the current production question in the Control treatment. For the new questions, Solar Panels and Sewer, where there is only one version of the test question, the same question will be asked in the Control and Test treatments. For the other new question, Electric Vehicles, since there are two versions of the test question, the Control treatment and Test treatment will each have have different versions of this question.

The primary purpose of the Roster Test treatment is to test the roster test question separately since changes in a household's roster could impact the results of person-level topics (Educational Attainment, Disability, etc.). Therefore, the analyses for Test Version 2 of the

---

[8] RESPSTR=the sampling stratum type: high versus low response stratum; FIPST=fips state code; FCNTY=fips countycode; SBSTR=ACS 2nd stage sampling stratum; STRM=ACS 1st stage sampling stratum.

Health Insurance Coverage, Labor Force, and Income questions could be impacted by these changes. As part of our analysis, we will compare the household sizes and person-level characteristics between treatments (see Section 3.2.1.3). We will report any significant differences that might confound the results. We will also adjust the person-level data from the Roster Test treatment based on any found differences to limit the confounding effects and allow comparisons of Test Version 2 of these topics to other treatments.

Table 3. 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 will be in both Test treatments to align with Labor Force and Income that also have a reference period change to the previous calendar year.

3.2   ANALYSIS

The purpose of the 2022 ACS Content Test is to field test new and revised questions in the ACS data collection instruments. Sample addresses for the three treatments were selected in a manner so that their response propensities are similar (see Section 3.1 for methodology). As part of our analysis, we will test this assumption for the original interview and the CFU reinterview using various unit-level response metrics. The details are presented in Section 3.2.1.

With respect to analysis of the new and revised questions, we will compute a variety of metrics. Section 3.2.2 discusses metrics that are common to all topics (topic-level analysis). Details of the specific metrics that will be calculated for each topic are discussed in Sections 4 through 13.

Section 3.2.3 discusses the editing, weighting, and hypothesis testing methodologies pertaining to the unit-level and topic-level metrics.

3.2.1 Unit-Level Analysis

The samples for the Control and Test treatments were designed to have similar response propensities on the unit-level (i.e., address level). Therefore, we do not expect their unit response rates to be different. As part of our analysis for the 2022 ACS Content Test, we will test this assumption. We will calculate and compare the unit response rate overall and by mode (internet, mail, and CAPI) for each treatment. We will perform this analysis for the original interview (mail, internet, and CAPI responses) and for the CFU reinterview (to ensure that the rates are high enough to provide appropriate measures of response error and response reliability; see Section 3.2.2.4 for details on these measures).

The samples for the Control and Test treatments are selected using the nearest neighbor method described in Section 3.1. As such, it is expected that these treatments will have similar socioeconomic and demographic distributions. Similar distributions allow us to conclude that any differences in the analysis metrics are more than likely attributable to differences in the test questions. To test the assumption of similar distributions, we will calculate and compare respondent characteristics for the control and test.

Details about these analyses are provided in Sections 3.2.2.1 through 3.2.2.3. This research will only be published if noteworthy results are found.

3.2.1.1 Unit Response Rates

The unit response rate is generally defined as the weighted proportion of sample addresses eligible to respond that provided a complete or sufficient partial response.[9] We will answer the following questions pertaining to the unit response rate:

Research Question 1

What are the overall unit response rates for the original Content Test interview in the Control and two Test treatments? How do they compare?

Research Question 2

What are the unit response rates by data collection mode for the original Content Test interview in the Control and two Test treatments? How do they compare?

---

[9] In general, a sufficient partial internet response is one that has at least minimal information, which indicates an attempt to respond. The specific definition of a sufficient partial internet response is sensitive and for Census Bureau internal use only.

Attachment C

Research Question 3

What are the unit response rates by high and low response areas for the original Content Test interview in the Control and two Test treatments? How do they compare?

For the Control and Test treatments, we will calculate the overall unit response rate (all modes of data collection combined) and unit response rates by mode: internet, mail, and CAPI. We will also calculate the total self-response rate by combining internet and mail modes together.

Some topic-specific analyses will focus on the different data collection modes, so it is important to include each mode in the response rates. In addition to those rates, we calculate the response rates for high and low response areas because analysis for some Content Test topics will be done separately for each of the two response areas, as well as overall. High and low response strata will be defined based on historical ACS response rates from the 2018 and 2019 panels.

The universe for both the overall unit response rates and self-response rates consists of all addresses in the initial sample that are eligible to respond to the survey. The universe for the CAPI response rates consists of a subsample of all remaining nonrespondents from the initial sample that were eligible to respond to the survey. Any nonresponding addresses that were sampled out of CAPI will not be included in any of the response rate calculations.

To calculate response rates and standard errors, we will use replicate base weights adjusted for CAPI sub-sampling (but not adjusted for nonresponse). All response rates will be weighted.

Overall unit response rates are calculated as follows for each treatment:

$$\text{Final Response Rate} = \frac{\text{Number of sample addresses that provided a response by mail, internet, or CAPI}}{\text{Total number of sample addresses that were eligible to respond to the survey}} \times 100$$

Unit response rates by mode are calculated as follows for each treatment:

$$\text{Self-Response Response Rate} = \frac{\text{Number of sample addresses that provided a response by mail or internet}}{\text{Total number of sample addresses in the universe that were eligible to respond to the survey}} \times 100$$

U.S. Census Bureau

$$\text{CAPI Response Rate} = \frac{\text{Number of sample addresses that provided a response by CAPI (phone or in person)}}{\text{Total number of sample addresses in the universe that were eligible to respond to the survey}} \times 100$$

We will test for differences for all unit response rates using two-tailed t-tests. We will calculate for each mode separately and overall.

3.2.1.2   CFU Response Rates

We will calculate the CFU response rates overall and by mode. We expect these rates to be similar between treatments; however, we will test this assumption. Any differences observed will more than likely be due to randomness.

<u>Research Question 4</u>

What are the response rates for the CFU reinterviews in the Control and two Test treatments? How do they compare overall and by mode?

For the CFU response rates, we must determine which original interview units are eligible for the CFU and calculate CFU response rates based on that universe. In addition, since the CFU is a telephone-only survey, we will calculate CFU response rates by the data collection mode of the original interview, in addition to the overall CFU response rate. For each type of rate (overall, internet, mail, CAPI, and total self-response) we will do a statistical comparison of rates amongst the treatments.

Overall CFU response rates are calculated as follows for each treatment:

$$\text{Final CFU Response Rate} = \frac{\text{Number of sample addresses that provided a CFU response}}{\text{Total number of sample addresses that were eligible for the CFU operation}} \times 100$$

CFU response rates by mode are calculated as follows for each treatment:

$$\text{CFU Response Rate for Original Self-Response Interviews} = \frac{\text{Number of sample addresses that provided a CFU response}}{\text{Total number of sample addresses that provided an original self-response original interview and were eligible for the CFU operation}} \times 100$$

$$\text{CFU Response Rate for Original CAPI Interviews} = \frac{\text{Number of sample addresses that provided a CFU response}}{\text{Total number of sample addresses that provided an original CAPI interview and were eligible for the CFU operation}} \times 100$$

For the CFU response rates, we will use replicate base weights adjusted for original interview CAPI sub-sampling and for original interview nonresponse but not adjusted for CFU nonresponse.

We will test for differences in CFU response rates using two-tailed t-tests. We will calculate for each original mode separately and overall.

3.2.1.3   Demographic Characteristics of Respondents

The 2022 ACS Content Test sample was designed so that respondents in both Control and Test treatments exhibit similar distributions of socioeconomic and demographic characteristics. Similar distributions allow us to compare the treatments and conclude that any differences are due to the experimental treatment instead of underlying demographic differences. To test the assumption of similarity of respondent characteristics among treatments we will answer the following question:

<u>Research Question 5</u>

How are responding units in the original Content Test interview distributed by socioeconomic and demographic characteristics? How do these distributions compare between the treatments? How do these distributions compare by mode?[10]

The universe for the distributions will be all persons for which we have reported data for the socioeconomic and demographic characterstics listed below:

- Hispanic Origin
- Race
- Age
- Sex

We will also examine these household-level characteristics:
- Tenure
- Language of response
- Average household size

---

[10] As part of our analysis, we will also examine cross classifications of these demographic characteristics.

The related metrics will be weighted using replicate base weights adjusted for CAPI sub-sampling.

Research Question 6

How are responding units in the CFU reinterview distributed by socioeconomic and demographic characteristics? How do these distributions compare between the treatments? How do these distributions compare by mode of original interview?

Education and household size are also indicators of response propensity; however, since the Educational Attainment and Household Roster topics are included in the Content Test and may yield different distributions among the treatments, we will not include those in this analysis. We will use chi-square tests of independence to test for differences in demographic distributions among all treatments.

3.2.1.4   Respondent Burden

Adding new questions to the data collection instruments may increase the average time it takes to complete the ACS. The same is true of the modified questions, some of which add details to to improve question clarity. The time to complete a survey is an important metric for assessing respondent burden. We will investigate respondent burden by answering the following questions.

Research Question 6

Is there a difference between completion times for the test questionnaires versus the control questionnaire? (internet and CAPI).

We will calculate and compare average completion times for these treatments by household size subgroups (1-2, 3-4, etc.).

Research Question 7

Is there a difference between completion times for the individual test topics? (internet and CAPI)?

We will calculate and compare average completion times for these individual test topics (Household Roster, Labor Force, etc.).

Research Question 8

Is there a difference in the number of times respondents access the help screens for the tested questions for the Test treatments versus the Control treatment? (internet only).

Accessing the help screen could indicate that the respondent did not understand the question being asked and needed further instructions. We will calculate and compare how frequently the help screens were accessed.[11]

<u>Research Question 9</u>

Is there a difference in breakoff rates for the tested questions? (internet and CAPI).

A breakoff occurs when the respondent leaves the interview before completing the survey. A breakoff can indicate potential problems with an individual question. If a breakoff occurs, we can see which screen the respondent was on and whether it was one of the topics being tested.

We will compare the breakoff rates among the Control and Test treatments.

<u>Research Question 10</u>

Is there a difference in the form completeness rates in the original interview for the Test treatments versus the Control treatment? (internet, paper, and CAPI).

For all respondent burden questions, we will compare treatments using two-tailed t-tests.

3.2.2   Topic-Level Analysis

To evaluate the new questions and proposed changes to existing questions, we will calculate a variety of metrics including benchmark comparisons, item missing data rates, response distributions, and response reliability. We will perform analyses of Spanish-language responses where sample size allows. Sections 3.2.2.1 through 3.2.2.5 describe the different metrics, in a general sense, and how they evaluate the effectiveness of a survey question. The metrics will be discussed with more specificity for each topic in Sections 4 through 13.

3.2.2.1   Benchmarks

We will compare data from our treatments to information from other data sources that are considered standard which provide a rough judgement of treatment accuracy (i.e., benchmarks). In general, we will perform observational comparisons of distributions and proportions which will allow us to determine if our results are grossly different from other sources.

Differences between the benchmark data sources and the 2022 ACS Content Test data such as universe, question wording, and type of data (survey or administrative records) will be noted in each analysis. This analysis plan discusses benchmark results that are available now, but the

---

[11] We exclude the CAPI response mode from this analysis because we are only interested in when the respondent accesses the help screen, not the interviewer.

U.S. Census Bureau

final analysis will use the latest or most appropriate benchmark data available at the time of the analysis.

3.2.2.2   Item Missing Data Rates

Respondents leave items blank for a variety of reasons including not understanding the question (clarity), unwillingness to answer a question as presented (sensitivity), and lack of knowledge of the data needed to answer the question. The item missing data rate (for a given item) is the proportion of eligible units (housing units for household-level items or persons for person-level items) for which a required response is missing.

For each item, it is important to carefully define both the universe of eligible units and the criteria that determine whether a response to that item is missing or not missing. Usually the definition of "missing" will include "Don't Know" and "Refused to Answer" from CAPI interviews, as well as paper and internet questionnaires where no answer was provided.

We will compare the item missing data rates via two-tailed t-tests.

3.2.2.3   Response Distributions

Comparing the response distributions between the Control version of a question and one of the Test versions of a question will allow us to assess whether the question change affects the resulting estimates.

For questions with categorical responses, proportion estimates will be calculated as:

$$\text{Category Proportion} = \frac{\text{Weighted count of valid responses in category}}{\text{Weighted count of all valid responses}}$$

For questions with numeric responses, we will define categories as ranges of valid responses (0, 1 to 99, 100 or higher, etc.). These range categories will be defined by the subject area analysts or based on the categories used for published tables. In addition, we may also compare means or medians for these types of questions.

When a research question asks about comparing the Control and Test results from a single yes/no question or item, we will conduct an independent two-tailed t-test based on the research question.

When a research question involves comparison between the Control and Test multi-category response distributions, comparisons will be made using a Rao-Scott chi-square test (first-order adjustment) that checks for a significant difference between two sample distributions (Rao & Scott, 1987). If the chi-square test indicates a significant difference between the Control and Test distributions, we will test for significant differences in the individual category proportions using two-tailed t-tests.

U.S. Census Bureau

3.2.2.4   Response Reliability

We measure the reliability of responses to a question by looking at response error. Response error occurs for a variety of reasons, such as flaws in the survey design, misunderstanding of the questions, misreporting by respondents, or interviewer effects. Simple response variance is the degree to which respondents answer a question inconsistently. Conversely, a question has good response reliability if respondents tend to answer the question consistently. Response bias is the degree to which respondents consistently answer a question incorrectly. Either response bias or simple response variance will be measured for a topic, not both. The reason for this is to minimize CFU respondent burden and breakoffs by avoiding two sets of questions for a given topic.

We will measure response error by comparing responses to the CFU reinterview with the corresponding original Content Test interview responses. The U.S. Census Bureau has frequently used content reinterview surveys to measure response error for large demographic data collection efforts, including previous ACS Content Tests (2010 and 2016), and the 1990, 2000, and 2010 Decennial Censuses (Dusch & Meier, 2012). Only units (households or persons) with a valid response in both the original interview and the CFU will be included in this analysis.

The measures of response error assume that those characteristics in question did not change between the original interview and the CFU reinterview. This assumption will not be true in a small percentage of cases. For instance, a household that reported having no health insurance in the original interview might have acquired health insurance before the CFU reinterview and then accurately report a different response in the CFU reinterview. To the extent that this assumption is incorrect, we assume that it is incorrect at similar rates between the Control and Test panels.

3.2.2.4.1   Response Variance

Re-asking the same question of the same respondent (or housing unit) allows us to measure simple response variance. The following measures of simple response variance are used:

- 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 have three or more mutually exclusive response categories, as a measure of overall reliability for the question. In the following table, let "Yes" indicate the unit is in the category of interest, according to the response from either the Content Test original interview or the CFU reinterview; "No" indicates the unit is not reported to be in the category.

Table 4. Interview/Reinterview Counts Used 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 will be weighted to make them more representative of the population.

We will calculate the gross difference rate for this response category as

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

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:

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

Attachment C

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

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

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

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

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

We can now express the IOI in terms of the count variables in Table 4:

$$IOI = \frac{GDR/2}{(p_1 q_2 + p_2 q_1)/2} \times 100 = \frac{n(b + c)}{(a + c)(c + d) + (a + b)(b + d)} \times 100$$

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

If the response categories do not match up exactly between the compared treatments (if the categories are redefined as part of the question revision), we will either collapse response categories to form equivalent categories for comparison, or we will do comparisons for the response categories where it makes sense.

So far, we have only discussed response error 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 will also measure the overall response reliability of a question using the L-fold index of inconsistency, IOI_L. We will look at the difference in IOI_L between treatments and test 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 5 shows a cross-tabulation of

the original interview and CFU results for a generic analysis topic. Note that if L = 2 then Table 5 is equivalent to Table 4.

Table 5. Cross-tabulation of Original Interview Results with CFU Results for a Question with Response Categories

		Original Interview categories						CFU totals
		1	2	...	J	...	L	
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

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

It can be shown that the IOI_L is a weighted sum of the L category IOI values (Biemer 2011, 50), but this formula is easier for calculation.

3.2.2.4.2   Response Bias

To measure response bias, the CFU questions need to elicit more accurate responses than in the original interview. This usually involves asking a "gold standard" question. [12] The CFU questions for the receipt of incomes for self-employment, rental property, public assistance, and retirement will come from the 2021 Annual Social and Economic Supplement of the

---

[12] A gold standard question is a question from an established survey or source where the response values are considered highly accurate.

Current Population Survey (CPS ASEC). For these topics, we consider the CFU responses to be the true value. We will calculate response bias via the net difference rate (NDR).

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:

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

The NDR can be negative, zero, or positive. If the NDR is significantly negative, this indicates that the Content Test original interview version of the question tends to result in an underestimate of the true proportion in a category. Conversely, if the NDR is significantly positive, the original interview question tends to result in an overestimate of the true proportion. If the NDR is zero (i.e., not significantly different from zero), this is an indication that the original interview question results in an unbiased estimate of the true proportion.

For topics measuring response variance, we will also calculate the NDR, but only to check that it is not significantly different from zero. 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 will use the following adjustment of the IOI, developed by Flanagan (2001):

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

3.2.2.5 Other Metrics

Some topics have additional analysis metrics that are topic-specific. Details about the additional metrics and associated research questions are discussed under each specific topic in Sections 4 through 13.

3.2.3 Editing, Weighting, and Hypothesis Testing

3.2.3.1 Editing

For the 2022 ACS Content Test, we will not be making our typical ACS edits. This is because we are interested in how changes to existing questions and differences between versions of new questions affect the unaltered responses provided directly from respondents. For this reason, we are also not imputing responses. Some edits may be applied to the non-topic data, such as calculating a person's age based on his or her date of birth, but edits should be minimal. [13]

---

[13] When talking about edits here we mean edits to the data sets before analysis. During the analysis phase, we may perform additional edits, such as collapsing categories, if it makes sense to do so.

3.2.3.2   Weighting

All estimates from the ACS Content Test will be weighted. The final content test weights will take into account the initial probability of selection (the base weight) and CAPI sub-sampling. There will also be an adjustment for CFU non-response for the CFU analysis. The weighting procedure will be the same one used for the 2016 ACS Content Test. For more information on the weighting procedure, see the ACS 2022 Content Test sampling and weighting specification (Keathley, 2022).

3.2.3.3   Hypothesis Testing

Most of our analyses will compare the Control and Test versions of questions using hypothesis testing. The two topics that are testing only one question version, Sewer and Solar Panels, will test for differences in specified subgroups using hypothesis testing. Hypothesis testing involves making an inference between two samples using means or proportions. Hypothesis testing involves a null hypothesis and an alternative hypothesis. In most of our tests, the null hypothesis is that the means or proportions of the two samples are not statistically different. The alternative hypothesis is that the observed estimates between the samples are different. This is called a two-tailed hypothesis test. All hypothesis testing for the 2022 Content Test analysis will be two-tailed t-tests.

3.2.3.3.1   Level of Significance and Statistical Power

The significance level for a single hypothesis test will be $\alpha=0.1$, which is the Census Bureau standard. The p-value is the probability of obtaining the sample statistic observed or one more extreme if the null hypothesis were true. If the hypothesis test results in a p-value of less than 0.1, we will reject the null hypothesis and conclude that the data strongly supports the alternative hypothesis. However, if the p-value is greater than or equal to 0.1, we can only "fail to reject" the null hypothesis. We cannot claim the null hypothesis to be true.

Statistical power is the probability of correctly rejecting the null hypothesis. We want the power to be as close to 1 as possible, while keeping the significance level at $\alpha=0.1$. This typically involves making sure the sample size is large enough to detect differences. However, larger sample sizes typically incur greater operational costs; the balance between statistical power and cost is carefully considered.

The 2022 ACS Content Test sample size provides enough statistical power (0.80) to detect a difference in the gross difference rates of at least two percentage points between the control and test groups (effect size). The sample size and power requirements for this test were established based on the questions for electric vehicles because it has the smallest universe of respondents among the topics being tested. We expect statistical testing on the other topics with larger universes to be able to detect smaller differences in the gross difference rates.

3.2.3.3.2   Multiple Comparisons

We will sometimes test a set of hypotheses simultaneously. For these multiple comparison tests, the probability of observing at least one significant result due to chance increases as the number of comparisons increases. We will attempt to control for this overall Type I error rate within a "family" (called a familywise error rate). The definition of a family will vary depending on what comparisons are being made. The most common families are across modes or across response categories.

To control the familywise error rate, we will adjust the resulting p-values using the Hochberg method. This procedure is an improvement upon the Bonferroni sequentially rejective procedure (Hochberg, 1988). PROC MULTTEST in SAS® provides a simple procedure to adjust the p-values using the Hochberg procedure (SAS, 2009). For more information on adjusting for multiple comparisons, see Westfall (2011).

4. HOUSEHOLD ROSTER

Authors: Kathleen Kephart (CBSM) and Broderick Oliver (DSSD)

The following section describes the specifics of the household roster question(s) in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the Test version of the question (Section 4.1).
- A discussion of the changes between the Control and Test versions (Section 4.2).
- The specific research questions and analysis methodology (Section 4.3).

4.1   LITERATURE REVIEW

The Census Bureau has never conducted a formal study of within household coverage in the ACS. The roster instructions used by the Census Bureau have changed very little since the late 1990s, while the complexity of household living arrangements has increased (Cherlin, 2010). Recent research produced by the Census Bureau's Undercount of Young Children Task Force reports that the 2009 ACS coverage ratio for all children ages 0 to 4 was 0.89, suggesting a substantial undercoverage problem (Jensen & Hogan, 2017; Jensen, 2019).

Further, the task force's research showed evidence of more undercoverage error for children who were not the biological or adopted child of the householder, such as grandchildren and children not related to the householder (for instance, foster children or the children of a roommate) (Jensen et al., 2018). These types of children may have more tenuous ties to the households in which they reside. The causes for the undercount of young children is unknown, but many conjecture that rostering error is one likely source.

In addition to undercounting young children, there is evidence of respondent confusion and burden when rostering a household. Cognitive testing conducted by Census Bureau staff indicates that respondents are sometimes confused about whom to include on the ACS roster (Ashenfelter et al., 2012; Clark, 2017).

The general approach of the ACS is to ask for a list of everyone living or staying at the address, and follow up with several coverage probes about anyone who may have been missed in this initial question. Research conducted in 2015 on the automated ACS instruments (internet, CATI and CAPI) shows that most people are rostered using a simple instruction such as, "The following questions are about everyone who is living or staying at <address>. First, create a list of people." However, over two million people (weighted count) were rostered as a result of one of the coverage probes (Clark, 2017). Without these probes, these people would have been omitted from the roster. Households with complex living situations have more difficulty with rostering.

The coverage probes are questions that are posed after the initial roster question at the beginning of the automated ACS instruments, such as, "Does anyone else live or stay there?" and "Other than the people listed below, is there anyone else staying there even for a short time?" Despite responding in the affirmative to these probes, 67.0 percent of respondents who said "yes" did not provide a name, and therefore, no one was added to the roster. This suggests that respondents either did not understand the coverage questions or did not want to provide the information for the additional person(s). Additionally, there is evidence of respondents likely not fully understanding the coverage probes, as respondents use the "back" button frequently while navigating the roster questions in the internet mode (Clark, 2017; Horwitz et al., 2013).

In addition to the question wording itself, some of the confusion about who to roster in a household may also be a result of the two-month rule that is used for the ACS, which is a modified de facto residence rule. The intent of the ACS current residence concept is to count everyone living or staying at the address on the day the survey is completed and to include people who might be away for a short vacation, business trip, or overnight sleepover. The use of the two-month reference in the instructions and rostering questions may be causing people who should be included in the household to be left off the roster when respondents incorrectly apply this reference period to a situation. For instance, there is an instruction to ignore the reference period if the person has no other place to live, but many respondents may not notice or may ignore this and leave these people off the roster.

Cognitive testing in Rounds 1 and 2 in preparation for the 2022 ACS Content Test found that the proposed version for the Content Test resulted in less confusion and burden for respondents than other proposed versions. A primary goal of the Content Test is to determine if the test version results in the inclusion of more groups that have been underrepresented in the ACS, without reducing data quality in other ways.

4.2 QUESTION CONTENT

Control and Test versions of each question are shown as they will appear on the paper questionnaire. The internet is shown after the paper version below. We removed the mention of "2 months" from the paper version and only included 2 months in the followup question if respondents said "yes" to the question on the *Another_home* screen in the internet or interviewer modes.

Figure 1. Control Version of the Household Roster Question (Paper)

How many people are living or staying at this address?

- **INCLUDE** everyone who is living or staying here for more than 2 months.

- **INCLUDE** yourself if you are living here for more than 2 months.

- **INCLUDE** anyone else staying here who does not have another place to stay, even if they are here for 2 months or less.

- **DO NOT INCLUDE** anyone who is living somewhere else for more than 2 months, such as a college student living away or someone in the Armed Forces on deployment.

Number of people

Fill out pages 2 – 7 for everyone, including yourself, who is living or staying at this address for more than 2 months. Then complete the rest of the form.

Figure 2. Test Version of the Household Roster Question (Paper)

How many people, including yourself, live or stay at this address?

INCLUDE...

- ✓ anyone not related to you, like roommates and other families.

- ✓ babies and children, related or unrelated, including grandchildren and foster children.

- ✓ everyone staying here now who has no other place to stay.

DO NOT INCLUDE anyone living somewhere else, such as...

- ✗ a college student living away.

- ✗ someone in the Armed Forces on deployment.

Number of people

Fill out pages 2 – 7 for everyone, including yourself, who is living or staying at this address. Then complete the rest of the form.

Attachment C

Figure 3. Roster_A Screen for the Control Version of the Household Roster Question (Internet)

Figure 4. Roster_A Screen for the Test Version of the Household Roster Question (Internet)

U.S. Census Bureau

Figure 5. Roster_B Screen for the Control Version of the Household Roster Question (Internet)

Figure 6. Roster_B Screen for the Test Version of the Household Roster Question (Internet)

Figure 7. Roster_C Screen for the Control Version of the Household Roster Question (Internet)

Figure 8. Roster_C Screen for the Test Version of the Household Roster Question (Internet)

Figure 9. Away_Now Screen for the Control Version of the Household Roster Question (Internet)

Figure 10. Away_Now Screen for the Test Version of the Household Roster Question (Internet)

Figure 11. Another_Home Screen for the Control Version of the Household Roster Question (Internet)

Figure 12. Another_Home Screen for the Test Version of the Household Roster Question (Internet)

Figure 13. More_Than_Two Screen for the Control Version of the Household Roster Question (Internet)

Figure 14. More_Than_Two Screen for the Test Version of the Household Roster Question (Internet)

Figure 15. Roster_Away Screen for the Test Version of the Household Roster Question (Internet)

Note: There is no Control version for this additional question that will be asked at the end of rostering if the respondent indicates that someone is not staying for more than 2 months.

4.3   RESEARCH QUESTIONS AND METHODOLOGY

The following section describes the specific analysis to be conducted on the household roster question(s) in the 2022 ACS Content Test. Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 4.3.1).
- Specifics of the item missing data rate analysis (Section 4.3.2).
- Specifics of the response distribution analysis (Section 4.3.3).
- Specifics of the response reliability analysis (Section 4.3.4).
- Other analysis planned for this topic (Section 4.3.5).

4.3.1   Benchmarks for Household Roster

None. There are no administrative records for roster that we can use as a benchmark.

4.3.2   Item Missing Data Rates for Household Roster

<u>Research Question 1</u>

Is the item missing data rate for population count, which asks how many people live at the address, different on paper returns for the Control and Test versions? (Analyze for paper mode).

<u>Note</u>: The population count of how many people live or stay is not asked on internet or CAPI. In these modes, the instrument begins with collecting the names of everyone living or staying there.

<u>Research Question 2</u>

Are the item missing data rates for the roster probes Roster_B, Roster_C, Away_now, Another_home, and More_than_two different on internet returns for the Control and Test versions? (Analyze for internet).

<u>Note</u>: We will compare the item missing data rate for each individual roster probe.

<u>Research Question 3</u>

Are the item missing data rates for the roster probes Roster_B, Roster_C, Away_now, Another_home, and More_than_two different on CAPI returns for the Control and Test versions? (Analyze for CAPI).

<u>Note</u>: We will compare the item missing data rate for each individual roster probe.

To test for the differences in item missing data rates in research questions 1-3, we will conduct two-tailed t-tests at the 0.10 level of significance for internet, paper, and CAPI.

4.3.3   Response Distributions for Household Roster

4.3.3.1   Count Discrepancy Research Questions

The paper questionnaire asks respondents to first report the total number of people living at the address (population count) and then asks the respondent to roster the people (report their names and demographics).

A *count discrepancy* occurs when the number of people provided by the respondent in the population count question does not match the number of people rostered.

A *high count discrepancy* occurs when the number of people rostered is greater than the number provided in the population count.

A *low count discrepancy* occurs when the number of people rostered is less than the provided population count.

Research Question 4

Are the number of cases with a count discrepancy for paper returns for the Control and Test versions different?

Research Question 5

Are the number of cases of high count discrepancy for paper returns for the Control and Test versions different?

Research Question 6

Are the number of cases of low count discrepancy for paper returns for the Control and Test versions different?

To test for differences in count discrepancy rates in research questions 4-6 (paper mode only), we will conduct two-tailed t-tests at the 0.10 level of significance.

4.3.3.2   Household Characteristics Research Questions

Research Question 7

Are the number of complex households for the Control and Test versions different? (Analyze overall and by mode: paper, internet, and CAPI).

A *complex household* is defined as a household with a structure other than one in which: 1) a householder lives alone; 2) a householder lives with a married or unmarried partner without children; 3) a householder lives with a married or unmarried partner with biological or adopted children; and 4) the householder is a single parent with biological or adopted children.

Examples of complex households include:
- Blended families (i.e., stepchildren)
- Multi-generational households (grandparents, parents, and children)
- Family with other relatives (i.e., aunts, uncles, cousins, etc.)
- Skip generation (grandparents with grandchildren, no parents present)
- Family with other nonrelative(s)

Research Question 8

Are the number of people with tenuous connections to the household for the Control and Test versions different? (Analyze overall and by mode: paper, internet, and CAPI).

A *tenuous connection* will be defined based on each person's relationship to Person 1.

This includes people who are not a partner or biological or adopted children of Person 1, such as a brother or sister, stepchildren, father or mother, parent-in-law, son-in-law or daughter-in-law, other relative, roommates, foster children, and other nonrelatives.

Research Question 9

Is there a difference in the number of children between the ages of 0-4 on the final roster for the Control and Test versions? (Analyze overall and by mode: paper, internet, and CAPI).

Research Question 10

Is there a difference in household size distributions between Control and Test versions? (Analyze overall and by mode: paper, internet, and CAPI).

We will compare the proportion of 1-person, 2-person, 3-person, 4-person, and 5-person or larger households.

Research Question 11

Are the person characteristics between the Control and Test versions different? (Analyze overall and by mode: paper, internet, and CAPI).

Characteristics include the relationship to reference person for each person, by relationship category. Partners will include married, unmarried, and opposite and same-sex partners. Biological and adopted children will also be grouped, and the other relationship categories will be examined separately. We will also look at person-level characteristics for age group, race, Hispanic origin, sex, educational attainment, and English-speaking ability for all members of the household.

4.3.3.3   Roster Add or Delete Research Questions

Research Question 12

Is there a difference in the number of people who were originally rostered on screen Roster_a between the Control and Test versions, independent of if they stayed on the roster or were later deleted? (Analyze by mode: internet and CAPI).

Research Question 13

Is there a difference between the Control and Test versions in how often internet respondents answer "Yes" to one of the filter questions (Away_now, Another_home, More_than_two, Roster_away), suggesting roster changes are needed, but do not provide a name to add/delete from the roster? (Analyze overall for each screen and analyze by mode: internet and CAPI).

Research Question 14

Is there a difference in the number of people who were added, and kept on the final roster between the Control and Test versions? (Analyze by mode: internet and CAPI).

Research Question 15

Is there a difference in the number of people who were added that are later deleted between Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Research Question 16

Is there a difference in the number of people who were deleted from the roster between the Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Research Question 17

Is there a difference in the number of households with an added household member between Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Research Question 18

Is there a difference in the number of households with a deleted household member between Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Research Question 19

Is there a difference in the number of young children (0-4 years) added to the roster between the Control and Test versions? (Analyze overall and by mode: internet and CAPI, and by screen Roster_A, Roster_B, and Roster_C).

Research Question 20

Is there a difference in the number of households that added a young child (0-4 years) to the roster between the Control and Test versions? (Analyze by mode: internet and CAPI, and by screen Roster_A, Roster_B, Roster_C).

Research Question 21

Across modes and screens, is there a difference in the distribution of biological and adopted children versus all other relationship statuses (relative to Person 1) for young children (0-4 years) who were added between the Control and Test versions?

Research Question 22

Across modes and screens, is there a difference in the distribution of biological and adopted children versus all other relationship statuses (relative to Person 1) for young children (0-4 years) who were removed between the Control and Test versions?

Research Question 23

Is there a difference in the number of households that deleted a young child (0-4 years) from the roster between the Control and Test versions? (Analyze by mode: internet and CAPI, and by screen Roster_A, Roster_B, Roster_C).

Research Question 24

For the universe of households that added a person, is there a difference in the number of persons per household added between the Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Research Question 25

For the universe of households that deleted a person, is there a difference in the number of persons per household deleted between the Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Research Question 26

Is there a difference in the characteristics of households that added someone to the roster between the Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Characteristics include household size, tenure, building type, if no one in household over the age of 14 speaks English very well, and household type (complex versus non-complex).

Research Question 27

Is there a difference in the characteristics of households that deleted someone from the roster between the Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Characteristics include household size, tenure, building type, if no one over the age of 14 speaks English very well, and household type (complex versus non-complex).

Research Question 28

Is there a difference in the characteristics of the people that are added to the roster between the Control and Test versions? (Analyze overall and by mode: internet and CAPI).

Characteristics include relationship to reference person, age, race, Hispanic origin, sex, educational attainment, and English-speaking ability.

4.3.4   Response Reliability for Household Roster

The CFU reinterview will collect an independent roster by asking the same roster questions again. Respondents who received the Test version will be asked the Test version roster questions again. Respondents who received the Control version will be asked the Control version roster questions again.

Research Question 29

When we match the Content Test roster to the CFU independent roster by name and date of birth for each person within the household, is the proportion of households with a mismatch between the Content Test and CFU different between the Control and Test versions? (Analyze overall and by mode of original interview: paper, CAPI, and internet).

Research Question 30

When we match at the person level between the Content Test roster and the CFU independent roster, what is the distribution of age groups (0-4, 5-17, 18-25, 25-65, 65 plus) for people who are not matched for the Control and Test versions? (Analyze overall and by mode of original interview: paper, CAPI, and internet).

Research Question 31

When we match at the person level between the Content Test roster and the CFU independent roster, what is the distribution of relationship type relative to Person 1 for people who are not matched for the Control and Test versions? (Analyze overall and by mode of original interview: paper, CAPI, and internet).

4.3.5   Other Metrics for Household Roster

Two additional exploratory questions are being asked in the CAPI and internet modes to understand more about situations where people add or delete someone from the roster. The results of these questions will not be used to make any decisions about the implementation of the Test version.

Research Question 32

Overcount Followup

[If yes to Away_now or Another_home]
We are conducting research to understand why people stay in more than one place. Earlier in the survey you indicated that <name from questions Away_now and Another_home> sometimes live(s) somewhere else or is(are) only staying here for a short time? Could you briefly explain <name from Away_now and Another_home>'s living situation?

Research Question 33

Undercount Followup

[If anyone is added on Roster_B or Roster_C]
We are conducting research to understand why people may be left off of a roster. Earlier in the survey you indicated that <additional person's NAME from Roster_b and Roster_c> were not initially listed as living or staying here/there. Could you briefly explain <additional person's NAME from Roster_b and Roster_c>'s living situation?

4.4 DECISION CRITERIA

The most important results of this analysis when drawing a conclusion about the Test version compared to the Control version are:

Table 6. Decision Criteria for Household Roster

Priority	Research Question	Decision Criteria
1	9	The number of children between the ages of 0-4 on the final roster is significantly higher in the Test version compared to the Control version.
2a	30	The proportion of households with a mismatch between the Content Test and CFU roster is lower in the Test version than the Control version.
2b	8	The number of people with tenuous connections to the household is significantly higher in the Test version compared to the Control version.
3	7	The number of complex households for the Test version is significantly higher than the Control version.
4	4	The number of cases with a count discrepancy for the Test version is lower than the number of cases with a count discrepancy for the Control version.
5	1, 2, 3	The item missing data rates for the Test version are the same or lower than the item missing data rates for the Control version.

Research questions not included in the decision criteria are for informational purposes only.

4.5 REFERENCES

Ashenfelter, K., Quach, V., Holland, T., Nichols, E., & Lakhe, S. (2012). *Report for round 3 of usability testing of the 2011 American Community Survey online instrument: Focus on login and roster features*. U.S. Census Bureau. Retrieved June 14, 2018, from https://www.census.gov/srd/papers/pdf/ssm2012-10.pdf

Cherlin, A. J. (2010). Demographic trends in the United States: A review of research in the 2000s. *Journal of Marriage and Family*, 72(3), 403-419. Retrieved February 23, 2022 from, https://www.doi.org/10.1111/j.1741-3737.2010.00710.x

Clark, S. (2017). *Analysis of the household roster questions on the ACS*. U.S. Census Bureau. Retrieved June 14, 2018, from https://www.census.gov/library/working-papers/2017/acs/2017_Clark_01.html

Jensen, E.B., & Hogan, H.R. (2017). The coverage of young children in demographic surveys. *Statistical Journal of the IAOS*, 33 (2), 321-333. Retrieved February 8, 2022, from https://www.doi.org/10.3233/sji-170376

Jensen, E., Schwede, L., Griffin, D., & Konicki, S. (2018). *Investigating the 2010 undercount of young children: Analysis of complex households*. U.S. Census Bureau. Retrieved February 8, 2022, from https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-2010-undercount-children-complex_households.pdf

Jensen, E. (2019). *Investigating the 2010 undercount of young children: Examining coverage in demographic surveys*. U.S. Census Bureau. Retrieved February 8, 2022, from https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-2010-undercount-children-examining_coverage_demo.pdf

Horwitz, R., Tancreto, J.G., Zelenak, M.F., & Davis, M. (2013). *Use of paradata to assess the quality and functionality of the American Community Survey internet instrument*. U.S. Census Bureau. Retrieved June 14, 2018, from https://www.census.gov/content/dam/Census/library/working-papers/2013/acs/2013_Horwitz_01.pdf

5. SOLAR PANELS

Authors: William Chapin (SEHSD) and Michael Risley (DSSD)

The following section describes the specifics of the Solar Panel question in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the addition of the Solar Panels question in the 2022 ACS Content Test (Section 5.1).
- A discussion of the question content (Section 5.2).
- The specific research questions and analysis methodology (Section 5.3).

5.1 LITERATURE REVIEW

The question on Solar Panels is a new question proposed for inclusion in the ACS and has never been asked in the Decennial Census. A similar question is asked on the American Housing Survey (AHS) as well as the Residential Energy Consumption Survey (RECS), but they only produce estimates on the national level. By asking this question on the ACS, we will be able to obtain data for operational solar panels on a housing unit level at smaller geography levels. This information will help the Energy Information Administration (EIA) match energy consumption to energy production across the United States. The differential adaptation of energy producing technologies such as rooftop solar (i.e., photovoltaic generating capacity) creates new, more variable demands and new potential for energy infrastructure for and in U.S. households.

There is only one version of this question being tested due to the succinct nature of both the question and the response categories. We will be comparing results from the Content Test with data obtained in the AHS as data is provided every two years. The question was first asked in the AHS in 2017, where 2.6 percent of occupied units reported having solar panels. That number increased to 3.0 percent in 2019 (American Housing Survey, 2020).

We will also compare data with the RECS where possible, as data is not provided in any report, but data may be obtained from the microdata.

5.2 QUESTION CONTENT

The Solar Panels question is shown as it appears on the paper questionnaire. Automated versions of the questionnaire will have the same content formatted accordingly for each mode. For the modes where it is possible, help text will be available.

Figure 16. Solar Panels Question (Paper)

> **Does this house, apartment, or mobile home use solar panels that generate electricity?**
>
> ☐ Yes
>
> ☐ No

Figure 17. Solar Panels Question (Internet/CAPI)

> **23. Does this <mobile home/house/apartment/unit> use solar panels that generate electricity?**
>
> o Yes
> o No

5.3   RESEARCH QUESTIONS AND METHODOLOGY

The following section describes the specific analysis to be conducted on the Solar Panels question in the 2022 Content Test. Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 5.3.1).
- Specifics of the item missing data rate analysis (Section 5.3.2).
- Specifics of the response distribution analysis (Section 5.3.3)
- Specifics of the response reliability analysis (Section 5.3.4).
- Other analysis planned for this topic (Section 5.3.5).

5.3.1   Benchmarks for Solar Panels

Research Question 1

How does the percentage of housing units with solar panels compare to the proportions found in the most recent American Housing Survey (AHS) overall and for owners and renters?

The AHS asks a similar question to obtain data for housing units on solar panels. However, the methodology of the AHS is different from the ACS in ways that make the estimates not directly comparable.

- The AHS is a longitudinal survey that follows the same houses over time and data is only collected every other year. Notably, the AHS last collected household data in 2021 and

the data will be an entire year older than the data collected in the Content Test. This is especially an issue for a rapidly changing topic like solar panels.

- The AHS has no internet or paper questionnaire response option, respondents are only able to respond by interview, either in-person or over the phone.

- AHS has different editing procedures that may lead to different estimates.

Because of the differences in methodology, we will compare the AHS data to the data collected in the Content Test nominally. Included in this review will be comparisons by tenure, which AHS includes in their table creation tool, looking at owned units versus rented units.

Research Question 2

How does the item missing data rate for the new solar panels question compare to the item missing data rate for the solar panels question on the most recent American Housing Survey (AHS)?

The AHS asks a similar question to obtain data for housing units on solar panels. Differences in the data collection methodology are described for Research Question 1. Notably for this question, AHS only collects data through interview, which tends to have different item missing data rates compared to internet and paper questionnaire modes. We will use the data on those that did not answer the solar panels question on the AHS to compare with the data collected in the Content Test.

Research Question 3

How does the percentage of housing units with solar panels compare to the percentages found in the most recent Residential Energy Consumption Survey (RECS) overall and for owners and renters?

The RECS collects data on units with solar panels, however the methodology of RECS data collection is different in ways that makes the estimates not directly comparable.

- The RECS household survey occurs roughly every five years. Notably, the RECS last collected household data in 2020 and the data will be two years older than the data collected in the Content Test. This is especially an issue for a rapidly changing topic like solar panels.

- In comparison to the ACS, the RECS household survey has a much smaller sample at 5,686 households in 2015. For solar panels, which has a relatively low prevalence, this would result in relatively high variance for the estimate.

- RECS has different editing and imputation procedures, especially due to the comparatively small sample size, which may lead to different estimates.

Additionally, the 2015 RECS did not produce any data tables on solar panel prevalence. If the next data release also does not include or if the next data release is not available in time, we will use the microdata to create data to use for comparison. Included in this review will be comparisons by type of building and tenure, which are both available on the microdata. For tenure, we will look at owned units versus rented units.

Rates will be compared nominally.

5.3.2   Item Missing Data Rates for Solar Panels

<u>Research Question 4</u>

Is there a difference in the item missing data rate different between modes of response?

We will compare the non-response rates for solar panels by the different modes of collection (self-administered versus interviewer-administered).

<u>Research Question 5</u>

Is there a difference in item missing data rate different between selected housing demographics?

We will compare by household demographics. Included in this review will be comparisons by type of building and tenure. For type of building there will be two categories compared: Single-Family house (attached and detached combined) and apartments. For tenure, we will look at owned units versus rented units.

5.3.3   Response Distributions for Solar Panels

<u>Research Question 6</u>

Are the percentages of housing units with Solar Panels different between selected housing demographics?

We will compare by householder and housing demographics. Included in this review will be comparisons by income, type of building, and tenure. For type of building there will be two categories compared: single-family house (attached and detached combined) and apartments. For tenure, we will look at owned units versus rented units.

5.3.4   Response Reliability for Solar Panels

<u>Research Question 7</u>
Is there a difference in response reliability between self-administered and interviewer administered responses?

We will compare response reliability (Gross Difference Rate and Index of Inconsistency) by type of response. This question should have relatively high response reliability compared to the other items tested, considering the dichotomous response categories and the low question sensitivity.

5.3.5   Other Metrics for Solar Panels

None

5.4   DECISION CRITERIA

The most important results of this analysis when evaluating this new question are, in order of priority:

Table 7. Decision Criteria for Solar Panels

Priority	Research Question	Decision Criteria
1	1, 3	The solar panel prevalence rate should be similar to the AHS and RECS benchmarks overall and for selected housing demographics. A difference between the prevalence and one of the benchmarks being larger than the difference between the two benchmarks would be of particular concern. Due to a possible correlation between solar panel usage and unit nonresponse, a prevalence rate that is lower than the benchmarks would be a larger concern than one that is higher.
2	2, 4, 5	The item missing data rate should be low and comparable to the item missing data rate in the AHS.
3	7	The question should have high response reliability. For IOI this would generally be a value less than 20.
4	6	Rates should be comparable when crossed by type of building within tenure.

Research questions not included in the decision criteria are for informational purposes only.

5.5   REFERENCES

U.S. Census Bureau (2020). *American Housing Survey Table Creator*. Retrieved November 29, 2021, from https://www.census.gov/programs-surveys/ahs/data/interactive/ahstablecreator.html

6. ELECTRIC VEHICLES

Authors: Molly Cromwell (SEHSD) and Lindsay Longsine (DSSD)

The following section describes the specifics of the Electric Vehicles questions in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the addition of the Electric Vehicles questions in the 2022 ACS Content Test (Section 6.1).
- A discussion of the differences between the two versions of the Electric Vehicles question (Section 6.2).
- The specific research questions and analysis methodology (Section 6.3).

6.1 LITERATURE REVIEW

The proposal for adding a question about the use of electric vehicles to the ACS was submitted by the Energy Information Administration (EIA). The EIA is tasked with determining the current national energy supply and whether it meets the demands of the country. These demands are projected to increase greatly in the next 10-20 years. The EIA is able to collect data from state and local governments but lacks the resources to collect information at smaller levels (e.g., housing units and public buildings).

Determining the prevalence of electric vehicles at the housing-unit level will help the EIA to understand the energy supply, demand, and current technology of energy resources when making projections for future energy needs. Understanding energy consumption at these lower levels also helps in making the technological and capital changes necessary within the energy infrastructure. Additionally, the prevalence of electric vehicle ownership may help to evaluate the effectiveness of energy and tax policies.

The administration has made it a goal that half of all vehicle sales in 2030 will be zero-emissions vehicles, which includes plug-in hybrid electric and all-electric vehicles. In addition to heavily promoting zero-emissions vehicle sales, the current administration will also prioritize electric vehicle manufacturing and the establishment of electric vehicle infrastructure in the country, which will include installing the first-ever national network of electric vehicle charging stations. By adding the electric vehicles question to the ACS, we will be able to provide data showing which communities have the highest needs for these charging stations (U.S. Census Bureau, 2018).

6.2 QUESTION CONTENT

The Electric Vehicles question is new to the ACS and has not been tested using the ACS framework. There will be two versions field tested in the 2022 ACS Content Test. The first version of the question separates electric vehicles down into plug-in electric and hybrid-electric categories. This version has been cognitively tested.[14] The second version asks respondents whether they own an either an all-electric vehicle or a plug-in electric vehicle. This version of the question uses the same wording as the electric vehicle question on the Residential Energy Consumption Survey (RECS).

Figure 18. Control Version of the Electric Vehicles Question (Paper)

Are any of the following types of electric
vehicles kept at home for use by members
of this household?

a. A plug-in electric vehicle?

☐ Yes

☐ No

b. A hybrid electric vehicle?

☐ Yes

☐ No

Figure 19. Test Version of the Electric Vehicles Question (Paper)

Do you or any member of this household own or
lease an electric vehicle? *Include both all-electric
and plug-in hybrid electric vehicles.*

☐ Yes

☐ No

---

[14] It is possible a housing unit will have two electric vehicles (one plug-in and one hybrid) or they could have one vehicle that is both a plug-in and hybrid. We will not be able to distinguish between these two scenarios. We really only care if any plug-in electric vehicle exists at the housing unit, not whether it is also a hybrid or the number of electric vehicles. The one scenario that could be problematic is if the housing unit has a vehicle that is both, but the respondent only marks yes to hybrid. This scenario did not come up in cognitive testing and we are unsure how often it would occur.

6.3   RESEARCH QUESTIONS AND METHODOLOGY

The following section describes the specific analysis to be conducted on the Electric Vehicles question in the 2022 Content Test. Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 6.3.1).
- Specifics of the item missing data rate analysis (Section 6.3.2).
- Specifics of the response distribution analysis (Section 6.3.3).
- Specifics of the response reliability analysis (Section 6.3.4).

6.3.1   Benchmarks for Electric Vehicles

Research Question 1

How does the percentage households with of plug-in electric vehicles compare to the proportions found in the most recent RECS?

The wording for Version 2 of the Electric Vehicles question is borrowed from the RECS. The RECS survey is conducted by the Energy Information Administration through in-person interviews, as well as internet and mail survey forms. Every five years, a nationally representative sample of households is surveyed regarding their energy use patterns and household demographics. Household energy suppliers are also surveyed to estimate energy cost and usage patterns. This sample is much smaller than the annual ACS sample (roughly 5,700 housing units for RECS vs roughly 3.5 million housing units for ACS). There are also differences in how the RECS and the ACS edit and impute data.

Because of these design differences, we cannot make direct comparisons between the electric vehicles rates. Thus, rates of plug-in electric vehicle usage found in the most recent RECS survey and those observed during field testing will be compared nominally overall and by mode.

6.3.2   Item Missing Data Rates for Electric Vehicles

Research Question 2

Is the overall item missing data rate different between Version 1 and Version 2?

For the mail and internet formats of the questions, a missing item response will be indicated when there are no response boxes checked.[15] For CAPI, a response of either "Don't Know" or "Refused" will be considered a missing item response. To test for differences in the rate of

---

[15] For Version 1, we will consider the response missing if no boxes are checked for both subsections of the question. This way we test the question as a "whole". Since we really care about plug-in electric vehicles, we may also compare item missing data rates using only part a of Version 1 (ignoring part b).

missing item responses between Version 1 and Version 2 of the question, our analysis will involve conducting two-tailed t-tests at the 0.1 level of significance.

<u>Research Question 3</u>

Are the item missing data rates different between Version 1 and Version 2 when dividing responses into self-administered (paper and internet) versus interviewer administered (CAPI)?

To test for differences between the item missing data rates of Version 1 and Version 2, we will conduct two-tailed t-tests at the 0.1 level of significance.

6.3.3   Response Distributions for Electric Vehicles

<u>Research Question 4</u>

Are the percentages of households with a plug-in electric vehicle different between Version 1 and Version 2?

For this question, we will compare the percentage of housing units that marked 'yes' to Version 2 to the percentage of households that marked 'yes' to part a of Version 1. To test for differences between the proportions of households with a plug-in electric vehicle in Version 1 and Version 2, we will conduct two-tailed t-tests at the 0.1 level of significance. We will also compare cross-sectionally by household demographics between Version 1 and Version 2. Some household demographics that will be included in the analysis are units in structure, tenure, year built, persons per household, and household income.

6.3.4   Response Reliability for Electric Vehicles

<u>Research Question 5</u>

Are the measures of response reliability different between Version 1 and Version 2?

To evaluate whether the measure of response reliability is different between Version 1 and Version 2, we will look at simple response variance (how consistently are respondents answering the questions). As with research question 4, we will only be using part a of Version 1 in our comparison. Statistical significance between the GDRs of each version will be determined using a two-tailed t-test at the 0.1 level of significance. We will compare by mode pending a sufficient number of CFU reinterviews.

6.3.5   Other Metrics for Electric Vehicles

None

6.4 DECISION CRITERIA

The most important results of this analysis when drawing a conclusion about the two Test versions are shown below in order of priority.

Table 8. Decision Criteria for Electric Vehicles

Priority	Research Question	Decision Criteria
1	4, 5	The version which has the highest rate of plug-in electric vehicles and lowest response variance is preferrable. We expect underreporting would be more likely than overreporting, so a higher estimate is more likey to be accurate, but only if the response variance is not statistically higher.
2	2, 3	The version that has lower item missing data rates is preferable.

Research questions not included in the decision criteria are for informational purposes only.

6.5 REFERENCES

U.S. Census Bureau (2018). *Preparing for cognitive testing of items included in the 2021 ACS Content Test*. Retrieved December 8, 2021, from https://sharepoint.ecm.census.gov/ teams/acso/MethodsPanel/2022ACT/Topics/Forms/AllItems.aspx

U.S. Census Bureau

7. SEWER

Authors: Molly Cromwell (SEHSD) and Broderick Oliver (DSSD)

The following section describes the specifics of the Sewer question in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the addition of the Sewer question in the 2022 ACS Content Test (Section 7.1).
- A discussion of the question content (Section 7.2).
- The specific research questions and the analysis methodology (Section 7.3).

7.1   LITERATURE REVIEW

The Sewer question is a new question proposed for inclusion in the ACS. The Sewer question appeared most recently on the 1990 Census long-form questionnaire.[16] Between 1970 and 2000, the decennial censuses utilized two forms for data collection: the short-form questionnaire and the long-form questionnaire. Most households received the short-form questionnaire while one in every six households received the long-form questionnaire, which allowed for the collection of more detailed demographic and household characteristics. Beginning in 2005, the ACS began collecting these long-form data. As a result, beginning with the 2010 Census, the decennial census collects only short-form questionnaire data.[17]

While the 1990 Census was the last time sewer data had been collected on the Census, the American Housing Survey (AHS) has been consistently asking respondents about their public sewer and septic tank status since the 1970s. However, the lowest level of geography available from the AHS is Metropolitan Statistical Areas. Consistent data on the decentralized wastewater infrastructure status in rural and other small communities is needed to protect public health and water quality.

Due to the level of geography at which the ACS data can be collected, adding this question to the ACS may help to address changes in housing development and to support regular planning and funding cycles at the local, state, and national level. Additionally, determining the prevalence of existing septic systems and obtaining periodic updates on new septic system construction at the local level would assist in meeting the country's growing infrastructure needs.

---

[16] Historical Census of Housing Tables: Sewage Disposal
[17] Questionnaires - History - U.S. Census Bureau

U.S. Census Bureau

7.2 QUESTION CONTENT

This is a potential new question on the American Community Survey. Only one version of this question will be tested. The paper version of this question is shown in Figure 20. The internet version of this question is shown in Figure 21. The CAPI version of this question is shown in Figure 22.

Figure 20. Sewer Question (Paper)

Is this house, apartment, or mobile home connected to a public sewer?

☐ Yes, connected to public sewer

☐ No, connected to septic tank

☐ No, use other type of system

Figure 21. Sewer Question (Internet)

Is this <mobile home/house/apartment/unit> connected to a public sewer?

○ Yes, connected to public sewer
○ No, connected to septic tank
○ No, use other type of system

Figure 22. Sewer Question (CAPI)

14a. Is this <mobile home/house/apartment/unit> connected to a public sewer?

○ Yes
○ No

14b. Is it connected to a septic tank?

○ Yes
○ No

7.3   RESEARCH QUESTIONS AND METHODOLOGY

The following section describes the specific analysis to be conducted on the Sewer question in the 2022 ACS Content Test. Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 7.3.1).
- Specifics of the item missing data rate analysis (Section 7.3.2).
- Specifics of the response distribution analysis (Section 7.3.3).
- Specifics of the response reliability analysis (Section 7.3.4).

7.3.1   Benchmarks for Sewer

Research Question 1

Is the proportion of households connected to a public sewer, septic tank, or other type of sewage system in the 2022 ACS Content Test different from the respective proportions in the American Housing Survey (AHS)?

We will compare national level estimates and subgroup estimates like building type and tenure (owner versus renter) for occupied households from both surveys. The comparisons will be nominal only due to differences in sampling, data collection, question wording, and other factors.

Since the control and test versions of the Sewer question are identical, we will combine the control and test samples to produce a single sample for this analysis.

7.3.2   Item Missing Data Rates for Sewer

Research Question 2

What are the missing item data rates for the Sewer question, overall and by these subgroups: renters and homeowners? How do these rates compare to the item missing data rate for the telephone availability question that follows the Sewer question on the questionnaire and has similar response categories?

The universe is occupied households.

For CAPI, the Sewer question will be considered not filled out if both questions are not filled out.

To test for differences, we will conduct two-tailed t-tests at the 0.10 level of significance. We will combine the control and test samples to produce a single sample for this analysis.

7.3.3   Response Distributions for Sewer

No questions. Research question 1 addresses the response distributions of interest adequately.

7.3.4   Response Reliability for Sewer

<u>Research Question 3</u>

Are the answers provided by respondents to the Sewer question in the original interview consistent with the answers they provided in the CFU?

We will combine the control and test samples to produce a single sample for this analysis. We will test response variance using the Gross Difference Rate and Index of Inconsistency metrics at the 0.10 level of significance.

The analysis will be conducted overall and by these subgroups: renters and homeowners.

7.3.5   Other Metrics for Sewer

None

7.4   DECISION CRITERIA

The most important results to consider in determining if the new question on Sewer should be included in the ACS are, in order of priority:

Table 9. Decision Criteria for Sewer

Priority	Research Question	Decision Criteria
1	1	The proportion of households connected to a public sewer, septic tank, or other type of sewage is about the same nominally as the respective proportions in the American Housing Survey (AHS).
2	2	The overall item missing data rate is about the same nominally as the item missing data rate for the telephone availability question that follows the Sewer question on the questionnaire.
3	3	The measure of simple response variance, GDR and IOI do not provide an indication of potential problems with questionnaire wording.

Research questions not included in the decision criteria are for informational purposes only.

7.5   REFERENCES

U.S. Census Bureau (1990). *Historical census of housing tables: Sewage disposal*. Retrieved December 8, 2021, from https://www.census.gov/data/tables/time-series/dec/coh-sewage.html

U.S. Census Bureau (2021). *Questionnaires*. Retrieved December 8, 2021, from https://www.census.gov/history/www/through_the_decades/questionnaires/

8. EDUCATIONAL ATTAINMENT

Authors: Kevin McElrath, Jacob Fabina (SEHSD), and Broderick Oliver (DSSD)

The following section describes the specifics of the Educational Attainment question in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the Test version of the question (Section 8.1).
- A discussion of the difference between the Control and Test versions (Section 8.2).
- The specific research questions and the analysis methodology (Section 8.3).

8.1   LITERATURE REVIEW

A question on Educational Attainment has been asked on the ACS since the survey began in 2005. The current version of the question has been asked since 2008. Ongoing research at the U.S. Census Bureau shows that a relatively high percentage of people are selecting the response category "No schooling completed" in the self-response modes of the ACS, including adults who probably have completed some schooling. This testing evaluates whether changes to the lowest response categories can reduce the number of cases where respondents with some schooling completed the "No schooling completed" response category.

Starting in 2008, headers were added for each section of educational attainment. These headers included "No schooling completed", which was added as a header during the 2006 ACS Content Test after cognitive testing suggested that respondents had difficulty locating the "No schooling completed" category (Crissey, Bauman, and Peterson, 2007). The Educational Attainment report for the 2006 ACS Content Test noted a higher rate of "No schooling completed" in the revised Educational Attainment question.

In 2007, 0.71 percent of persons 25 years and older who responded by mail selected "No schooling completed." In 2008, following the revision, 1.52 percent of persons 25 years and older who responded by mail selected "No schooling completed"; and in 2016, 1.22 percent of persons 25 years and older who responded by internet and 2.26 of persons who responded by mail selected "No schooling completed". The percent of persons 25 years and older who selected "No schooling completed" with the CATI and CAPI response modes were 0.68 and 0.88 percent, respectively in 2007; 0.66 and 0.83, respectively in 2008; and 0.65 and 0.85, respectively in 2016.

Immigrants are likely to have different levels of educational attainment than the native born, as immigrants often immigrate to the U.S. to fill specific high- or low-skill labor niches (Lee and Zhou, 2015). Hence, the foreign-born population is an important subpopulation to test the lower response categories in the Educational Attainment question.

Figure 24. Test Version of the Educational Attainment Question (Paper)

What is the highest grade of school or degree this person has **COMPLETED?** *If currently enrolled, select the previous grade or highest degree received.* Mark (X) ONE box.

LESS THAN GRADE 1

☐ Less than grade 1

GRADE 1 THROUGH GRADE 12

☐ Grade 1 through 11 – *Specify grade 1 – 11*

☐☐

☐ 12th grade – **NO DIPLOMA**

HIGH SCHOOL GRADUATE

☐ Regular high school diploma

☐ GED or alternative credential

COLLEGE OR SOME COLLEGE

☐ Some college credit, but less than 1 year of college credit

☐ 1 or more years of college credit, no degree

☐ Associate's degree *(for example: AA, AS)*

☐ Bachelor's degree *(for example: BA, BS)*

AFTER BACHELOR'S DEGREE

☐ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*

☐ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*

☐ Doctorate degree *(for example: PhD, EdD)*

8.3 RESEARCH QUESTIONS AND METHODOLOGY

The following section describes the specific analysis to be conducted on the Educational Attainment question in the 2022 Content Test. Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 8.3.1).
- Specifics of the item missing data rate analysis (Section 8.3.2).
- Specifics of the response distribution analysis (Section 8.3.3).
- Specifics of the response reliability analysis (Section 8.3.4).

8.3.1 Benchmarks for Educational Attainment

Research Question 1

How do the estimates of the percentage of the population with educational attainment less than grade 1 in the Control and Test treatments compare with the estimate in the Current Population Survey (CPS)?

The benchmark measure will come from the CPS October Supplement, Table 1 of the CPS Educational Attainment tables: "Educational Attainment of the Population 18 Years and Over, by Age, Sex, Race, and Hispanic Origin". We will make nominal comparisons between the 2022 ACS Content Test estimates and the CPS estimates of the percentage of the population with educational attainment that is less than grade 1.

8.3.2 Item Missing Data Rates for Educational Attainment

Research Question 2

Is the missing data rate different for the Test treatment than for the Control treatment?

The universe is all persons 3 years and over (total population).

The percentage of within universe persons without a valid response to this question in the Test treatment will be compared to the corresponding percentage from the Control treatment for the following populations:

- Total population (across mode)
- Total population (self-response modes, internet and paper combined)
- Total population (CAPI)

To test for these differences, we will conduct two-tailed t-tests at the 0.10 level of significance.

8.3.3 Response Distributions for Educational Attainment

Research Question 3

Is the percentage of persons with educational attainment less than grade 1 for the Control and Test treatments different?

The universe is all persons 3 years and over (total population).

For comparison purposes, we will collapse the "No schooling completed", "Nursery school", and "Kindergarten" response categories in the Control treatment into a single response category, "Less than Grade 1".

U.S. Census Bureau

Attachment C

We will compare the percentage with less than grade 1 educational attainment for these populations:

- Total population
- Children (3-17)
- Adults (18 and older) (overall, self-response, and CAPI)
- Adults who reported that they are foreign born (overall)

To test for these differences, we will conduct two-tailed t-tests at the 0.10 level of significance.

Research Question 4

Are these distributions of educational attainment for the Test and Control treatments different?

The universe is all persons 3 years and over (total population).

(i)   Distribution for the response categories on the questionnaire

The "No schooling completed", "Nursery school", and "Kindergarten" response categories in the Control treatment will be collapsed into a single response category, "Less than Grade 1".

(ii)   Distribution where the response categories are:
    o Less than grade 9
    o Grades 9 through 12 (no diploma)
    o High school graduate (regular high school diploma, GED or alternative credential)
    o Some college credit, no degree
    o Associate's degree
    o Bachelor's degree or higher

We will compare these response distributions for the following populations using the first-order Rao-Scott adjustment of the chi-square test at the 0.1 level of significance. If this test indicates a significant difference between the control and test distributions, we will test for significant differences in the individual response category proportions.

- Total population
- Children (3-17)
- Adults (18 and over) (overall, self-response, CAPI)
- Adults (18 and over) who reported that they are foreign born

8.3.4   Response Reliability for Educational Attainment

<u>Research Question 5</u>

Are the measures of response reliability (Gross Difference Rate, Index of Inconsistency) for the Control and Test treatments different?

The universe is adult (18 and over) respondents.

We will collapse the "No schooling completed", "Nursery school", and "Kindergarten" response categories in the Control treatment into a single response category, "Less than Grade 1".

We will measure response reliability for each treatment by comparing the responses in the original interview to the responses in the CFU reinterview (conducted via telephone). We will compute a GDR and IOI value for each response category in each treatment.

The analysis will be conducted across mode to ensure that we have sufficient sample in the individual response categories. We will test for differences between the Control and Test treatments in the individual response categories using a t-test at the 0.1 level of significance.

8.3.5   Other Metrics for Educational Attainment

None

8.4   DECISION CRITERIA

The most important results to consider in determining if the Test version of the Educational Attainment questions should be adopted are shown below, in order of priority:

Table 10. Decision Criteria for Educational Attainment

Priority	Research Question	Decision Criteria
1	3	The Test version has a lower percentage of persons with educational attainment less than grade 1.
2	4	The distribution of educational attainment across response categories for the Control and Test versions are different.
3	1	The estimate of the percentage of the population with educational attainment less than grade 1 in the Test version is nominally close to the estimate in the CPS.
4	2	The missing data rate for the Test version is the same or lower than the missing data rate for the Control version.
5	5	The measures of response variance (GDR, IOI) for the Test version is the same or lower than that of the Control version in the less than grade 1 response category.

Research questions not included in the decision criteria are for informational purposes only.

8.5   REFERENCES

Crissey, S., Bauman, K., & Peterson, A. (2007). *Evaluation report covering educational attainment*. U.S. Census Bureau. Retrieved November 16, 2021, from https://www.census.gov/content/dam/Census/library/working-papers/2007/acs/2007_Crissey_02.pdf

Lee, J., & Zhou, M. (2015). Immigration hyper-selectivity and second-generation convergence. In *The Asian American achievement paradox* (pp. 21-49). Russel Sage Foundation.

U.S. Census Bureau

9. HEALTH INSURANCE COVERAGE

Authors: Laryssa Mykyta, Sharon Stern, Jonathan Vespa (SEHSD), and Samantha Spiers (DSSD)

The following section describes the specifics of the Health Insurance Coverage question in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the Test versions of the question (Section 9.1).
- A discussion of the difference between the Control and Test versions (Section 9.2).
- The specific research questions and the analysis methodology (Section 9.3 and 9.4).

9.1 LITERATURE REVIEW

The Census Bureau first introduced a question collecting information on a person's health insurance coverage on the ACS in 2008. The purpose of the question is "to enable the Department of Health and Human Services (HHS) and other federal agencies to more accurately distribute resources and better understand state and local health insurance needs" (U.S. Census Bureau, 2007).

The purpose of testing the revised health insurance question is to enhance question reliability and validity. Since implementation in 2008, analyses have revealed some limitations in the current measure. Among these, research has found that Medicaid and other means-tested programs are underreported (O'Hara, 2010; Boudreaux et al., 2011; Lynch et al., 2011; Boudreaux et al., 2013; Boudreaux et al., 2014). Other research has found that direct purchase coverage is overreported, in part due to misreporting of non-comprehensive health plans, such as plans that cover only dental, vision, or prescription drug expenses and do not include hospital or physician coverage (not in scope in the ACS), and reporting multiple coverage types for the same plan (Mach & O'Hara, 2011; Lynch et al., 2011; Boudreaux et al., 2014).

Moreover, revisions to the health insurance coverage question would help capture changes to the health insurance landscape that occurred with and since the passage of the Patient Protection and Affordable Care Act (United States Congress, 2010).

The primary objectives of revising the health insurance coverage question are:
- to improve measurement of public coverage and accuracy of direct purchase coverage,
- to reduce over-count of single-service insurance plans, and
- to reduce erroneous reports of multiple coverage.

This revised question would enhance question reliability and validity.

9.2   QUESTION CONTENT

Control and Test versions of each question are shown as they will appear on the paper questionnaire. Automated versions of the questionnaire will ask each category as an individual question for the Control and Test Version 1 or have respondents choose coverage options from a list (with a show card in CAPI) for Test Version 2.

Figure 25. Control Version of the Health Insurance Coverage Question (Paper)

Is this person **CURRENTLY** covered by any of the following types of health insurance or health coverage plans? *Mark "Yes" or "No" for EACH type of coverage in items a – h.*

	Yes	No
a. Insurance through a current or former employer or union (of this person or another family member)	☐	☐
b. Insurance purchased directly from an insurance company (by this person or another family member)	☐	☐
c. Medicare, for people 65 and older, or people with certain disabilities	☐	☐
d. Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability	☐	☐
e. TRICARE or other military health care	☐	☐
f. VA (enrolled for VA health care)	☐	☐
g. Indian Health Service	☐	☐
h. Any other type of health insurance or health coverage plan – *Specify* ↙	☐	☐

Figure 26. Test Version 1 of the Health Insurance Coverage Question (Paper)

Is this person CURRENTLY covered by any of the following types of health insurance or health coverage plans?

Do NOT include plans that cover only one type of service, such as dental, drug, or vision plans.

Mark "Yes" or "No" for EACH type of coverage in items a – h.

	Yes	No
a. Insurance through a current or former employer, union, or professional association (of this person or another family member)	☐	☐
b. Medicare, for people 65 and older, or people with certain disabilities	☐	☐
c. Medicaid, Children's Health Insurance Program (CHIP), or any kind of government-assistance plan for those with low incomes or a disability	☐	☐
d. Insurance purchased directly from an insurance company, a broker, or a State or Federal Marketplace, such as Healthcare.gov	☐	☐
e. Veteran's health care (enrolled for VA)	☐	☐
f. TRICARE or other military health care	☐	☐
g. Indian Health Service	☐	☐
h. Any other type of health insurance or health coverage plan – *Specify* ⬏	☐	☐

Figure 27. Test Version 2 of the Health Insurance Coverage Question (Paper)

Is this person CURRENTLY covered by any of the following types of health insurance or health coverage plans? *Do NOT include plans that cover only one type of service, such as dental, drug, or vision plans.*

YES, INSURED

Mark (X) for all that apply.

☐ Insurance through a current or former employer, union, or professional association (of this person or another family member)

☐ Medicare, for people 65 and older, or people with certain disabilities

☐ Medicaid, Children's Health Insurance Program (CHIP), or any kind of government-assistance plan for those with low incomes or a disability

☐ Insurance purchased directly from an insurance company, a broker, or a State or Federal Marketplace, such as Healthcare.gov

☐ Veteran's health care (enrolled for VA)

☐ TRICARE or other military health care

☐ Indian Health Service

☐ Any other type of health insurance or health coverage plan – *Specify* ⬏

NO, UNINSURED

☐ No health insurance or health coverage plan

The changes to the Test versions of the health insurance question include the following improvements:

Attachment C

1. **Instruction to exclude single-service coverage plans:**

In order to address the objective of reducing the overcount of single-service coverage plans, Test Versions 1 and 2 of the health insurance question added an instruction to help focus respondents' attention on comprehensive coverage (e.g., coverage for hospital and physician services), and, therefore, to reduce overreporting of direct purchase coverage. The instruction was positioned after the title question but before the health insurance type response choices and reads:

Do NOT include plans that cover only one type of service, such as dental, drug, or vision plans.

2. **Changes to question order of health insurance types:**

The order of the health insurance types in the Control version (which is the same as the current ACS production) is as follows: (a) employer provided, (b) direct purchase, (c) Medicare, (d) Medicaid, (e) TRICARE, (f) Veteran's health care (VA), (g) Indian Health Service, (h) other (write-in).

The order in Test Versions 1 and 2 of the question is as follows: (a) employer provided, (b) Medicare, (c) Medicaid, (d) direct purchase, (e) Veteran's health care (VA), (f) TRICARE, (g) Indian Health Service, (h) other (write-in).

The direct purchase option was moved down the list (from second (b) to fourth (d) position), and Medicare and Medicaid were shifted up one position each. This reordering was designed to reduce over-reporting of direct purchase insurance (Mach & O'Hara, 2011) and improve reporting of public coverage. The reordering was found to reduce direct purchase coverage rates in the 2016 ACS Content Test (Berchick et al., 2017).

Round 1 of the Content Test cognitive testing revealed confusion among interviewees regarding military health care (TRICARE) and Veteran's health care (VA) health care based on interview observations (RTI International, 2021). In order to improve reporting, TRICARE and Veteran's health care (VA) were reordered on the Test versions in later cognitive testing rounds and for the field test.

3. **Rewording of response options**

A. Key terms were added to the direct purchase question in Test Versions 1 and 2 to improve measurement of health coverage in the current health insurance landscape. The description in the Control version (i.e., production) is:

Insurance purchased directly from an insurance company (by this person or another family member)

The description of direct purchase coverage in the Test versions now reads:

Insurance purchased directly from an insurance company, a broker, or a state or federal Marketplace, such as healthcare.gov

    i.   In order to improve measurement of direct purchase coverage obtained through the Health Insurance Marketplace, "..or a state or federal Marketplace, such as healthcare.gov" was added to the direct purchase coverage question. Since the ACA was implemented in 2014, people have been able to buy health insurance through the Health Insurance Marketplace, usually through healthcare.gov or a state-specific website. Although this coverage is direct purchase, individuals may not know how to report it. This change is intended to reduce reporting of Marketplace plans as Medicaid or as both Medicaid and direct purchase. The inclusion of the terms "Marketplace" and "healthcare.gov" may help orient respondents to identify this coverage as direct purchase coverage.[18]

    ii.   In order to improve measurement of direct purchase coverage, the term "a broker" was added to the direct purchase question. Some individuals purchasing coverage directly may purchase their coverage through a broker. Adding the term "a broker" may help orient respondents to identify this coverage as direct purchase coverage.

    iii.   In order to simplify the question text and to make space for additional references to types of direct purchase health insurance, the phrase "(by this person or another family member)" was removed. Removing this phrase highlights the different types and sources of direct purchase coverage and may aid respondents in recognizing their coverage as direct purchase.

B.   Reference to the Children's Health Insurance Program (CHIP) was added to the Medicaid question in Test Versions 1 and 2 to improve reporting of Medicaid and other means-tested coverage. This addition will aid respondents in recognizing CHIP as a means-tested insurance program, selecting Medicaid coverage and therefore, potentially reducing the undercount of Medicaid coverage. The term "Medical Assistance" was removed and reference to CHIP was added so that the description of Medicaid coverage now reads:

Medicaid, Children's Health Insurance Program (CHIP), or any kind of government-assistance plan for those with low incomes or a disability

C.   The phrase "or professional association" was added to the employer provided insurance coverage question in Test Versions 1 and 2 to improve reporting of coverage under group plans

---

[18] Many state portals that provide access to Marketplace coverage also provide access to Medicaid coverage if individuals meet income eligibility and other criteria for these programs. Therefore, it is possible that some respondents may misreport Medicaid coverage accessed through a state portal as Marketplace coverage.

(e.g., plans sponsored by trade associations or professional groups). The description of employer provided coverage now reads:

Insurance through a current or former employer, union, or professional association (of this person or another family member)

D.   "Veterans' health care" was substituted for "VA" in Test Versions 1 and 2 of the health insurance question. This change should improve reporting of Veteran's health care through the Veterans' Administration and mitigate respondent confusion of Veteran's health care and other forms of military coverage, such as TRICARE. The description of Veteran's health care now reads:

Veteran's health care (enrolled for VA)

4.   **Test Version 2 includes a "No, not insured" option as part of the health insurance question.**

In the Round 1 briefing report from cognitive testing for the 2022 ACS Content Test, RTI recommended including a "No coverage, Uninsured" option, and Round 2 Version 2 of the health insurance question tested this option. The Round 1 briefing report suggested that respondents would prefer a response category that allows them to specifically report no coverage (RTI International, 2021). When this option was tested in Version 2 of the health insurance question in Round 2 of cognitive testing, however, the results were inconclusive given the small number of interviews. The Test Version 2 of the health insurance question incorporates a "No, not insured" option as part of the health insurance question. This change should improve reporting of coverage by enabling respondents to actively select no coverage as an option and mitigate against erroneous reporting of no coverage by including the option as the last item on the insurance type list.

5.   **Test Version 2 is a check-all-that-apply format**

Due to adding the "No coverage, Uninsured" option, Test Version 2 has a different layout than Test Version 1 and the Control version. The question layout is a check-all-that-apply or check "no" format instead of individual yes/no questions for each coverage type (letters a-h). This change is expected to be most impactful in the self-response modes, paper and internet. In CAPI, we include a show card with the question, which allows the respondent to select from the list of health insurance options. In the current production ACS and in the Content Test Control and Test Version 1, field representative's read each item separately and ask for a yes/no response (and therefore, no show card is needed).

9.3 RESEARCH QUESTIONS AND METHODOLOGY: TEST VERSION 1 VS. CONTROL

The following section describes the specific analysis to be conducted on the Health Insurance coverage questions in the Test treatment (Test Version 1) of the 2022 Content Test.[19] Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 9.3.1).
- Specifics of the item missing data rate analysis (Section 9.3.2).
- Specifics of the response distribution analysis (Section 9.3.3).
- Specifics of the response reliability analysis (Section 9.3.4).
- Other analysis planned for this topic (Section 9.3.5).

Given the changes to question order in the treatments as described in Section 9.2, all comparisons will be based on coverage type and not on question order. Table 11 shows the comparisons that will be made for Test Version 1 (Test treatment) and the Control version (Control treatment) for particular types of coverage.

Table 11. Comparisons between Test Control and Version 1 Based on Question Order for Health Insurance Coverage

Health Coverage Type	Test Version 1	Control
Employer provided	a	A
Medicare	b	C
Medicaid	c	D
Direct purchase	d	B
VA	e	F
TRICARE	f	E
Indian Health Service	g	G
Other – fill	h	H

Research questions looking at "any health insurance" or "no health insurance" are complicated by the following two issues.

- People who report only Indian Health Service and no other type of health coverage are considered uninsured.
- "Other" is not a type of health coverage. Responses could reflect:
  - comprehensive health insurance coverage that is assigned to one of the types of coverage defined in questions (a) through (f);
  - a non-comprehensive health plan or a single-service plan or other type of insurance (e.g., long term care) reported erroneously (assigned to uninsured); or
  - no coverage and be assigned to uninsured.

---

[19] Research questions describing analyses to be conducted on the health insurance coverage questions in the Roster Test treatment (Test Version 2) are specified in Section 9.4.

Therefore, we specify how the categories "any health insurance" or "no health insurance" are defined and how comparisons are constructed in the research questions below.

9.3.1 Benchmarks for Health Insurance Coverage

Research Question 1

How do the proportions of persons with any health insurance coverage in the Test treatment and the Control treatment compare to the proportions found in the most recent Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and the National Health Interview Survey (NHIS)?

Research Question 2

How do the proportions of persons with Medicaid coverage in the Test treatment and the Control treatment compare to the proportions found in the most recent Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and the National Health Interview Survey (NHIS)? How do the proportions of persons with direct purchase coverage compare?

The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) is the most widely used source of national health insurance estimates released by the Census Bureau, and the National Health Interview Survey (NHIS) is the principal source of information on the health of the U.S. population released by the National Center for Health Statistics (NCHS). Estimates of health insurance coverage from these surveys may be different than those from the ACS for a variety of reasons. First, while the ACS and the NHIS measure health insurance coverage at the time of the interview, the CPS is collected between February and April of each year and captures coverage in the previous calendar year.[20] Second, different survey modes may lead to different results, and the ACS is the only survey among the three that includes a paper mode and internet mode. Third, the context of the survey may differ, which could prime respondents in different ways when they answer survey questions. For example, the NHIS focuses on health, while the CPS ASEC focuses on income. Fourth, different surveys have different editing procedures that may lead to different estimates. Additionally, there are other differences (such as question wording) that could create differences in estimates of health insurance coverage.

Historically, the uninsured estimate in the ACS has been higher than that in the CPS ASEC but similar to the uninsured estimate in the NHIS. However, estimates of private coverage in the NHIS have been lower than that in the ACS, and estimates of public coverage in the NHIS have

---

[20] The CPS ASEC also includes a measure of coverage reported at the time of interview. However, the CPS ASEC is collected between February and April of each year and the current coverage measure thus reflects health coverage early in the survey year. In contrast, the ACS is collected throughout the year and provides a measure of annual coverage.

been slightly higher than that in the ACS. These trends may change with the 2019 NHIS redesign.

We will make general comparisons to the most recent CPS ASEC and NHIS data that are available. We will compare the proportion of people with any health coverage and with coverage by Medicaid and by direct purchase in the Control treatment and the Test treatment with benchmark estimates. A person will be defined as having any health coverage if they select "Yes" to having employment-based coverage, Medicare, Medicaid, direct purchase coverage, TRICARE, coverage through the VA, or if they select Other and specify an in-scope coverage or a plan.

9.3.2   Item Missing Data Rates for Health Insurance Coverage

Given the structure of the health insurance questions on the Control version and Test Version 1, there are three classes of responses that are of interest when analyzing the item missing data rate.

1. Overall missing: Person records with no response to any part of the question (completely blank)
2. Partial missing (or partial response): Person records with at least one item with a "yes" or "no" box marked and at least one item with neither box selected
3. Complete response: Person records with either "yes" or "no" marked for each coverage type.

Research Question 3

Is the overall item missing data rate different between the Test treatment and the Control treatment?

The overall item missing data rate is defined as the proportion of eligible people who fail to provide any type of response to items a through h (completely blank) on paper or internet questionnaires or respond with "Don't Know" or "Refused" in CAPI. We will compare the proportion of overall item missingness between treatments. We will exclude persons that have all detailed person questions missing or persons with early breakoffs (i.e., stopped answering the questionnaire at any of the questions before the health insurance questions).

Research Question 4

Is the proportion of partial responses different between the Test treatment and the Control treatment?

We will also compare the proportion of people with partial responses (those with at least one item with a "yes" or "no" box marked and at least one item with neither box selected for self-

response modes or at least one item marked don't know or refusal for CAPI) between the Control and Test treatments of the health insurance question. These partial responses create the need for assumptions in editing. More complete data is preferred.

Research Question 5

Are the overall item missing data rate and proportion of partial responses different between the Test treatment and the Control treatment when dividing responses by mode (paper, internet, CAPI)?

Self-administered questionnaires (notably paper responses) have been shown to have higher rates of nonresponse than other modes (Clark, 2014). We will compare the item missing data rate and proportion of partial responses in the above research questions between treatments while keeping the response mode constant.

Research Question 6

Are the item missing data rates for the Medicare, Medicaid, and direct purchase items different between the Test treatment and the Control treatment?

Research Question 7

Are the item missing data rates for the specific combinations of response different between the Test treatment and the Control treatment? Specifically examining Medicare-Medicaid partial response and Medicaid-Direct Purchase partial response.

One motivation for changing the health insurance coverage question is to improve the accuracy of reported Medicaid and direct purchase coverage. We are also concerned about the interaction between the Medicare and Medicaid boxes.

In Research Question 6, we will examine the rate of missing responses for three items: Medicaid, direct purchase, and Medicare boxes. We will examine whether a respondent marked either checkbox for the item of interest and compare the item missing data rates by health insurance type between the Control and Test treatments.

In Research Question 7, we will examine all missing and partial missing for the combined sets: Medicaid combined with direct purchase and Medicaid combined with Medicare. A partial response is one in which only one box (yes or no) is checked among the four possible (two response options (yes or no) for each item).

Research Question 8

Are the item missing data rates for the TRICARE and VA boxes different between the Test treatment and the Control treatment?

Because the Test treatment has the order of TRICARE and VA switched, we want to examine if there is a difference in item missing data rates between the treatments for these boxes. Similar to the previous research questions, we will examine whether a respondent marked a checkbox, treating an implied "no" as missing data.

For all research questions, statistical significance between versions will be determined using a two-tailed t-test.

9.3.3   Response Distributions for Health Insurance Coverage

Research Question 9

Are rates of having any health insurance coverage different between the Test treatment and the Control treatment?

We will compare the rates of any health insurance coverage between treatments. Health insurance coverage excludes people with only Indian Health Service. Additionally, if someone only marks "yes" to the write-in box and their entry is coded as "out-of-scope", then they do not have coverage.

The table below explains the distributions being compared in Research Question 9.

Table 12. Definitions of Any Coverage and No Coverage

With or Without Coverage	Test and Control
Any Coverage (This definition may include people with partial missing)	Selecting "yes" to one or more of the items a-f or writing-in an in-scope coverage or plan a = yes OR b = yes OR c = yes OR d = yes OR e = yes OR f = yes OR (h = yes AND write-in is in-scope)
No Coverage (This definition excludes people with any missing responses to a-f)	Selecting "no" to all items a-h, selecting "no" to all items a-f and writing-in an out-of-scope coverage plan, or selecting "yes" to items g and h and writing-in an out-of-scope coverage or plan (a = no AND b = no AND c = no AND d = no AND e = no AND f = no AND (g = yes OR no) AND h = no) OR (a = no AND b = no AND c = no AND d = no AND e = no AND f = no AND h = yes AND write-in is out-of-scope) OR (g = yes AND a-f = no AND h = yes AND write-in is out-of-scope)

Attachment C

Research Question 10

Are rates of coverage by employer provided insurance, Medicare, Medicaid, direct purchase insurance, VA, and TRICARE different between the Test treatment and the Control treatment?

We will compare the rates of specific types of health insurance coverage between treatments. We want to see if the rates of Medicaid coverage and direct purchase coverage are different between the Test treatment and the Control treatment from reordering the answer choices and clarifying the language in the Test treatment. We also want to see if the rates of VA and TRICARE coverage are different between treatments from reordering the answer choices in the Test treatment question. We will compare the rates of employer provided coverage between treatments to see if the rates are affected due to changes in the language of the Test treatment question. Table 1 provide the items being compared. For example, item b on Test Version 1 will be compared with item c on the Control version, because both response choices are Medicare.

Research Question 11

Are the proportions of persons with multiple types of health insurance coverage different between the Test treatment and the Control treatment?

We will compare the proportions of persons with multiple types of health insurance coverage between treatments. A person will be defined as having multiple coverage if they select "yes" to more than one type of health coverage except for Indian Health Service (i.e., if they select any combination of employment-based coverage, Medicare, Medicaid, direct purchase coverage, TRICARE, coverage through the VA, or other and specify an in-scope coverage type or plan).

Research Question 12

Are the proportions of persons who reported having both Medicaid and direct purchase insurance different between the Test treatment and the Control treatment?

One motivation for changing the health insurance coverage question is to improve the accuracy of reported Medicaid and direct purchase coverage. We will compare the proportion of persons who report having both Medicaid and direct purchase insurance between treatments. A person will be defined as having both Medicaid and direct purchase insurance if they selected "yes" to both the Medicaid and the direct purchase insurance options (options c and d in the Test treatment). Having Medicaid and direct purchase coverage at the same time is an unlikely combination.

Attachment C

Research Question 13

Are the proportions of persons who reported having both Medicare and direct purchase insurance different between the Test treatment and the Control treatment? Are the proportions of persons who reported having both Medicare and Medicaid insurance different between the Test treatment and the Control treatment?

Persons over 65 are nearly universally eligible for Medicare coverage, yet many persons may supplement their Medicare coverage with a direct purchase plan or may purchase a Medicare Advantage plan. We will compare the proportion of persons who report having both Medicare and direct purchase insurance between treatments. A person will be defined as having both Medicare and direct purchase insurance if they selected "yes" to both the Medicare and the direct purchase insurance options (options b and d in the Test treatment).

We are also concerned about the interaction between the Medicare and Medicaid boxes. Cognitive testing revealed some individuals with dual coverage reported only Medicare or only Medicaid. We will compare the proportions of persons who report having both Medicare and Medicaid between treatments. A person will be defined as having both Medicare and Medicaid if they selected "yes" to both the Medicare and Medicaid options (options b and c in the Test treatment).

Research Question 14

Are the proportions of persons who reported having both employer provided and direct purchase insurance different between the Test treatment and the Control treatment?

We want to see if the proportions of persons who report having both employer provided and direct purchase insurance are different between treatments. The Test treatment includes several question changes that may reduce confusion between the two insurance options, such as moving the direct purchase response option further down the list and including an instruction to exclude single-service plans. A person will be defined as having both employer provided and direct purchase insurance if they selected "yes" to both the employer provided and the direct purchase insurance options (options a and d in the Test treatment).

Research Question 15

Are the proportions of persons who write-in "other" type of health insurance coverage different between the Test treatment and the Control treatment?

The content of the write-in box will not be assessed as part of this question; only the presence of a reply in the write-in box will be assessed.

Research Question 16

Do the proportion of write-ins that reference "CHIP" or a state name for the CHIP program, "Marketplace", "healthcare.gov", "ACA" or other terms associated with the ACA such as "Obamacare" differ between the Test treatment and the Control treatment?

As described in Section 9.2, additional text was added to several response options to better define the health insurance types for respondents. For example, "Children's Health Insurance Program (CHIP)" was added to the Medicaid option and "through a state or federal Marketplace, such as healthcare.gov" was added to the direct purchase option. The addition of this text should reduce the number of write-ins that reference these terms in Test Version 1.

Research Question 17

Do the proportion of write-ins that are out-of-scope differ between the Test treatment and the Control treatment?

As described in Section 9.2, an instruction was added to exclude single-service plans from reporting of health insurance coverage in response to the health insurance question. Many write-ins coded as out-of-scope specify single-service plans, such as prescription plans, vision, or dental plans, accidental coverage, or coverage for specific conditions. The addition of the instruction should reduce the number of write-ins that are out-of-scope.

Research Question 18

Are the above rates and proportions described in Q9 through Q15 different between the Test treatment and the Control treatment when dividing responses by mode (paper, internet, CAPI)?

Research Question 19

Are the above rates and proportions described in Q9 through Q15 different between the Test treatment and the Control treatment when dividing responses by age (under 19, 19-64, 65+) or by state group (i.e., whether or not the state expanded Medicaid eligibility)?

The set of tests would be unnecessarily high if we generate all possible comparisons by age and collection mode in Research Questions 18 and 19. Table 13 presents the subset of comparisons that we need to address the purpose of the test. It shows which of the Research Questions 9-15 need to be repeated by collection mode, age group, and Medicaid-expansion state status.

Table 13. Comparisons by Mode, Age Group, and Medicaid-Expansion State Status

Research Question	Mode	Age Group*	Medicaid-Expansion State Status
Research Question 9: Any health insurance coverage	Yes	Age Groups 1, 2, 3	Yes
Research Question 10 Individual types of coverage			
Medicaid	Yes	Age Groups 1, 2, 3	Yes
Direct Purchase	Yes	Age Groups 1, 2, 3	Yes
Employer Provided	Yes	Age Groups 2, 3	No
Medicare	Yes	Age Group 3	No
Research Question 11 Multiple coverage – any combination	Yes	No	No
Research Question 12 Multiple coverage - Medicaid and Direct Purchase	Yes	No	Yes
Research Questions 13 Multiple coverage – Medicare and…	No	Yes	No
..Direct Purchase		Age Group 3	
..Medicaid insurance		Age Groups 2, 3	
Research Question 14 Multiple coverage – Employer Provided and Direct Purchase	No	Age Groups 2, 3	No
Research Question 15 *With a write-in "other" type of health insurance coverage*	Yes	No	Yes

* Age groups are as follows: Group 1 = People under age 19, Group 2 = People 19 to 64 years, and
  Group 3 = People age 65 and over

We will compare the rates of any coverage and specific types of health insurance coverage between treatments by mode because health insurance coverage historically varies by response mode. We will compare rates by age group because certain types of health insurance are specific to age groups (e.g., Medicare for people 65 years old or older). We will also compare rates by Medicaid-expansion state status between treatments because knowledge of health coverage types may differ across states due to differences in outreach related to public program eligibility.

For all research questions, statistical significance between versions will be determined using a two-tailed t-test.

9.3.4   Response Reliability for Health Insurance Coverage

We will measure response reliability, specifically simple response variance, through re-asking the health insurance coverage question versions on the CFU reinterview.

U.S. Census Bureau

Research Question 20

Are the measures of response reliability (GDR, IOI) different between the Test treatment and the Control treatment, for individual coverage types (items a to f) or no coverage?

Research Question 21

Are the measures of response reliability (GDR, IOI) different between the Test treatment and the Control treatment within each collection mode (paper, internet, CAPI)?

We will compare measures of response reliability between treatments. We will compare individual items as presented in Table 11. We will create GDR and IOI for "no coverage" and "any coverage" between the original interview and reinterview as described in Table 12. Response reliability for items g and h are covered by the "no coverage" and "any coverage" comparison. We will make the same comparisons of GDR and IOI between treatments within each of the three data collection types: paper, internet, and CAPI.

Statistical significance between the GDRs and IOIs of each version will be determined using a two-tailed t-test.

9.3.5   Other Metrics for Health Insurance Coverage

Research Question 22

Is there a difference in help text use on the internet and CAPI instruments between the Test treatment and the Control treatment?

Research Question 23

Is there a difference in behavior on the internet instrument between the Test treatment and the Control treatment with regards to:

a.   Switching between answer choices while on the same screen?
b.   Going back to previous screens?
c.   Time spent on the question screen?

Analysis for the last two research questions will be partially covered by a separate planned respondent burden analysis using internet instrument paradata for all topics in the 2022 Content Test (see Section 3.2.1).

For all research questions, statistical significance between versions will be determined using a two-tailed t-test.

Research Question 24

How does the number of persons with Medicaid coverage in each treatment compare with Medicaid enrollment based on administrative records from the Center for Medicaid and Medicare Services?

We may also compare estimates from both Test Version 1 and the Control version with Medicaid enrollment data from administrative data released through the Center for Medicaid and Medicare Services. Although ACS estimates of the number of people covered by Medicaid and CHIP tend to be lower than enrollment counts from CMS, we will compare the number of people with Medicaid coverage in the Control and Test treatments with enrollment data from CMS.

9.4 RESEARCH QUESTIONS AND METHODOLOGY: TEST VERSION 1 VS. VERSION 2

The following section describes the specific analysis to be conducted between Test Version 1 (Test treatment) and Test Version 2 (Roster Test treatment) of the health insurance coverage question in the 2022 Content Test. Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 9.4.1).
- Specifics of the item missing data rate analysis (Section 9.4.2).
- Specifics of the response distribution analysis (Section 9.4.3).
- Specifics of the response reliability analysis (Section 9.4.4).
- Other analysis planned for this topic (Section 9.4.5).

Most of the analysis for the following research questions have been described in the research questions for the Control treatment versus the Test treatment (Section 9.3). Instead of repeating the information, we refer to the applicable parts of that section when needed.

9.4.1 Benchmarks for Test Version 1 vs. Test Version 2

Research Question RT1

How do the proportions of persons with any health insurance coverage in each treatment compare to the proportions found in the most recent Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and the National Health Interview Survey (NHIS)?

Research Question RT2

How do the proportions of persons with health insurance coverage by Medicaid in each treatment compare to the proportions found in the most recent Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and the National Health Interview Survey (NHIS)? How do the proportions by direct purchase compare?

We will make general comparisons to the most recent CPS ASEC and NHIS data that are available. As in Section 9.3.1, we will compare the proportion of people with any health coverage and with coverage by Medicaid and by direct purchase in Test Version 1 and Test Version 2 with benchmark estimates. A person will be defined as having any health coverage if they select "yes" to (in Test Version 1) or select (in Test Version 2) employment-based coverage, Medicare, Medicaid, direct purchase coverage, TRICARE, coverage through the VA, or if they select other and specify an in-scope health coverage type or health insurance plan.

9.4.2   Item Missing Data Rates for Test Version 1 vs. Test Version 2

In Section 9.3.2, we define missing data for Test Version 1 as follows:

4. Overall missing: Person records with no response to any part of the question (completely blank)
5. Partial missing (or partial response): Person records with at least one item with a "yes" or "no" box marked and at least one item with neither box selected
6. Complete response: Person records with a marked either "yes" or "no" for each coverage type.

Research Question RT3

Is the overall item missing data rate in Test Version 2 different from the rate of overall missing or partial missing in Test Version 1?

Research Question RT4

Is the overall item missing data rate in Test Version 2 different from the rate of overall missing or partial missing in Test Version 1 when dividing responses by mode (paper, internet, CAPI)?

The overall item missing data rate is the proportion of eligible people who fail to provide any type of response (completely blank) on paper or internet questionnaires or respond with "Don't Know" or "Refused" in CAPI. Since Test Version 2 is a check-all-that-apply format and not individual yes/no questions, the overall item missing data rate for this version also includes the partial missingness defined for Test Version 1. Therefore, to have an equivalent comparison, we will compare the overall item missing data rates in Test Version 2 to the combined proportion or overall missingness and partial missingness in Test Version 1.

As in Section 9.3.2, we will exclude persons that have all detailed person questions missing or persons with early breakoffs (i.e., stopped answering the questionnaire at any of the questions before the health insurance questions). Self-administered questionnaires (notably paper responses) have been shown to have higher rates of nonresponse than other modes (Clark 2014). We will compare the item missing data rates in the above research question between treatments while keeping the response mode constant for Research Question RT4.

For all research questions, statistical significance between versions will be determined using a two-tailed t-test.

9.4.3   Response Distributions for Test Version 1 vs. Test Version 2

Table 12 in Section 9.3.3 describes how we measure "any coverage" or "no coverage" in the Test and Control treatments. It defined any coverage as selecting "yes" to one or more of the items a-f or writing-in an in-scope coverage or plan in item h. This definition also applies to Test Version 2: checking a check box to one or more of the first six items or writing-in an in-scope coverage or plan in the other field.

Table 12 also defined no coverage for Test Version 1 as selecting "no" to all items a-f, which excludes people with missing responses to all items. Since Test Version 2 has a no health insurance option, there could be more people selecting the option instead of leaving the question blank entirely. To better examine this possibility, we will compare two different ways of constructing "no coverage" in Test Version 1, as shown in Table 14.

Table 14. Definitions of No Coverage for Test Version 1

No Coverage Definition 1	No Coverage Definition 2
Selecting "no" to all items a-h, selecting "no" to all items a-f and writing-in an out-of-scope coverage plan, or selecting "yes" to items g and h and writing-in an out-of-scope coverage or plan	Not selecting "yes" to any items a-h, not selecting "yes" to any items a-f and writing-in an out-of-scope coverage plan, or selecting "yes" to items g and h and writing-in an out-of-scope coverage or plan
(a = no AND b = no AND c = no AND d = no AND e = no AND f = no AND (g = yes OR no) AND h = no) OR (a = no AND b = no AND c = no AND d = no AND e = no AND f = no AND h = yes AND write-in is out-of-scope) OR (g = yes AND a-f = no AND h = yes AND write-in is out-of-scope)	(a ≠ yes AND b ≠ yes AND c ≠ yes AND d ≠ yes AND e ≠ yes AND f ≠ yes AND (g = yes OR no OR blank) AND h ≠ yes) OR (a ≠ yes AND b ≠ yes AND c ≠ yes AND d ≠ yes AND e ≠ yes AND f ≠ yes AND h = yes AND write-in is out-of-scope) OR (g = yes AND a-f ≠ yes AND h = yes AND write-in is out-of-scope)

Research Question RT5

Are rates of having any health insurance coverage different between Test Version 1 and Test Version 2?

We will compare the rates of any health insurance coverage between Test Version 1 and Test Version 2. Health insurance coverage excludes people with only Indian Health Service. Additionally, if someone only marks "yes" to the write-in box and their entry is coded as "out-of-scope" or "no coverage", then they do not have coverage.

Research Question RT6

Are rates of coverage by employer provided insurance, Medicare, Medicaid, direct purchase insurance, VA, and TRICARE different between Test Version 1 and Test Version 2?

We will compare the rates of specific types of health insurance coverage between Test Version 1 and Test Version 2. Table 11 in Section 9.3 shows the order of the individual types on Test Version 1. The order is the same on Test Version 2. A "yes" box checked in Test Version 1 will be compared to selecting the check box next to the analogous item. We want to see if the rates are affected by changes in the question format.

Research Question RT7

Are the proportions of persons with multiple types of health insurance coverage different between Test Version 1 and Test Version 2?

We will compare the proportions of persons with multiple types of health insurance coverage between treatments. A person will be defined as having multiple coverage if they select "yes" to (in Test Version 1) or select (in Test Version 2) more than one type of health coverage except for Indian Health Service (i.e., if they select any combination of employment-based coverage, Medicare, Medicaid, direct purchase coverage, TRICARE, coverage through the VA, or other and specify coverage or a plan).

Research Question RT8

Are the proportions of persons who write-in an "other" type of health insurance coverage different between Test Version 1 and Test Version 2?

The content of the write-in box will not be assessed as part of this question; only the presence of a reply in the write-in box will be assessed.

Research Question RT9

Is the proportion of persons with no health insurance coverage in Test Version 1 different from the proportion of uninsured persons in Test Version 2?

We will compare the proportions of persons with no health insurance coverage in Test Version 1 with the proportion of uninsured persons in Test Version 2. In Test Version 2, a person is uninsured if they select "NO, UNINSURED: No health insurance or health coverage plan". We will examine two different definitions of no coverage for Test Version 1 (see Table 14). For both versions, a person will also be considered as having no coverage if they select "yes" to or check other and their write-in entry is coded "out-of-scope" or "no coverage". We will exclude cases

that have all detailed person questions missing or cases with early breakoffs (i.e., stopped answering the questionnaire at any of the questions before the health insurance questions).

Research Question RT10

Are the above rates and proportions different between Test Version 1 and Test Version 2 when dividing responses by mode (paper, internet, CAPI)?

Research Question RT11

Are the above rates and proportions different between Test Version 1 and Test Version 2 when dividing responses by age (under 19, 19-64, 65+)?

We will compare the rates of any coverage and specific types of health insurance coverage between treatments by mode because health insurance coverage historically varies by response mode. We will also compare rates by age group because certain types of health insurance are specific to age groups (e.g., Medicare for people 65 years old or older). See Table 13 in Section 9.3.3 for the specific comparisons for Research Questions RT6 and RT8.

For all research questions, statistical significance between versions will be determined using a two-tailed t-test.

9.4.4   Response Reliability for Test Version 1 vs. Test Version 2

Research Question RT12

Are the measures of response reliability (GDR, IOI) different between Test Version 1 and Test Version 2, overall and when dividing responses by mode (paper, internet, CAPI)?

We are concerned that the measurement of response reliability, specifically simple response variance, between Test Version 1 and Test Version 2 will be impacted by the telephone interview mode of the CFU reinterview. For Test Version 2, the CAPI mode includes a show card which respondents will not be able to access during the CFU telephone reinterview.

We will compare measures of response reliability between versions both overall and when dividing responses by mode. We will conduct an initial analysis of each individual health insurance types as well as uninsured, followed by a secondary analysis involving combinations of responses which will account for the possibility of multiple health insurances.

Statistical significance between the GDRs and IOIs of each version will be determined using a two-tailed t-test.

However, due to the concerns noted above, it is not clear whether we will be able to draw any conclusions regarding the comparative response reliability of the versions.

9.4.5 Other Metrics for Test Version 1 vs. Test Version 2

Research Question RT13

Is there a difference in help text use on the internet and CAPI instruments between Test Version 1 and Test Version 2?

Research Question RT14

Is there a difference in behavior on the internet instrument between Test Version 1 and Test Version 2 with regards to:

    a.   Switching between answer choices while on the same screen?
    b.   Going back to previous screens?
    c.   Time spent on the question screen?

Analysis for the last two research questions will be partially covered by a separate planned respondent burden analysis using internet instrument paradata for all topics in the 2022 Content Test (see Section 3.2.1).

For all research questions, statistical significance between versions will be determined using a two-tailed t-test.

9.5 DECISION CRITERIA

As stated in previous sections, the primary objectives of revising the health insurance questions are to improve measurement of public coverage and accuracy of direct purchase coverage, to reduce overcount of single-service insurance plans, and to reduce erroneous reports of multiple coverage. These misreports may arise in large part through respondent confusion in distinguishing coverage type or not recognizing non-comprehensive coverage as such. Therefore, the decision criteria as specified below adjudicates which version best reduces respondent confusion and improves respondent reports of coverage type.

9.5.1 Decision Criteria for Test Version 1 vs. Control

Preferences for higher or lower rates in Test Version 1 compared with the Control version derive from the literature review section (see Section 9.1). These are included in the decision criteria only when the quality measures identified as priority 1, 2, 3 and 4 decision criteria do not lead to a conclusive decision.

Research Questions 1, 2, 13, 14, 22, 23 and 24 are only for information and will not be a factor in the decision.

The most important results of this analysis when drawing a conclusion about Test Version 1 to the Control version are, in order of priority:

Table 15. Decision Criteria for Health Insurance Coverage: Test Version 1 vs. Control

Priority	Research Question*	Decision Criteria
1	15, 16, 17	We expect the proportion of write-ins to be lower when the respondent understands the question and/or how to categorize their insurance. In addition, coding write-in responses is time-consuming and costly. The test version incorporates two changes which may be expected to reduce the volume of write-ins: (1) the inclusion of the instruction to exclude single-service plans; and (2) the references to "CHIP", "Marketplace" and "healthcare.gov" in the health insurance question text. Therefore, A lower proportion of people who write-in a type of health insurance coverage is preferable. A lower proportion of write-ins that reference the terms that the new wording attempts to clarify is preferable. A lower proportion of write-ins that are determined to be out-of-scope is preferable.
2	11	A difference in proportions of multiple coverage could indicate a reduction in possible respondent confusion on how to categorize their coverage. Therefore, a lower rate of multiple coverage is preferable.
3	3-8	A decrease in overall missingness is preferable. Specifically, fewer responses with all parts left blank is preferable. In general, lower item missing data rates are preferred.
4	20, 21	Higher test-retest reliability is preferable. A lower GDR and IOI are preferable.
5	9	External evaluations suggest the ACS current coverage rate is still too low. If Test Version 1 and the Control version do not differ on other quality measures, the version which has a lower uninsured rate is preferable.
	10	External evaluations suggest the ACS public coverage rate is too low. If Test Version 1 and the Control version do not differ on other quality measures, then the version which has a higher rate of Medicaid coverage is preferable. It may be unclear if a change in estimates means an improvement in reporting (i.e., reduction in bias), and so these changes alone are not sufficient and must be considered along with the other priorities.
	12	Having <u>Medicaid and direct purchase coverage</u> concurrently is highly unlikely. Thus, the version which has fewer cases where both are selected is preferable. In some states, Medicaid and direct purchase Marketplace plans may be accessed through the same online portal. Therefore, some respondents may be answering correctly from their perspective – it is one plan that is Medicaid and purchased through a state

Priority	Research Question*	Decision Criteria
		website. This is a situation which can be managed in the editing process. Hence, while we generally prefer fewer reports of both Medicaid and direct purchase, when analyzed in the fuller context, we may accept higher reports of both types.
	18, 19	Preference for higher or lower rates by mode, age, and Medicaid-expansion state status will depend on the coverage type. See the decision criteria for Research Questions 9-12 for more information.

*Research questions within a priority are organized with smallest research question number listed first.
 Research questions not included in the decision criteria are for research purposes only.

9.5.2   Decision Criteria for Test Version 1 vs. Test Version 2

As described in Section 9.2, Test Version 2 was introduced in Round 2 of cognitive testing for the 2022 ACS Content Test to address some of the feedback received from Round 1. Specifically, some respondents wanted a response category that allowed them to report no coverage.

We will consider the decision criteria below if Test Version 1 is shown to be an improvement over the Control version based on the decision criteria in Table 15. Hence, the proposed decision criteria listed below will be used primarily to determine if Test Version 2 resolves issues identified with the current question format without introducing additional error.

Research Questions RT1, RT2, RT13, and RT14 are only for information and will not be a factor in the decision.

The most important results of this analysis when drawing a conclusion about Test Version 1 to Test Version 2 are, in order of priority:

Table 16. Decsion Criteria for Health Insurance Coverage: Test Version 1 vs. Test Version 2

Priority	Research Question*	Decision Criteria
1	RT3, RT4	Filling missing data with logical edits or allocations requires making assumptions and adding uncertainty that are not measured in the final statistics. In general, lower item missing data rates are preferred. However, the check-all-that-apply question format is almost guaranteed to have less missing data because any mark is considered a complete response. In order to determine that Test Version 2 is an improvement over Test Version 1, we compare missingness within mode. For all modes but especially self-response modes, a decrease in overall missingness is preferable.
2	RT12	Ask/Re-ask reliability measures consistency of response. Generally, if people are able to answer the same question with the same answer several weeks apart, this is evidence of a well-understood question. Both Test Version 1 and Test Version 2 use the same wording. If the check-all-that-apply approach in Test Version 2 is successful, the GDR and IOI should be lower than Test Version 1. However, due to the potential mode impact on Test Version 2, we may not be able to draw any conclusions regarding the comparative response reliability of the versions.
3	RT5, RT9	External evaluations suggest the ACS current coverage rate is still too low. Test Version 2 provides respondents the option to select "NO, UNINSURED: No health insurance coverage or plan". The version which has a lower uninsured rate is preferable.
	RT6	External evaluations suggest the ACS public coverage rate is too low. The version which has a higher rate of Medicaid coverage is preferable.
	RT7	A difference in proportions of multiple coverage could indicate a reduction in possible respondent confusion on how to categorize their coverage. The version which has a lower rate of multiple coverage is preferable.
	RT8	We expect the proportion of write-ins to be lower when the respondent understands the question and/or how to categorize their insurance. In addition, coding write-in responses is time-consuming and costly. Therefore, the version which has a lower proportion of people who write-in a type of health insurance coverage is preferable.
	RT10, RT11	Preference for higher or lower rates by mode and age will depend on the coverage type.

*Research questions within a priority are organized with smallest research question number listed first.

Research questions not included in the decision criteria are for research purposes only.

9.6   REFERENCES

Berchick, E., O'Hara, B., Heimel, S., & Chase Sawyer, R. (2017). *2016 American Community Survey evaluation report: Health insurance coverage, premiums and subsidies*. U.S. Census Bureau. Retrieved February 23, 2022, from https://www.census.gov/library/working-papers/2017/acs/2017_Berchick_01.html

Boudreaux, M., Call, K., Turner, J., & Fried, B. (2014). Estimates of direct purchase from the ACS and Medicaid misreporting: Is there a link? State Health Access Data Assistance Center. Retrieved February 23, 2022, from https://www.shadac.org/sites/default/files/publications/SHADACBrief38_DirectPurchase_Web.pdf

Boudreaux, M. H., Call, K. T., Turner, J., Fried, B., & O'Hara, B. (2015). Measurement error in public health insurance reporting in the American Community Survey: Evidence from record linkage. *Health Services Research*, 50(6), 1973-1995. https://doi.org/10.1111/1475-6773.12308

Boudreaux, M., Ziegenfuss, J. Y., Graven, P., Davern, M., & Blewett, L. A. (2011). Counting uninsurance and means-tested coverage in the American Community Survey: A comparison to the Current Population Survey. *Health Services Research*, 46(1), 210-231. https://doi.org/10.1111/j.1475-6773.2010.01193.x

Lynch, V., Kenney, G. M., Haley, J., & Resnick, D. M. (2011). *Improving the validity of the Medicaid/CHIP estimates on the American Community Survey: The role of logical coverage edits*. U.S. Census Bureau. Retrieved February 24, 2022, from https://www.census.gov/library/working-papers/2011/demo/improving-the-validity-of-the-medicaid-chip-estimates-on-the-acs.html

Mach, A., & O'Hara, B. (2011). *Do people really have multiple health insurance plans? Estimates of nongroup health insurance in the American Community Survey*. U.S. Census Bureau. Retrieved February 24, 2022, from https://www.census.gov/content/dam/Census/library/working-papers/2011/demo/SEHSD-WP2011-28.pdf

O'Hara, B. (2010). *Is there an undercount of Medicaid participants in the 2006 ACS Content Test?* U.S. Census Bureau. Retrieved February 24, 2022, from https://www.census.gov/library/working-papers/2010/acs/2010_OHara_01.html

RTI International (2021). *Cognitive testing for the 2022 ACS Content Test: Rounds 1 and 2 combined briefing report*. U.S. Census Bureau. Revised August 30, 2021.

U.S. Census Bureau. (2007). *Subjects planned for the 2010 Census and American Community Survey: Federal legislative and program uses*. Retrieved February 28, 2022, from https://www2.census.gov/programs-surveys/acs/operations_admin/ Final_2010_Census_and_American_Community_Survey_Subjects_Notebook.pdf

10. DISABILITY

Authors: Sharon Stern, Laryssa Mykyta, Natalie Young, Amy Steinweg (SEHSD), Samantha Spiers, and Lauren Contard (DSSD)

The following section describes the specifics of the Disability questions in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the Test version of the question (Section 10.1).
- A discussion of the difference between the Control and Test versions (Section 10.2).
- The specific research questions and the analysis methodology (Section 10.3).

10.1 LITERATURE REVIEW

The National Center for Health Statistics (NCHS) is proposing that the Census Bureau modify the disability questions in the ACS. NCHS proposed that the Census Bureau use the Washington Group Short Set on Functioning (WG-SS) as a replacement to the existing disability question set in the ACS.

10.1.1 Background

The test version of the disability questions was developed by the Washington Group on Disability Statistics (WG), a city group created by the UN Statistical Commission to improve the quality and international comparability of disability statistics worldwide.[21] The group was constituted in response to issues raised at the 2001 International Seminar on the Measurement of Disability about differences in how disability is conceptualized and operationalized within surveys internationally, and the resulting variation in the validity, reliability, and cross-national comparability of national disability estimates. A primary goal of the WG was to establish agreed upon standards for the measurement of disability in national surveys and censuses, in part to improve cross-survey and cross-national comparability of disability estimates (Washington Group on Disability Statistics, 2020). The WG, whose Secretariat is located at NCHS, includes representatives from statistical agencies from across the world. The first product developed by the WG was a short set of questions for use in censuses and surveys – the WG-SS. The WG-SS was endorsed by the Washington Group in 2006, following cognitive testing in the U.S. and fifteen other countries, as well as field testing in the U.S., Argentina, Brazil, Gambia, Paraguay, and Vietnam (Madans, 2017). Since 2006, the WG-SS has been included in censuses and surveys in over 90 countries worldwide (Miller et al., 2020). It is the question set recommended by the United Nations for measuring disability, monitoring the UN Convention on the Rights of Persons with Disabilities, and disaggregating Sustainable Development Goal Indicators and other

---

[21] For more information about the Washington Group on Disability Statistics, visit: https://www.washingtongroup-disability.com/

international commitments (United Nations, 2017; United Nations Economic Commission for Europe, 2015; United Nations General Assembly, 2007; United Nations General Assembly, 2015; United Nations Economic and Social Commission for Asia and the Pacific, 2012). The WG-SS is currently collected as part of the National Health Interview Survey (NHIS)[22] and the National Health and Nutrition Examination Survey (NHANES)[23] and will replace the current disability question set in the National Survey of Family Growth in 2022.[24] In addition, four questions from the WG-SS have been included in the U.S. Census Bureau's Household Pulse Survey.[25]

A related but parallel effort to improve disability statistics collected in federal surveys was initiated by OMB. This effort emerged in response to data user concerns about disability questions included in the 2000 Census and in the ACS. OMB's Interagency Committee for the ACS established an ACS Subcommittee on Disability Measurement in 2003. NCHS was asked to spearhead an evaluation of the ACS disability question set, with the help of other federal agencies (Brault, 2009). The ACS Subcommittee reviewed agency mandates and determined that information on disability was necessary for at least two major reasons: 1) to monitor whether persons with disabilities are being prevented from full participation in society as outlined in the 1990 Americans with Disabilities Act and 2) to estimate the number of persons eligible for service programs offered by state and federal governments (Brault et al., 2007). Similar to the WG, the Subcommittee used the International Classification of Functioning, Disability and Health (ICF) as a conceptual guide for identifying disability domains and an approach to question construction based on a definition of disability located at the person-level, conceptualizing limitations or difficulties as possible risk factors associated with restrictions to full participation in society (World Health Organization, 2001). In January-March 2006, a content test was conducted to evaluate the disability question set proposed by the ACS Subcommittee on Disability Measurement. Both the Census Bureau and NCHS were involved in content testing efforts. In 2008, the disability question set developed by the ACS Subcommittee replaced the previous question set in the ACS. In 2011, in response to the Affordable Care Act signed into law in 2010, the U.S. Department of Health and Human Services (HHS) made the new ACS question set the standard for use in all surveys that are conducted or sponsored by

---

[22] In addition to the WG-SS, the NHIS survey instrument includes a number of questions from the Washington Group Extended Set on Functioning (WG-ES). A copy of the 2021 NHIS survey instrument can be found here: https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2021/EnglishQuest.pdf

[23] Some modifications were made to the WG-SS for inclusion in the NHANES, such as moving the question about communication difficulty to a different place in the question series. A copy of the 2019-2020 NHANES survey instrument can be found here: https://wwwn.cdc.gov/nchs/data/nhanes/2019-2020/questionnaires/FNQ_K.pdf

[24] The planned replacement of the disability question set in the National Survey of Family Growth with the WG-SS is mentioned in the following document: https://www.cdc.gov/hearingloss/CDC-Surveillance-Survey-Questions-on-Hearing-Loss.html

[25] A copy of the Phase 3.3 Household Pulse survey instrument can be found here: https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase3-3_Questionnaire_12_01_21_English.pdf

U.S. Census Bureau

HHS.[26] Thus, the new ACS disability question set was also incorporated into a number of other surveys, including the Survey of Income and Program Participation (SIPP), the National Health Interview Survey (NHIS), and the National Survey of Family Growth.[27]

Since 2008, however, a number of federal surveys have either switched or are planning to switch from the disability question set developed by the ACS Subcommittee in 2006 to the Washington Group question set. Due to a shared conceptual basis and approach to measuring disability, the WG-SS and ACS disability question sets are similar in many ways (see Section 10.2). The main difference between the sets is in the response categories. The WG-SS uses graded response categories, while the ACS uses a dichotomous yes/no response. Although some members of the ACS Subcommittee argued in 2006 that graded response categories would more accurately reflect the continuum of functional abilities, the strict space requirement of the ACS paper form used at that time required that the new questions could take no more space on the form than the existing set of disability questions occupied. As a result, a dichotomous yes/no response was used for each of the six questions and, in some cases, the word "serious" was added to the question stem wording.

10.1.2   Overview of the Washington Group Short Set on Functioning

The WG-SS is a set of six questions designed to identify (in a census or survey format) people with a disability, namely those at greater risk than the general population for participation restrictions due to the presence of difficulties in six core functional domains, if appropriate accommodations are not made. The questions ask whether people have difficulty performing basic universal activities in six domains of functioning: vision, hearing, ambulation, cognition, self-care, and communication. The response options allow for a continuum of difficulty to be reported: 1) no difficulty, 2) some difficulty, 3) a lot of difficulty, 4) cannot do at all. The WG-SS is designed to provide comparable data cross-nationally for populations living in a variety of cultures with varying economic resources.

Although the WG-SS question set includes a question about cognition in the form of difficulty concentrating or remembering, there are psychosocial and cognitive disabilities that do not affect these cognitive activities but that affect communication. The WG-SS question on communication difficulty identifies difficulty with speech and with psychosocial and cognitive issues that affect communication. Unlike the ACS disability question set, the WG-SS does not include a question on independent living difficulty, due to the difficulty in developing a question that would produce internationally comparable data.

---

[26] HHS Implementation Guidance on Data Collection Standards for Race, Ethnicity, Sex, Primary Language, and Disability Status | ASPE

[27] For a more extensive list of surveys that include the disability question set developed by the ACS Subcommittee (or that included the question set in the past), see https://www.cdc.gov/ncbddd/disabilityandhealth/datasets.html

10.1.3   NHIS Analysis: Comparing the WG-SS and ACS Disability Question Set

The Washington Group set of six questions represents a valid, reliable measure of disability with cross-national comparability based on findings from cognitive interviews and field tests conducted by NCHS and other stakeholders within the U.S., as well as in other countries (Altman, 2016). In addition, beginning in 2010, the WG-SS was added to the NHIS instrument.

The NHIS included both the WG-SS and the ACS question set using split samples but in 2011-2012 both questions sets were asked of the same respondent. This provided an opportunity to compare the performance of the WG-SS against that of the existing 6-item ACS disability question set (Weeks et al., 2021). Both question sets were designed to identify the population with disabilities. Disability status in the ACS is defined as reporting difficulty with at least one of the 6 activities in the question set – that is, providing an answer of "yes" on at least one question. Different definitions of disability status can be created using the WG-SS, but if using a dichotomous measure, the WG recommends defining a respondent as having a disability if they report "a lot of difficulty" or "cannot do at all" for at least one activity in the question set. Those who report "no difficulty" or "some difficulty" for all questions are considered to _not_ have a disability.

Weeks et al. (2021) presents results using this unique dataset. The first step in the analysis of the performance of the WG-SS and ACS question sets was to assess consistency in disability status as determined by the WG-SS and the ACS questions. Given the high percentage of respondents reporting "no difficulty" on both question sets, nonagreement between the two measures was generally low, however, the ACS questions identified a larger percentage of the population as having a disability.

A detailed analysis of the response categories was then conducted to assess the nature of the differences in identification. Responses of "no difficulty" were highly concordant with "no" responses to the ACS questions, and responses of "a lot of difficulty" or "cannot do at all" to the WG-SS questions were highly concordant with "yes" responses to the ACS questions. In contrast, respondents who reported "some difficulty" to at least one WG-SS question did not consistently fall into either the "yes" or "no" response category for the ACS questions: 63% responded "no" to all ACS questions, while 37% responded "yes" to at least one ACS question. Notably, respondents who fell into the "some difficulty" category for the WG-SS and responded "yes" to at least one ACS question were classified differently by the two question sets: they were considered to have a disability according to the ACS definition, while they were not considered to have a disability based on the dichotomous WG-SS measure. This discrepancy accounts for the higher estimate of disability prevalence using the ACS question set, relative to the WG-SS. In their report, Weeks et al. (2021) concluded that the population with disability defined by the ACS questions is more heterogenous in functional level than that defined by the WG–SS questions.

10.1.4   Cognitive Interviews in an ACS-like Environment

Two important limitations of the earlier testing of WG-SS warrant discussion. The first limitation relates to the mode of survey administration. The WG-SS has been tested extensively in the context of an interviewer administered data collection and less so in a self-response environment. ACS data collection includes self-administered versions of the questionnaire, including a self-administered paper instrument and a self-administered internet instrument. In recent years, the internet instrument has become increasingly popular, highlighting the importance of assessing whether the WG-SS performs well within the context of a self-administered questionnaire. The second limitation regards the use of proxy reporting. The WG-SS was initially designed to ask individuals about difficulties they experience with activities, such as seeing, hearing, self-care, and communication. The ACS question set, on the other hand, is designed for both self- and proxy-reports of disability; that is, the individual completing the survey is not only asked about their own difficulties with activities, but also about difficulties other members of the household experience.

Prior to the current field test, cognitive interviewing was conducted by both NCHS and the Census Bureau to evaluate the performance of the WG-SS in an ACS-like environment. In addition to assessing other aspects of the WG-SS, one goal of these cognitive interviews was to improve understanding of how the question set performed in self-administered instruments and when used for proxy reporting of disability. Results from the two sets of cognitive interviews are summarized in the sections below.

10.1.5   NCHS-Directed Evaluation

Between August 2019 and February 2020, NCHS conducted 43 cognitive interviews to further investigate differences between the ACS disability question set and the WG-SS, and in the context of the ACS environment. Respondents completed a self-administered paper version of the ACS instrument, with half of the respondents receiving a copy of the form that included the ACS disability question set, and the other half receiving a copy with the WG-SS instead. The two questinonnaires used in NCHS-directed cognitive testing were designed specifically for testing disability content. They had minimal non-disability content and the two versions were the same in all respects except with regard to the disability item. To assess differences between the two question sets in proxy reporting, respondents not only answered the questions for themselves, but also on behalf of other household members.

In order to focus testing on the effect of the answer categories, the WG-SS was modified slightly for use in the cognitive interviews. Specifically, the wording of the ambulatory difficulty and self-care difficulty questions was adjusted to match the wording used in the ACS question set (i.e., "bathing" was used in lieu of "washing all over" and "climbing stairs" was used instead of "climbing steps") and the communication question was removed. Since the question about independent living difficulty in the ACS question set is not in the WG-SS, it was removed from the ACS disability question set for the cognitive interviews. As such, the primary difference

between the two disability question sets tested in the cognitive interviews concerned answer categories. For the WG-SS, respondents could choose among four response options ("no difficulty"; "some difficulty"; "a lot of difficulty"; "cannot do at all"), while the ACS question set used binary response options ("yes"; "no"). After completing the ACS form, respondents were probed to better understand how they interpreted the questions, as well as how they decided among the response categories. Respondents who received the WG-SS version of the disability question set were also asked how they would have responded to the ACS version, and vice versa.

Results from the cognitive interviews suggest that there is some subjectivity in respondents' evaluation of difficulty with activities, both for the ACS and the WG-SS questions sets. Two respondents who experience a similar level of difficulty seeing, for example, may not provide the same answer to the question about vision difficulty. This issue appeared to be more pronounced, however, for the ACS question set. NCHS noted that some respondents who experienced minor difficulty with an activity responded "no" to the relevant ACS question, while others responded "yes." This situation resulted in a high degree of heterogeneity in the functional abilities of people falling into the "yes" category for the ACS disability question set, as well as some heterogeneity among those who responded "no."

In their report, Miller et al. (2020) find that respondents who reported "a lot of difficulty" or "cannot do at all" when presented with the WG-SS version were a more homogenous group, relative to those who responded "yes" to the ACS version. Specifically, respondents who selected "a lot of difficulty" reported experiencing difficulty frequently and in numerous contexts, and those who reported "cannot do at all" said they could not perform the activity in any context. Overall, findings from the cognitive interviews suggest that the dichotomous WG-SS measure of disability captures a population with more severe disability (i.e., in terms of frequency, intensity, and impact) than the population captured by the ACS definition, as well as a group with less heterogeneity in functional ability. Miller et al. (2020) also present evidence that the WG-SS is better able to depict the range of functioning with more consistency, compared to the ACS question set. Finally, in terms of proxy reporting, the cognitive interviews suggested that respondents draw on similar information when responding to either the ACS question set or the WG-SS on behalf of other household members. Specifically, they drew on their own observations of household members to assess frequency and severity of difficulty with activities, or on teachers' or doctors' evaluations if they felt their own observations were inadequate.

10.1.6   Census-Directed Evaluation

The Census Bureau routinely conducts cognitive interviews as part of the process of pre-testing new and modified survey questions. In preparation for the field test portion of the 2022 ACS Content Test, the disability content was tested along with changes in the health insurance and educational attainment questions. While separate evaluations were written for each topic area,

all interviewees had the three topic areas included in the interview. This process was different from the NCHS-directed testing which focused exclusively on disability.

A total of 115 cognitive interviews were conducted in two Rounds (45 interviews in Round 1; 70 interviews in Round 2). In Round 1 of cognitive interviewing, two versions of the instrument were tested across two different modes of survey administration. Respondents were randomly assigned to either Version 1 or Version 2 of the instrument and to either the paper or CAI (Computer-Assisted Interviewing) mode of administration.[28] The two versions of the instrument were then revised in response to findings from Round 1 and underwent further testing in Round 2. Although a third round of cognitive interviews will be conducted for Puerto Rico and group quarters, their results are not yet available and will not inform design of the field test. For more information on the cognitive testing procedures and results from Round 1 and Round 2, see RTI International (2021).

In terms of testing disability content, the cognitive interviews were designed to evaluate different versions of the WG-SS, as well as to assess how respondents decide between the four response categories ("no difficulty"; "some difficulty"; "a lot of difficulty"; "cannot do at all"). In Round 1, two versions of the WG-SS were compared: a version that used the same wording as the question set developed by the WG (Version 2), and a version with some wording modifications to reduce differences with the ACS disability question set currently in production (Version 1). The two versions of the WG-SS were then revised after Round 1 and underwent further testing in Round 2.

In some cases, the cognitive interviews supported the original WG-SS wording, while in other cases, the modified wording performed better. In measuring ambulatory difficulty, for example, Round 1 cognitive interviews indicated that there was some variation in the interpretation of the original WG-SS wording, which asked about "difficulty climbing steps," while the modified WG-SS wording, "difficulty climbing stairs," was interpreted in the same way by nearly all respondents (RTI International, 2021). In Round 2, the disability sub-committee sought to obtain additional information about the performance of the original WG-SS version of the question ("difficulty climbing steps") by only including this version of the question in the Round 2 materials. Overall, most respondents understood the question as intended, leading the sub-committee to recommend that the original WG-SS wording be used in the field test. The field test may provide additional insight into the performance of this version of the ambulatory difficulty question.

The two rounds of cognitive interviews were also an opportunity to evaluate the performance of the WG-SS question about communication difficulty, which is not included in the current ACS

---

[28] In Round 1, Spanish-language cognitive interviews were all CAI interviews, while English-language interviews were split between paper and CAI modes. In Round 2, Spanish-versions of both the self-administered paper questionnaire and the CAI instrument were tested (as well as English versions of both modes).

disability question set. Overall, most respondents who completed the English-language ACS questionnaire understood the communication question as intended, including both monolingual and bilingual individuals. In contrast, the Spanish translation of the communication question did not perform well in Round 1. Some monolingual Spanish speakers misinterpreted the question as asking about difficulty they experienced communicating with non-Spanish speakers (RTI International, 2021). A new translation of the question was proposed and tested in Round 2. Results from Round 2 indicated that the revised Spanish translation was generally understood as intended. Based on these results, the disability sub-committee selected the original WG-SS version of the communication question for inclusion in the English-language survey instruments that will be administered during the field test, and the revised Spanish translation of the question for inclusion in the Spanish-language field test instruments.

In addition, cognitive interviews were used to assess the performance of two versions of the ACS question about independent living difficulty when asked alongside the WG-SS. In the production ACS instrument, the question about independent living difficulty ("difficulty doing errands alone…") is preceded by text intended to reduce reports of independent living difficulty that are due to issues other than disability. Specifically, respondents are asked to only report difficulty "due to a physical, mental, or emotional condition," as opposed to difficulty due to transportation issues, language barriers, or other out-of-scope causes. Since no other questions in the WG-SS include a preamble, in Round 1 a version of the independent living difficulty without a preamble was compared to the ACS production version. Overall, respondents who received the version without a preamble appeared more likely to misinterpret the question. This was particularly true among monolingual Spanish speakers, some of whom reported having difficulty doing errands alone because they are not native English speakers, which is out-of-scope (RTI International, 2021). In Round 2, only the version of the question with a preamble was included in the survey instruments. Results indicated that most respondents understood the question as intended. As such, the disability sub-committee recommended that the field test version of the question about independent living difficulty include a preamble.

Finally, the two rounds of cognitive interviews provided an additional opportunity to assess how respondents choose among the four response categories in the WG-SS. Consistent with findings from cognitive interviews conducted by NCHS (Weeks et al., 2021), most respondents selected "some difficulty" or "no difficulty" when answering the WG-SS questions, while reports of "a lot of difficulty" and "cannot do at all" were relatively uncommon and restricted to individuals with very severe disability. Notably, the cognitive interviews suggested that the "some difficulty" category captured a rather heterogenous group in terms of functional ability, ranging from individuals with outdated eyeglass prescriptions to adults who currently receive disability benefits.

Overall, these results provide further evidence that the dichotomous WG-SS disability measure will likely capture a population with more severe disability than the population captured by the

current ACS disability measure. The WG-SS measure appears to primarily capture individuals who perceive themselves as having an impairment that severely curtails their ability to complete activities on a daily basis, while generally excluding individuals who feel their impairment limits their activities infrequently, to a minimal degree, or only in certain contexts. While this approach is expected to result in lower estimates of disability prevalence in the U.S., it may be more consistent with the social model of disability, which conceptualizes disability as arising only if an individual's impairment leads them to encounter barriers to participating in society (Davis, 2006).

10.1.7   Summary

Advocates for the full inclusion of persons with disabilities have long wanted the ability to disaggregate measures of wellbeing by an overall disability indicator and for specific disability types and severities, reinforcing the importance of including high-quality questions on disability in the ACS. Based on the results of the analysis of the NHIS data and the cognitive interview study, researchers from NCHS concluded that the WG-SS was preferable because the set provides more granular responses that better describe the functional characteristics of the population. Rather than treating disability as a dichotomy, the question set is designed to obtain information on the full extent of difficulties in each domain. Thus, it is possible to disaggregate equity and other measures by an overall indicator of disability, by the severity of difficulties, and by level of difficulty in each of the functional domains, which may increase the policy relevance of the information. Cognitive interviewing was also conducted by the Census Bureau as part of the test for new items to be used on the ACS. The cognitive interviews focused on two versions of the WG-SS, with the intention of evaluating several proposed wording modifications to the standard WG-SS and the graded response options. Many of the findings from the Census-directed cognitive interviews reaffirmed the findings of the NCHS and earlier WG cognitive tests showing that the questions obtained information on the intended constructs in a consistent way across populations. The cognitive interviews also provided valuable information regarding the relative performance of different versions of the WG-SS.

The current field test extends the evaluation of the performance of the WG-SS in the ACS instrument relative to the ACS-6 disability question set. While the Census-directed cognitive interviews were in part designed to assess validity of the WG-SS in the ACS environment, the primary aim of the field test is to evaluate other aspects of performance. The remainder of this document outlines the key research questions for the field test, the analyses that will be conducted to address these research questions, and the criteria for accepting the proposed WG-SS for inclusion in the ACS.

10.2   QUESTION CONTENT

Control and Test versions of each question set are shown as they will appear on the paper questionnaire. Automated versions of the questionnaire will have the same content formatted accordingly for each mode.

Figure 28. Control Version of the Disability Question (Paper)

18 a. Is this person deaf or does he/she have serious difficulty hearing?

☐ Yes

☐ No

b. Is this person blind or does he/she have serious difficulty seeing even when wearing glasses?

☐ Yes

☐ No

H Answer questions 19a – c if this person is 5 years old or over. Otherwise, SKIP to the questions for Person 2 on page 19.

19 a. Because of a physical, mental, or emotional condition, does this person have serious difficulty concentrating, remembering, or making decisions?

☐ Yes

☐ No

b. Does this person have serious difficulty walking or climbing stairs?

☐ Yes

☐ No

c. Does this person have difficulty dressing or bathing?

☐ Yes

☐ No

I Answer question 20 if this person is 15 years old or over. Otherwise, SKIP to the questions for Person 2 on page 19.

20 Because of a physical, mental, or emotional condition, does this person have difficulty doing errands alone such as visiting a doctor's office or shopping?

☐ Yes

☐ No

Figure 29. Test Version of the Disability Question (Paper)

The changes to the disability questions include the following:

- Changes in wording, such as removing the word "serious"
  - VISION
    - Control Version: Is this person blind or does he/she have serious difficulty seeing even when wearing glasses?
    - Test Version: Does this person have difficulty seeing, even if wearing glasses?
  - HEARING
    - Control Version: Is this person deaf or does he/she have serious difficulty hearing?
    - Test Version: Does this person have difficulty hearing, even if using a hearing aid?
  - COGNITION
    - Control Version: Because of a physical, mental, or emotional condition, does this person have serious difficulty concentrating, remembering, or making decisions?

- Test Version: Does this person have difficulty remembering or concentrating?
    - o AMBULATION
        - Control Version: Does this person have serious difficulty walking or climbing stairs?
        - Test Version: Does this person have difficulty walking or climbing steps?
    - o SELF-CARE
        - Control Version: Does this person have difficulty dressing or bathing?
        - Test Version: Does this person have difficulty with self-care, such as washing all over or dressing?
- Changes in answer choices
    - o Control Version: Yes/no
    - o Test Version: No difficulty/Some difficulty/A lot of difficulty/Cannot do at all
- A new question asking about communication difficulty
- A change in question order

Control Version	Test Version
▪ Hearing	▪ Vision
▪ Vision	▪ Hearing
▪ Cognition	▪ Ambulation
▪ Ambulation	▪ Cognition
▪ Self-care	▪ Self-care
▪ Independent living	▪ Communication
	▪ Independent living

10.3   RESEARCH QUESTIONS AND METHODOLOGY

The following section describes the specific analysis to be conducted on the Disability questions in the 2022 Content Test. Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 10.3.1).
- Specifics of the item missing data rate analysis (Section 10.3.2).
- Specifics of the response distribution analysis (Section 10.3.3).
- Specifics of the response reliability analysis (Section 10.3.4).
- Other analysis planned for this topic (Section 10.3.5).

We have numerous research questions, yet relatively few decision criteria. This is due to the unique nature of the disability question change. As described in Section 10.2, the differences between the two question sets can be displayed as question wording and question order changes. However, as laid out in Section 10.1, the question set to be tested is an alternative question set already validated and widely used. Thus, the main goals in the field test are 1) to understand the implications of this change in terms of who is identified as having a disability (and who is not) and how this change in disability measurement will impact the original time

series (for example, by leading to a sharp drop in estimated disability prevalence, relative to prior years), and 2) to evaluate how this question set performs in the ACS context, specifically verifying that no major detriment occurs in terms of missingness or reliability.

These research questions will provide critical insight into how the implementation of the WG questions into the unique ACS environment will affect the production ACS data, and how to interpret the resulting data after the fact. The implementation of the WG questions will be the biggest change in how ACS disability data are collected since 2008 and is expected to significantly change the resulting disability rates. We expect data users will have a lot of questions about the new disability measure, and how it compares to the previous measure. As such, the field test is a critical opportunity to develop a more in-depth understanding of response distributions for the two measures of disability and how they relate to each other. However, the actual decision criteria themselves are focused on identifying any unacceptable data quality issues, as opposed to further validating these already well-established survey questions.

As described in Section 10.1, a major difference between the ACS context and that of the other surveys using the WG question set, is that of mode. For example, the National Health Interview Survey (NHIS) uses exclusively interviewer-administered modes. In contrast, ACS is a multi-mode survey, wherein respondents self-select (non-randomly) into one of three main modes: self-response internet survey, self-response paper (mailed) questionnaire, or computer-assisted interview with a field representative (CAPI). Based on many years of evaluation of ACS disability data, mode is complexly interwoven into disability measurement. Collected disability rates vary dramatically by mode, as do the item-missing rates of the disability questions. This is due to a combination of 1) the characteristics of the sample that selects into each mode, 2) the actual effects of each mode-environment, and 3) the "observer" effect of the interviewer in the CAPI mode as opposed to that of internet and paper. Given the non-random selection into mode, it is difficult to know the specific contributions of each of the above sources of difference. To add to this complexity, the relative prevalence of each mode is changing over time, creating the potential for changes in the underlying characteristics of the subpopulation in each mode across years.

Thus, throughout the research questions, figures are calculated separately by survey mode, and comparisons are made between test and control treatments *using the same mode*. This allows for an apples-to-apples comparison of test versus control. It also allows us to tease out if the difference between question sets varies more in one mode (for any of the reasons previously noted) compared to another mode. We also include figures calculated for the sample overall (i.e., not separated by mode). Given that some modes are more prevalent than others, this allows us to see how any differences across individual modes is reflected as a relative footprint within in the total estimates.

To more fully understand the effect of proxy reporting on question performance, we will calculate estimates separately for self-reported versus proxy-reported disability data. For the internet and CAPI modes, the respondent is identified during the interview and questions are tailored to the household member by filling "you" or the person's name in the question. Hence the evaluation of proxy reporting is most relevant for those modes. For the paper mode, we will assume that respondent is the first person listed.

All estimates based on this construct of self- versus proxy-response are limited due to potential error in the identification of the respondent. For example, we will not know whether the person filling out the internet form is answering for other people or simply recording answers provided by the other household members. Due to these limitations, analyses regarding self-reported versus proxy-reported responses will only be used to understand the data. They will not be used in the decision criteria.

The questions on vision and hearing difficulty apply to all persons. The questions on ambulation, cognition, self-care, and communication difficulty apply to persons ages five and older. The question on independent living applies to persons ages 15 and older.

10.3.1 Benchmarks for Disability

There is no official source of disability statistics that can serve as a true benchmark. A number of surveys collect information on disability. Some use the WG-SS; some use the ACS set; and others use varieties of questions tailored to the specific population of interest. It is well-established in the disability literature that different question sets, or the same questions in different surveys particularly those focusing on specific topics, result in different prevalence estimates (Albrecht et al., 2001; Altman & Bernstein, 2008). In addition, as disability exists on a continuum, the choice of a cut point will greatly affect prevalence. As such, we have no way to evaluate the field test results in the context of their consistency (or inconsistency) with established benchmarks.

10.3.2 Item Missing Data Rates for Disability

Respondents may choose to skip certain items on a questionnaire for a variety of reasons. When this happens, the Census Bureau uses editing methods to allocate a response based on related information provided by the respondent. A respondent's answer is preferred, meaning lower item missing data is generally preferable. However, the Test treatment could have higher item missing data rates than the Control treatment because of the more complex response choices and the addition of communication difficulty (more items, more chance of missing data). Because the WG-SS is being proposed to improve comparability of disability statistics across national surveys (see Section 10.1), a small increase in item missing data rates is an acceptable consequence.

As mentioned in Section 10.2, there are differences in question order between the Control and Test treatments. Since question order affects item missingness, there may be differences in the item missing data rates between the treatments simply due to question order. Additionally, in the ACS we observe effects on item-missingness across the six questions related to varying age universes, and proximity to skip instructions on prior questions within paper mode (which may be mis-read, resulting in over-skipping of subsequent questions).

We will compare the item missing data rates by each disability type[29] as listed in Table 17 instead of by question order. Ideally, we want to examine if the wording differences affect item nonresponse, and not if question order affects it. Including the mode break-down will be especially helpful here, as we observe different patterns of subsequent item missingness across modes, with higher proportions of select item-missingness in self-administered modes (e.g., paper).

Research Question 1

How does the item missing data rate differ between the Test treatment and the Control treatment? This should be examined for both overall disability and for each disability type, and for the three modes individually as well as overall.

We will compare the percentage of eligible people who did not provide a response to any of the disability questions (all items are missing) in the Control treatment with the corresponding percent in the Test treatment. We will also compare the percentage of eligible people who did not provide a response for each of the disability questions separately. (Table 17 provides additional information on the comparison groups.)

On the paper and internet questionnaires, missing responses are those where no boxes are marked. For the CAPI instrument, a response of either "Don't Know" or "Refused" will be considered missing.

For all sets of comparisons, we will exclude persons that have all detailed person questions missing or persons with early breakoffs (i.e., stopped answering the questionnaire before reaching the disability questions).

---

[29] The term "type" is used in Section 10 to reference a specific item in the questionnaire.

Table 17. Alignment of Disability Type in the Control and Test Treatments

Type	Control Treatment	Test Treatment	Eligible Universe
Hearing	**Paper Form 18a**. Is this person deaf or does he/she have serious difficulty hearing?	**Paper Form 18b.** Does this person have difficulty hearing, even if using a hearing aid?	All
Vision	**Paper Form 18b**. Is this person blind or does he/she have serious difficulty seeing even when wearing glasses?	**Paper Form 18a.** Does this person have difficulty seeing, even if wearing glasses?	All
Cognition	**Paper Form 19a.** Because of a physical, mental, or emotional condition, does this person have serious difficulty concentrating, remembering, or making decisions?	**Paper Form 19b.** Does this person have difficulty remembering or concentrating?	Ages 5 and over
Ambulation	**Paper Form 19b.** Does this person have serious difficulty walking or climbing stairs?	**Paper Form 19a.** Does this person have difficulty walking or climbing steps?	Ages 5 and over
Self-Care	**Paper Form 19c.** Does this person have difficulty dressing or bathing?	**Paper Form 19c.** Does this person have difficulty with self-care, such as washing all over or dressing?	Ages 5 and over
Communication	N/A	**Paper Form 19d.** Using his or her usual language, does this person have difficulty communicating, for example understanding or being understood?	Ages 5 and over
Independent Living	**Paper Form 20.** Because of a physical, mental, or emotional condition, does this person have difficulty doing errands alone such as visiting a doctor's office or shopping? (Same on Control and Test)		Ages 15 and over

Because the group of eligible individuals varies across the questions, the overall item missing data rate definition varies by age. In addition, the Test treatment includes an extra question about communication difficulty that is not in the Control treatment, which may affect the overall item missing data rate comparison. The Eligible Universe column lists the items included in the overall item missing data rate for each age group and treatment. For example, the table shows that the overall item missing data rate for people ages 5 to 14 years includes <u>five</u> items in the Control treatment: hearing, vision, cognition, ambulation, and self-care. For people in the same age group in the Test treatment, the overall item missing data rate includes <u>six</u> items: hearing, vision, cognition, ambulation, self-care, and communication. It is worth noting that in at least some modes, the ACS sees higher disability item-allocation rates for children than for

adults. This has implications for comparisons of the allocation rates of disability items with different age universes.

Research Question 2

When calculated separately for proxy and non-proxy reports, how does the item missing data rate differ between the Test treatment and the Control treatment? This should be examined for both overall disability and for each disability type, and for the three modes individually as well as overall.

This research question is performed the same as for research question 1, except that it calculates the proportions of item-missing for two distinct sub-samples: 1) respondents self-reporting their own disability status, and 2) proxy-reports of disability status through another household member.

We will compare item missing data rates between versions using a two-tailed t-test.

10.3.3   Response Distributions for Disability

The following research questions regarding response distributions will provide additional information to data users but are NOT decision criteria.

For information purposes, we will compare mode-specific proportions of people with a disability between the Test and Control treatments, first by disability type and then overall.

Table 17 in the previous section describes which individual items will be compared between treatments. It also explains which disability items are included in the overall disability measure, which varies by age.

As previously stated, there is no benchmark for disability prevalence and prior literature suggests that different measures of disability result in different estimates. Consequently, these comparisons of disability prevalence between the Test and Control treatments will not be a formal part of the decision criteria. Additionally, based on prior cognitive testing and the analysis of NHIS data, the NCHS determined that the graded response categories (as in the test version) had superior validity to the current ACS production dichotomous response categories, resulting in more homogenous groups in terms of functional status, which may be important for evaluating the status of persons with disability.

For the Control treatment, a person has a disability if they answer "yes" to the relevant disability-related question. Respondents who answer "yes" to the question about hearing difficulty, for example, are considered to have a disability, while respondents who answer "no" to the hearing difficulty question do not have this type of disability.

For the Test treatment, a person has four response options on an ordinal response scale. The Washington Group on Disability Statistics (WG) recommends dichotomizing disability status as follows: a person has a disability if they answer "a lot of difficulty" or "cannot do at all" to the relevant disability item, and does not have a disability otherwise.[30] For the purposes of understanding the field test data, we will examine two different definitions of a disability as shown in the following table with the expectation that neither categorization of the Test version will 'match' the yes/no options of the Control version.

Table 18. Definitions of a Disability in the Control and Test Treatments

Definition of a Disability	Disability Category	Control Treatment	Test Treatment
First definition	With a disability	Percentage of people with "yes"	Percentage of people with "a lot of difficulty" or "cannot do at all"
	Without a disability	Percentage of people with "no"	Percentage of people with "no difficulty" or "some difficulty"
Second definition	With a disability	Percentage of people with "yes"	Percentage of people with "some difficulty"; "a lot of difficulty"; or "cannot do at all"
	Without a disability	Percentage of people with "no"	Percentage of people with "no difficulty"

An overall measure of disability will also be constructed. For this measure, a person will be defined as having a disability if they were categorized as having one or more of the individual disability types (i.e., they answered "yes" to at least one question in the disability question set). A person will be defined as not having a disability if they were determined not to have any of the individual disability types (i.e., the did not answer any "yes" responses, and answered at least one "no" to any question in the disability question set).

Research Question 3

When using the first definition of disability, are the proportions of persons with a disability different between the Test treatment and the Control treatment? This should be examined for both overall disability and for each disability type, and for the three modes individually as well as overall.

---

[30] See the Washington Group recommendations online, https://www.washingtongroup-disability.com/analysis /analysis-overview/

We will compare the percentage of people with a disability in the Test treatment versus the Control treatment using the first definition of disability in the Test treatment (see Table 18). First, we will compare disability rates by disability type (e.g., the percentage of people with a hearing-related disability in the Test treatment versus the corresponding percentage in the Control treatment). We will then compare the percentage with any disability type in the Test treatment versus the Control treatment. Based on the results of the NCHS test and analysis we expect the proportions to differ between the Test and Control treatments.

As stated earlier, due to very different estimates yielded from the different survey modes within ACS, we must include a break-down by mode to be able to interpret the comparison results.

Research Question 4

When using the second definition of disability, are the proportions of persons with a disability different between the Test treatment and the Control treatment? This should be examined for both overall disability and for each disability type, and for the three modes individually as well as overall.

We will conduct the same comparisons of disability prevalence between the Control and Test treatments that were described for Research Question 3. However, for Research Question 4, we will use the second definition of disability for the Test treatment (see Table 18). Based on the results of the NCHS test and analysis we expect the proportions to differ between the Test and Control treatments.

Research Question 5

In the Test treatment, what is the distribution of disability responses by the four answer categories? This should be calculated for overall disability, for each disability type, and for overall disability and disability type by mode.

We will examine the distribution of answer choices and how they vary across mode to better understand disability responses in the Test treatment.

We will compare proportions between Control and Test treatments using a two-tailed t-test.

10.3.4   Response Reliability for Disability

We will also evaluate differences in response reliability between the Control treatment and Test treatment. Reliability and validity are the two cornerstones of good survey measures (Fowler, 2014). That is, a good question should measure what it is intended to measure (validity), and it should consistently produce the same response each time it is asked (reliability). We are unable to use data from the field test to compare the validity (and

conversely, the response bias) of the two disability measures; cognitive testing is better suited for this purpose and the conclusions from the cognitive test conducted by NCHS suggested superior validity of the WG-SS questions on which the Test version is based. However, the field test can evaluate how response reliability compares across the two question sets through measuring response variance. A measure with higher response reliability (and conversely, lower response variance) is generally preferable. However, any advantage in reliability must be evaluated against the differences in validity based on cognitive testing.

Research Question 6

When using the first definition of disability, are the measures of response reliability (GDR, IOI) different between the Test treatment and the Control treatment? This should be examined for both overall disability and for each disability type, and for the three modes individually as well as overall.

Research Question 7

When using the second definition of disability, are the measures of response reliability (GDR, IOI) different between the Test treatment and the Control treatment? This should be examined for both overall disability and for each disability type, and for the three modes individually as well as overall.

To address Research Question 6 and Research Question 7, we will compare measures of response reliability between the Test and Control treatments overall and for each disability type. Table 17 describes which disability type items are being compared between treatments.

We will also compare measures by mode of original interview. In addition to differences by mode previously discussed, interviewer follow-up after completing the survey by self-response may also impact interpretation of results.

The simple response variance measures (GDR and IOI) are calculated for individual response categories (e.g., yes or no). The Control treatment has only two response categories, and thus the response categories do not need to be collapsed to calculate GDR and IOI. The Test treatment, however, has four response categories, and thus the response categories will be collapsed into yes/no using either the first or second definition of a disability as described in Table 18. Then GDR and IOI will be calculated as discordance in yes/no between the original interview and reinterview.

For Research Questions 6 and 7, we will compare response reliability measures between versions using a two-tailed t-test.

Research Question 8

In the Test treatment, are the measures of response reliability (GDR, IOI) consistent between the two different dichotomizations of disability? This should be examined for both overall disability and for each disability type, and for the three modes individually as well as overall.

While not part of the decision criteria, we will also compare measures of response reliability across the two Test treatment definitions of disability to better examine how each definition functions in the context of the ACS instrument. That is, is response reliability in the Test treatment higher when we use one definition of disability over another?

Research Question 8.1

When calculated for same versus different respondent in follow-up interview: When using the first definition of disability, are the measures of response reliability (GDR, IOI) different between the Test treatment and the Control treatment? This should be examined for both overall disability and for each disability type, and for the three modes individually as well as overall.

This research question is identical to that of Research Question 6, but it calculates numbers separately by whether or not the follow-up was reported by the same individual who responded to the original questionnaire. We expect that there will be more inconsistency between reporting in initial versus follow-up interview when different respondents report. Thus, this research question attempts to net out the effect of that change in reporting person.

Research Question 8.2

In the Test treatment, what are the measures of response reliability (GDR, IOI) when using the four answer categories (rather than the dichotomization into yes/no)? This should be examined for each disability type, and for the three modes individually as well as overall.

Finally, while not a formal research question, we would like to more fully understand response variability in the Test Treatment. Therefore, we will examine the response variance of each particular response category between original interview and reinterview. For example, we want to examine for how many people a response of "no difficulty" in the initial interview remains "no difficulty" in the reinterview, a response of "some difficulty" in the initial interview remains "some difficulty" in the reinterview, etc. In other words, we are interested in learning whether some individual response categories are more consistent than others. We may also calculate IOI_L for the Test treatment to know the overall response reliability of the disability questions.

10.3.5   Other Metrics for Disability

The following research questions are to provide additional information but are NOT decision criteria.

Changes to survey instruments should not unnecessarily increase respondent burden, since high levels of respondent burden can lead to measurement error. To assess an aspect of respondent burden in the Control and Test treatments, we will make use of paradata from the internet and CAPI instruments.

Research Question 9

Is there a difference in help text use on the internet and CAPI instruments between the Test treatment and the Control treatment?

Greater usage of the help text on the internet and CAPI instruments can be considered an indicator of respondent burden. That is, respondents who are confused by a question are more likely to click on help text than are respondents who understand a question as initially asked.

To address Research Question 9, we will compare the percentage of people using help text (on any screen within the disability section) on the internet version of the Test treatment to the corresponding percentage for the internet version of the Control treatment. We will also compare the percentage of people using help text on the CAPI version of the Test treatment to the percentage using help text on the CAPI version of the Control treatment.

Research Question 10

Is there a difference in behavior on the internet instrument between the Test treatment and the Control treatment with regards to:

a. *Switching between answer choices while on the same screen*
b. *Going back to previous screens*
c. *Time spent on each question screen*
d. *Time spent on the overall disability question set*

The ACS internet instrument collects paradata that can shed light on respondent burden. In particular, respondents who are confused by a question or unsure of how to answer may be expected to switch between answer choices, move back and forth in the survey instrument, or spend a greater amount of time on each survey question.

We will compare these proxy measures of respondent burden between the Test treatment and Control treatment. For example, is there a difference in the proportion of respondents in the Test and Control treatments who switch between answer choices?

Analysis for Research Questions 9 and 10 will be partially covered by a separate planned respondent burden analysis using internet instrument paradata for all topics in the 2022 Content Test (see Section 3.2.1).

For all research questions, statistical significance between versions will be determined using a two-tailed t-test.

10.4   DECISION CRITERIA

Section 10.3 presents ten research questions with explanations providing what we hope to learn from the analysis. All of the research questions will be used to address the first goal described in Section 10.3: understanding who is identified as having a disability and how this change in disability measurement will impact the original time series. Due to the unique nature of the proposed change to the WG-SS, the decision criteria are specifically focused on identifying any major unexpected quality implications. The decision criteria, therefore, focus on missingness and reliability.

Due to the limitations described previously with the identification of the respondent, Research Question 2 will not be used in the decision criteria.

The most important results of this analysis when drawing a conclusion about the Test version compared to the Control version are, in order of priority:

Table 19. Decision Criteria for Disability

Priority	Research Question*	Decision Criteria
1	6, 7	A lower GDR or IOI indicates higher reliability and is preferable. Because we expect more variability between individual responses in the Test treatment (4 choices offer more opportunity for switching), the decision criteria will rely on a comparison of the GDR/IOI for the summary measure of overall disability - "with a disability" or "without a disability". Although a lower GDR or IOI is preferable, lower reliability in the Test treatment may be acceptable when considering the documented cognitive testing validity of the Test treatment. Because the self-response option was not available in earlier tests of the proposed questions (Test Version), it is particularly important that the GDR and IOI for the internet and paper original interview modes be of similar order of magnitude as interviewer mode when compared to the Control version within the same modes.

Priority	Research Question*	Decision Criteria
2	1	In general, lower item missing data rates are preferable. However, the Test treatment could have higher item missing data rates than the Control treatment simply because of the more complex response choices and the addition of communication difficulty (more items, more chance of missing data). Thus, a small increase in item missing data rates is an acceptable consequence. Because the self-response option was not available in earlier tests of the proposed questions (Test version), it is particularly important that the item missing data rates for the self-response modes (internet and paper) be of similar order of magnitude as interviewer mode when compared to the Control version within the same modes.

*Research questions within a priority are organized with smallest research question number listed first. Research questions not included in the decision criteria are for research purposes only.

10.5 REFERENCES

Albrecht, G. L., Seelman, K., & Bury, M. (Eds.). (2001). *Handbook of disability studies*. SAGE Publications.

Altman, B. M., & Bernstein, A. (2008). *Disability and health in the United States, 2001-2005*. National Center for Health Statistics. https://stacks.cdc.gov/view/cdc/6983

Altman, B. M. (Ed.). (2016). *International measurement of disability: Purpose, method, and application.* Springer.

Brault, M. W. (2009). *Review of changes to the measurement of disability in the 2008 American Community Survey*. U.S. Census Bureau. Retrieved February 24, 2022, from https://www.census.gov/library/working-papers/2009/demo/brault-01.html

Brault, M., Stern, S., & Raglin, D. (2007). *Evaluation report covering disability* (2006 American Community Survey Content Test Report P.4). U.S. Census Bureau. Retrieved February 24, 2022, from https://www.census.gov/library/working-papers/2007/acs/2007_Brault_01.html

Davis, L. J. (2006). *The disability studies reader*. Routledge.

Fowler, F. J. Jr. (2014). *Survey Research Methods*. SAGE Publications.

Miller, K., Vickers, J. B., & Scanlon, P. (2020). *Comparison of American Community Survey and Washington Group disability questions*. National Center for Health Statistics. https://wwwn.cdc.gov/QBank/Report.aspx?1216.

Madans, J. H., Loeb, M. E., & Altman, B. M. (2011). Measuring disability and monitoring the UN Convention on the Rights of Persons with Disabilities: The work of the Washington Group on Disability Statistics. *BMC Public Health*, 11(S4). https://doi.org/10.1186/1471-2458-11-S4-S4

Madans, J. H. (2017, July 14). *Mainstreaming disability data: The Washington Group on Disability Statistics* [PowerPoint presentation]. UNSD Expert Group Meeting on Guidelines and Principles for the Development of Disability Statistics, New York, NY, United States.

RTI International (2021). *Cognitive testing for the 2022 ACS Content Test: Rounds 1 and 2 combined briefing report*. U.S. Census Bureau. Revised August 30, 2021.

United Nations. (2017). *Principles and recommendations for population and housing censuses: Revision 3*. https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf

United Nations Economic Commission for Europe. (2015). *Conference of European Statisticians recommendations for the 2020 census of population and housing*. https://unece.org/DAM/stats/publications/2015/ECECES41_EN.pdf

United Nations General Assembly (2007, January 24). *Convention on the Rights of Persons with Disabilities: Resolution / adopted by the General Assembly* (A/RES/61/106). https://www.refworld.org/docid/45f973632.html

United Nations General Assembly (2015, October 21). *Transforming our world: The 2030 Agenda for Sustainable Development* (A/RES/70/1). https://www.refworld.org/docid/57b6e3e44.html

United Nations Economic and Social Commission for Asia and the Pacific (2012). *Incheon strategy to "make the right real" for persons with disabilities in Asia and the Pacific*. https://www.unescap.org/sites/default/d8files/knowledge-products/Incheon%20Strategy%20%28English%29.pdf

Washington Group on Disability Statistics (2020). *An introduction to the Washington Group on Disability Statistics questions sets*. https://www.washingtongroup-disability.com/fileadmin/uploads/wg/Documents/primer.pdf

Weeks, J. D., Dahlhamer, J. M., Madans, J. H., & Maitland, A. (2021). *Measuring disability: An examination of differences between the Washington Group Short Set on Functioning and the American Community Survey disability questions* (National Health Statistics Reports No. 161). National Center for Health Statistics. https://www.cdc.gov/nchs/data/nhsr/nhsr161-508.pdf

World Health Organization (2001). *International Classification of Functioning, Disability and Health*. https://icd.who.int/dev11/l-icf/en

11. SUPPLEMENTAL NUTRITION ASSISTANCE PROGRAM (SNAP)

Authors: Lindsay Monte, Tracy Loveless (SEHSD), and Dorothy Barth (DSSD)

The following section describes the specifics of the Supplemental Nutrition Assistance Program (SNAP) question in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the Test version of the question (Section 11.1).
- A discussion of the changes between the Control and Test versions (Section 11.2).
- The specific research questions and the analysis methodology (Section 11.3).

11.1   LITERATURE REVIEW

Changes are being made to the SNAP question to more closely align the reference period for administrative records and with the labor force and income questions. See the Literature Review for the Labor Force and Income questions in Sections 12.1 and 13.1, respectively, for a detailed explanation on why this change is occurring.

11.2   QUESTION CONTENT

Control and Test versions of each question are shown as they will appear on the paper questionnaire. Automated versions of the questionnaire will have the same content formatted accordingly for each mode. The only change made to the SNAP question is the year of reference, from "THE PAST 12 MONTHS" to the prior calendar year. For the 2022 ACS Content Test, the prior calendar year will be 2021.

Figure 30. Control Version of the SNAP Question (Paper)
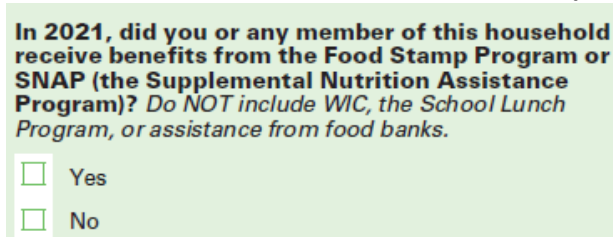
IN THE PAST 12 MONTHS, did you or any member of this household receive benefits from the Food Stamp Program or SNAP (the Supplemental Nutrition Assistance Program)? *Do NOT include WIC, the School Lunch Program, or assistance from food banks.*

☐ Yes

☐ No

Figure 31. Test Version of the SNAP Question (Paper)

In 2021, did you or any member of this household receive benefits from the Food Stamp Program or SNAP (the Supplemental Nutrition Assistance Program)? *Do NOT include WIC, the School Lunch Program, or assistance from food banks.*

☐ Yes

☐ No

11.3 RESEARCH QUESTIONS AND METHODOLOGY

The following section describes the specific analysis to be conducted on the SNAP question in the 2022 Content Test.

11.3.1 Benchmarks for SNAP

For the SNAP question, we are testing a change to the reference period from "the past 12 months" to the previous calendar year. Using the prior calendar year aligns with the reference period for administrative records and with the labor force and income questions. (See Sections 12 and 13.) Our hypothesis is that this change will NOT affect reported SNAP recipiency. We will assess the validity of this hypothesis by comparing content test estimates (Control and Test) to estimates from the CPS-ASEC SNAP questions. We will also compare estimates to official counts from the Food and Nutrition Service (FNS) at USDA. The ACS has traditionally undercounted SNAP receipt relative to FNS official counts; we want to see if the change in reference period affects our estimation relative to those counts. We will also compare Content Test estimates to prior years of ACS data (2017-2019). These will be only observational comparisons to determine if our results are grossly different from other sources; we will not perform statistical testing.

Research Question 1

How do the estimated proportions of households receiving SNAP in the Test treatments compare with estimates from the 2021 CPS ASEC, 2021 FNS counts, and ACS data from 2017 to 2019?

Since 2020 ACS data were experimental, we will not include it in this analysis.

Research Question 2

How do the estimated proportions of household receiving SNAP in the Control treatment compare with estimates from the 2021 CPS ASEC, 2021 FNS counts, and ACS data from 2017 to 2019?

11.3.2 Item Missing Data Rates for SNAP

To evaluate item nonresponse, we will calculate item missing data rates for the SNAP question and compare the Test and Control treatments. We will make comparisons for each response mode separately and all response modes combined.

Research Question 3

Is there a difference in item missing data rates for the test and control versions of the SNAP question?

All households in sample are eligible to answer the SNAP question, since it is in the household section of the survey. It is a Yes/No question, so when no box is checked the item response will be considered missing. For CAPI, a "Don't Know" or "Refusal" will be considered a missing response.

11.3.3   Response Distributions for SNAP

Research Question 4

Is the proportion of households reporting that they receive SNAP benefits different for the test versions than for the control version?

We will calculate and compare proportions between Control and Test treatments for each response mode separately and all combined.

11.3.4   Other Metrics for SNAP

There are no additional metrics for the SNAP questions. SNAP is not being asked in CFU. The primary goal for changing the SNAP question is to align the reference period of the questions more closely with administrative records. Since the only change to the question is the reference period, the reliability of responses will be determined by benchmark comparisons of Content Test data to administrative records and other known data sources.

11.4   DECISION CRITERIA

The most important results of this analysis when drawing a conclusion about the Test version compared to the Control version are shown below, in order of priority:

Table 20. Decision Criteria Changing the Reference Period for SNAP

Priority	Research Question	Decision Criteria
1	4	Response Distributions: We hope to see either no difference in the proportion or a greater proportion of people receiving SNAP.
2	3	Item Missing Data Rates – We hope to see no difference or lower item nonresponse.
3	1, 2	Benchmarks – We expect to see comparable estimates.

Regardless of the decision criteria outlined above, the reference period used for SNAP will align with the reference period used for the ACS Income questions.

11.5   REFERENCES

Not applicable.

12. LABOR FORCE

Authors: Mark Klee, Brian Mendez Smith (SEHSD), and Dorothy Barth (DSSD)

The following section describes the specifics of the Labor Force questions in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the Test version of the question (Section 12.1).
- A discussion of the changes between the Control and Test versions (Section 12.2).
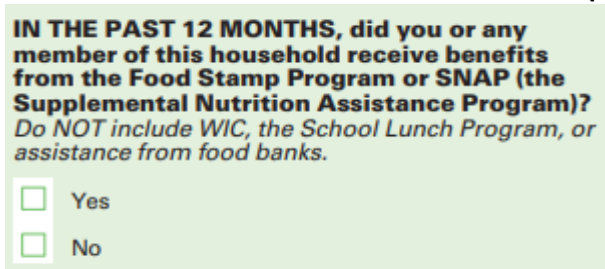- The specific research questions and analysis methodology (Section 12.3).

12.1 LITERATURE REVIEW

A question collecting information on a person's employment has been asked on the census in some form since 1850. The current version of the question series was introduced in 2008 and was updated in 2019 after being tested in the 2016 ACS Content Test (Smith, Howard, and Risley, 2017). The changes proposed to the Labor Force question series are linked to the changes proposed to the Income series of questions, which change the reference period from "the past 12 months" to the prior calendar year. For example, for the 2021 ACS "In the past 12 months" would become "In 2020." Changing the reference period to the prior calendar year will allow the ACS to better align with administrative records, which, if used in production, could improve the quality of ACS estimates.

The reference period for the Number of Weeks Worked and Number of Hours Worked questions has been "during the past 12 months" continuously since the initial ACS questionnaire in 1996. However, there is precedence for a prior year reference period for the Number of Weeks Worked and Number of Hours Worked questions The Census 2000 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?" "During the weeks WORKED in 1999, how many hours did this person usually work each WEEK?". Income questions in the Census 2000 long form also referred to 1999. Consequently, one might arguably conceive of the proposed change in reference period not only as paving the road to the ACS of the future, but also as returning to a past convention that has already been tried and tested.

Another difference between the Census 2000 long form questions and the ACS questions for 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, we will do analysis to see if the test affects the data collected for those questions.

Administrative data sources on employment, income, and public assistance benefits from the Internal Revenue Service, Social Security Administration, 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 and employment. This test is a significant step in that process.

Jarmin (2019) laid out a vision for how administrative records could help the federal statistical system respond to daunting challenges such as declining survey response rates. According to this vision, "survey design at statistical agencies should over time be optimized to take advantage of the administrative data available to the agencies, as well as new alternative sources. This optimized survey design involves minimizing the use of surveys to collect information for which administrative or alternative source data are available and prioritizing those items not captured in non-survey sources for inclusion in survey data collections."

The "Agility in Action" report series summarizes areas of research into challenges and opportunities for the American Community Survey to take advantage of administrative data available to the agencies. These reports detail the progress of research into whether and how administrative records on wages and self-employment income may be incorporated into the ACS. Exactly how these records would be incorporated into ACS remains undetermined. As noted in the Census Bureau's Agility in Action (1.0) report, "for some topics, records may be better suited in assisting with imputation, whereas for other topics the records may be used for direct substitution of a survey question (for all or a subset of the ACS respondent pool)." In addition, "in some cases, the quality of the records may be more accurate than the respondent's recollection (e.g., W2 information for wages). In some 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, Ortman, Pharris-Ciurej, and Clark (2018) argue that "the utility of IRS records depends largely on the degree to which the IRS tax year coincides with an ACS respondent's period of reference in the survey. Despite temporal alignment issues, IRS administrative records provide substantial coverage for the incidence of income from various sources (e.g., wages, self-employment income, and dividend/interest income) included on 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 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)." 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." These results further underscore both the feasibility and the potential benefits of using administrative records to increase the coverage and quality of employment data in the ACS. Changing the ACS reference period to align with the calendar year will remove the existing temporal barrier to administrative record linkage.

Existing research on Survey of Income and Program Participation (SIPP) data, which collects monthly income and weekly employment that can be aggregated to create a calendar year reference period, 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. 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 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 distribution of positive earnings. 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.

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. (forthcoming) 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.

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 report, RTI International further cognitively tested versions of the questions specifically for this Content Test (RTI International, 2021). The improvements made to the questions, as a result of cognitive testing, are outlined in Section 12.2 below.

12.2   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?"
- 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 Weeks Worked and Hours Worked questions.
- Added the instruction "Include all jobs for pay." to the questions on Weeks Worked.
- Modified the instructions for Version 1 to use a note with bullet points to see if it makes the question instructions clearer and easier to understand and read. (Paper only.)
- 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 and Test versions of each question are shown as they will appear on the paper questionnaire. Automated versions of the questionnaire will have the same content formatted accordingly for each mode. Test Version 1 and Test Version 2 only differ on paper; the internet and CAPI questions are identical. Version 1 on paper uses a note and bullet points to draw readers' attention to important instructions. Version 2 on paper keeps all instructions under the questions; this version is more similar to the way the test questions will be asked on the internet and in CAPI.

Figure 32. Control Version of the 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 33. Test Version 1 (Left) and Test Version 2 (Right) of the Labor Force Questions (Paper)

Test Version 1 (Left)

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 44*

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

☐☐☐

Test Version 2 (Right)

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 44*

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

☐☐☐

12.3 RESEARCH QUESTIONS AND METHDOLOGY

The analysis for Labor Force will adhere to the following definitions:

YEAR-ROUND WORKERS are those who report working 50, 51, or 52 weeks of work for the year (either by marking "Yes" to 52 weeks or by writing in the number of weeks). Part-year workers are those who report working less than 50 weeks for the year.

FULL-TIME WORKERS are those who report usually working 35 or more hours per week. Part-time workers are those who report usually working less than 35 hours per week.

12.3.1 Benchmarks for Labor Force

The CPS and SIPP full-time, year-round statistic and the CPS distributions of weeks worked will be compared to the Test treatments (combined) and Control results. The results cannot be statistically compared but similarities and trends can be discussed.

Research Question 1

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

Research Question 2

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

This ratio is the number of civilians employed (either full- or part-time, either full-year or part-year) in 2021 divided by the total civilian population age 16+ (in 2022).

Research Question 3

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

Research Question 4

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

Research Question 5

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

Research Question 6

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

Research Question 7

How does the distribution of weeks worked (age 16+) for the Test treatments compare with SIPP estimates?

Research Question 8

How does the distribution of hours worked (age 16+) for the Test treatments compare with SIPP estimates?

12.3.2 Item Missing Data Rates for Labor Force

All comparisons of the Test treatments, Version 1 vs. Version 2, will only be calculated for mail responses since the only difference between the two versions is on the paper questionnaire. For all other analyses, we will compare the Control to Version 1.

Research Question 9

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?

We will calculate and compare item missing rates for Version 1 vs. Control. The rates will be calculated for each data collection mode separately and for all modes combined. We will calculate and comare the rates for mail responses for Version 1 vs. Version 2.

Research Question 10

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?

We will calculate and compare item missing rates for Version 1 vs. Version 2 mail responses.

Research Question 11

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?

We will calculate and compare item missing data rates for Control and Version 1. The rates will be calculated for each data collection mode separately and for all modes combined. For Version 1 vs. Version 2 we will only calculate the rates for mail responses.

Research Question 12

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?

We will calculate and compare item missing data rates for Version 1 vs Control. The rates will be calculated for each data collection mode separately and for all modes combined. For Version 1 vs. Version 2 we will calculate rates for mail responses only.

For each question, statistical significance between treatements will be determined using a two-tailed t-test.

12.3.3   Response Distributions for Labor Force

All comparisons of the Test treatments, Version 1 vs. Version 2, will only be calculated for mail responses since the only difference between the two versions is on the paper questionnaire. For all other analyses, we will compare the Control to Version 1.

Research Question 13

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?

We will calculate and compare response distributions for Version 1 vs. Control. The rates will be calculated for each data collection mode separately and for all modes combined. Version 1 vs. Version 2 will only calculate and compare mail responses.

The response categories are: Within the past 12 months, 1 to 5 years ago, and Over 5 years ago or never worked.

Research Question 14

Is there a difference in the proportion of people working in 2021 between Test treatments?

We will calculate and compare proportions for Version 1 vs. Version 2. The rates will be calculated for each data collection mode separately and for all modes combined. This is a Yes/No question. We will only use mail responses.

Research Question 15

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

We will calculate and compare the proportion of full-time and year-round workers for Version 1 vs Control. The rates will be calculated for each data collection mode separately and for all modes combined. For Version 1 vs. Version 2 we will only calculate for the mail responses.

Research Question 16

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?

We will calculate and compare the proportion of part-year workers for Version 1 vs Control. The rates will be calculated for each data collection mode separately and for all modes combined. For Version 1 vs. Version 2 we will only include mail responses.

The continuous number of weeks worked will be aggregated into categories to facilitate the comparison of distributions. The categories will be 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. We will also conduct a second set of comparisons to examine whether the longer recall period induces more heaping on round responses. This analysis will recategorize continuous weeks worked into two categories: values ending in "0" or "5" and all other values.

Research Question 17

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?

We will calculate and compare response distributions for Version 1 vs. Control. The rates will be calculated for each data collection mode separately and for all modes combined. Version 1 vs. Version 2 will only include mail responses.

The continuous number of hours worked will be aggregated into categories to facilitate the comparison of distributions. The categories will be 1 to 14 hours, 15 to 34 hours, 35 to 39 hours, 40 hours, 41 to 49 hours, 50 to 59 hours, and 60 or more hours. Since this distribution is to be analyzed by full-time, year-round worker status (i.e., full-time, year-round vs not full-time, year-round), the 1 to 14 hours and 15 to 34 hours bins will be empty mechanically for workers who are full-time, year-round. There will be no mechanically empty bins for workers who are not full-time, year-round because full-time workers are not full-time, year-round if they work only part of the year. We will also conduct a second set of comparisons to examine whether the longer recall period induces more heaping on round responses. This analysis will recategorize continuous hours worked into two categories: values ending in "0" or "5" and all other values.

Attachment C

12.3.4   Response Reliability for Labor Force

Research Question 18

Is the GDR of When Last Worked different between the Control and Version 1?

We will calculate and compare gross different rates for Control and Version 1. The rates will be calculated for each data collection mode (of the original interview) separately and for all modes combined.

Research Question 19

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

We will calculate and compare GDR for Control and Version 1. The rates will be calculated for each data collection mode (of the original interview) separately and for all modes combined.

Research Question 20

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

We will calculate and compare GDR for Control and Version 1. The rates will be calculated for each data collection mode (of the original interview) separately and for all modes combined.

Statistical significance between the GDRs will be determined using a two-tailed t-test.

Reliability will be assessed by the proportion of persons with a difference of between the original response and the CFU response. This means comparing the actual write-in values of the original response and the CFU response, which will provide insight into how responses change. However, if responses change only slightly, there may be limited impact on estimates in official products. Consequently, we also want to assess reliability by comparing whether the original response and the CFU response fall into the same bin of weeks worked and hours worked that our official products use.

For Weeks Worked, reliability will also be assessed by the proportion of persons falling into 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 Hours Worked, reliability will also be assessed by the proportion of persons falling into different weeks 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.

12.3.5   Other Metrics for Labor Force

Results from both Control and Test versions (combined) will be compared with administrative data, contingent upon the receipt and matching of the Longitudinal Employer-Household Dynamics (LEHD) administrative data. LEHD Data are based on different administrative sources, primarily Unemployment Insurance (UI) earnings data and the Quarterly Census of Employment and Wages (QCEW), and censuses and surveys. Firm and worker information are combined to create job level quarterly earnings history data.

The LEHD data can indicate whether someone participated in the labor force in a given quarter for all states, assuming their job was covered by unemployment insurance.

Since the only difference between both Test Versions is a minimally different design on the paper version of the question, we will combine the Test Treatments for this analysis. Both test versions will produce data that comes from using the calendar year as a reference.

Research Question 21

Is there a difference in the mean number of quarters worked between the Test treatments and the Control treatment?

We will calculate and compare the mean number of quarters worked for the Control vs. the Test treatments combined.

Research Question 22

Do the Test and Control distributions of When Last Worked, Number of Weeks Worked, and Number of Hours Worked compare to the administrative data in a similar fashion?

With some assumptions, the administrative data can give upper and lower bound estimates for the numbers of people who worked year-round and part-year.

To the extent feasible, the hours and weeks data available in certain states will be used to assess the quality of both the Control and Test versions. We should be cautious of these estimates, however. Dividing by usual hours worked introduces additional measurement error (Borjas, 1980). There is also potential lumping in the LEHD data, where employers report the quarterly hours as the number of weeks in the quarter x 40. Because this information is only available for certain states, and the information is collected differently for different states, sample sizes may be too small to be meaningful. Therefore, while this could prove to be a useful analysis, it should not be considered the most important benchmark for the continuous distribution.

One important caveat of using the LEHD to assess weeks worked is that it suffers from important coverage gaps, including self-employed individuals and workers not covered by

unemployment insurance. Other administrative records that may be used to assess the quality of both the Control and Test versions for individuals falling into these coverage gaps include the Information Returns Master File – W2 and the 1040 Returns Master File. However, these datasets are expected to yield limited additional insight into how Test and Control distributions compare to the administrative data. The Information Returns Master File – W2 is available only at the person-year level. Consequently, it can provide insight into When Last Worked and, with some assumptions, Weeks Worked. The 1040 Returns Master Family suffers from these same drawbacks, and additionally it is available only at the tax filing unit level and it does not contain information on self-employment earnings. Consequently, it can only offer insight into When Last Worked.

12.3.5.1   Additional Analyses for Class of Worker, Industry, and Occupation Question Series

Significant changes were made to the Labor Force series of questions, including a change in the reference period. The section following Labor Force is the industry, occupation, and class of worker question set, which uses different reference periods based on employment in the last week or in the last five years. The Census Bureau publishes estimates of full-time, year-round workers by class of worker, industry, and occupation, so we will analyze if there is an impact on the resulting estimates. We will also analyze if there is an impact to item response to the Class of Worker (COW) and Industry and Occupation (I&O) questions because of the changing reference periods.

The universe to calculate item missing data rates for the COW and I&O questions are persons 15 years old or older who (1) did any work last week, (2) worked within the past 12 months, or (3) worked 1-5 years ago.

When Version 1 is referred to below, we mean the set of COW and I&O questions that follow Version 1 of the Labor Force questions. Similarly, Version 2 refers the set of COW and I & O questions that follow Version 2 of the Labor Force questions. There are no differences in the wording for the COW and I&O questions.

<u>Research Question 23</u>

Is there a difference in the item missing data rates for Class of Worker between the Control and Test treatments?

For the Class of Worker question, we will calculate and compare item missing rates for Version 1 vs. Control and Version 2 vs. Control. The rates will be calculated for each data collection mode separately and for all modes combined.

The class of worker question is "Which one of the following best describes this person's employment last week or the most recent employment in the past 5 years? Mark ONE box." For the purposes of this analysis, we will count Mail mode responses where multiple (two or more)

class of worker categories are selected (checked) as responses, even though the variable is blanked in normal ACS processing to create the unedited data file.

Research Question 24

Is there a difference in the item missing data rates for Industry between the Control and Test treatments?

For the Industry of work question, we will calculate and compare item missing rates for Version 1 vs. Control and Version 2 vs. Control. The rates will be calculated for each data collection mode separately and for all modes combined.

The industry questions are: "What was the name of this person's employer, business, agency, or branch of the Armed Forces?" and "What kind of business or industry was this?" We will consider a response to be missing only if both industry write-ins questions are blank.

Research Question 25

Is there a difference in the item missing data rates for Occupation between the Control and Test treatments?

For the Occupation question, we will calculate and compare item missing rates for Version 1 vs. Control and Version 2 vs. Control. The rates will be calculated for each data collection mode separately and for all modes combined.

The occupation questions are: "What was this person's main occupation?" and "Describe the person's main activities or duties." We will consider a response to be missing only if both occupation write-ins questions are blank.

Research Question 26

Is there a difference in the section missing data rates for all part of the Class of Worker, Industry, and Occupation series of questions between the Control and Test treatments?

We will consider the entire section to be "missing" if there are no valid answers for any of the questions in the section. We will calculate and compare the rates for Version 1 vs. Control and Version 2 vs. Control. The rates will be calculated for each data collection mode separately and for all modes combined.

The item missing data rate is the weighted sum of eligible persons for whom the response is missing divided by the weighted sum of all eligible persons. We will test for significance of differences between Control and Test using two-tailed t-tests.

Research Question 27

Are the distributions of full-time, year-round workers by NAICS industry sectors different between Test treatments and the Control Treatment?

The North American Industry Classification System (NAICS) industry sectors are defined by the 2017 North American Industry Classification System. The 27 industry sectors and their corresponding census industry code are indicated in Table 21.

We will compare each pair of distributions of full-time, year-round workers among the NAICS industry sectors, Control versus each Test treatment, using a chi-square test. If the chi-square statistic is significant, we will also compare category proportion estimates, using the Hochberg technique to control family-wise error (FWE).

Table 21. Census Industry Codes Corresponding to NAICS Industry Codes

Description of NAICS Industry Sector	Range of Census Industry Codes
Agriculture, forestry, fishing and hunting	0170-0290
Mining quarrying, and oil and gas extraction	0370-0490
Construction	0770
Manufacturing	1070-3990
Wholesale trade	4070-4590
Retail trade	4670-5790
Transportation and warehousing	6070-6390
Utilities	0570-0690
Information	6470-6780
Finance and insurance	6870-6992
Real estate and rental and leasing	7071-7190
Professional, scientific, and technical services	7270-7490
Management of companies and enterprises	7570
Administrative and support and waste management services	7580-7790
Educational services	7860-7890
Health care and social assistance	7970-8470
Arts, entertainment, and recreation	8561-8590
Accommodation and food services	8660-8690
Other public services, except public administration	8770-9290
Public administration	9370-9590
Military	9670-9870

Research Question 28

Are the distributions of full-time, year-round workers by SOC major groups different between Test treatments and the Control treatment?

The Standard Occupational Classification (SOC) major groups are defined by the 2018 Standard Occupational Classification Manual. The 23 occupational groups and their corresponding census occupation codes are indicated in Table 22.

We will compare each pair of distributions of full-time year-round persons among the SOC major groups, Control versus each Test treatment, using a chi-square test. If the chi-square statistic is significant, we will also compare category proportion estimates, using the Holm-Bonferroni technique to control family-wise error (FWE).

Table 22. Census Occupation Codes Corresponding to SOC Major Groups

Description of SOC Major Group	Range of Census Occupation Codes
Management occupations	0010-0440
Business and financial operations occupations	0500-0960
Computer and mathematical occupations	1005-1240
Architecture and engineering occupations	1305-1560
Life, physical, and social science occupations	1600-1980
Community and social services occupations	2001-2060
Legal occupations	2100-2180
Education, training, and library occupations	2205-2555
Arts, design, entertainment, sports, and media occupations	2600-2970
Healthcare practitioner and technical occupations	3000-3550
Healthcare support occupations	3601-3655
Protective service occupations	3700-3960
Food preparation and serving related occupations	4000-4160
Building and grounds cleaning and maintenance occupations	4200-4255
Personal care and service occupations	4330-4655
Sales and related occupations	4700-4965
Office and administrative support occupations	5000-5940
Farming, fishing, and forestry occupations	6005-6130
Construction and extraction occupations	6200-6950
Installation, maintenance, and repair occupations	7000-7640
Production occupations	7700-8990
Transportation and material moving occupations	9005-9760
Military specific occupations	9800-9830

12.4 DECISION CRITERIA

The most important results of this analysis when drawing a conclusion about the Test versions compared to the Control version are shown below, in order of priority:

Table 23. Decision Criteria for Labor Force: Changing the Reference Period while also Modifying Question and Instructional Wording; Control vs. Test Treatments

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

Table 24. Decision Criteria for Labor Force: Paper Questionnaire Design—Test Version 1 vs. Test Version 2 mail responses

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

Table 25. Decision Criteria for Class of Worker and Industry & Occupation

Priority	Research Question	Decision Criteria
1	22, 23, 24, 25	Item missing data rates: We hope to see no difference or a decrease in item nonresponse.
2	26	Response Distributions: We hope to see no difference in distributions of full-time, year-round workers.
3	27, 28	Benchmark Estimate: We hope to see comparable estimates with CPS ASEC for full-time, year-round workers.

The above criteria will be used to decide on the "best" version of the Labor Force questions. However, before implementing any new version in production or before deciding to continue using the production version, we will also take into account the other topics that are related to this one: SNAP, Class of Worker, Industry and Occupation, and Income.

12.5   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.

Census Bureau. (2015). *Agility in action 1.0: A snapshot of enhancements to the American Community Survey.* Retrieved January 27, 2022, from https://www.census.gov/programs-surveys/acs/operations-and-administration/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. (Forthcoming). *Self-employment status: Imputations, implications, and improvements*. U.S. Census Bureau.

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

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

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

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

U.S. Census Bureau

13. INCOME

Authors: Kirby Posey, Gloria Guzman (SEHSD) and Dorothy Barth (DSSD)

The following section describes the specifics of the Income Series of questions in the 2022 ACS Content Test. Information in this section includes:

- A summary of the background and literature supporting the Test versions of the question (Section 13.1).
- A discussion of the changes between the Control and Test versions (Section 13.2).
- The specific research questions and methodology intended for analysis (Section 13.3).

13.1 LITERATURE REVIEW

Income questions in the ACS have always had a reference period of "the past 12 months." The "past 12 months" varies, depending on the month that a household responds to the survey. For example, questions asked in September have a reference period of October of the prior year to September of the current year. Similarly, questions asked in October have a reference period of November of the prior year to October of the current year. To better align with administrative data sources on many types of income and transfer benefits, we are testing a change in the reference period from "past 12 months" to a prior calendar year. The calendar year of reference for this test will be 2021, since the test will occur in 2022.

Survey methodologists have conducted studies showing that "sharpening the boundaries of a reference period" can improve recall and therefore the accuracy of reporting (Tourangeau, Rips and Rasinski, 2000). In this case, defining the reference period as the prior calendar year, instead of a sliding reference period of the past 12 months, has the goal of improving recall and accuracy of reporting. Additionally, using a frame of reference that matches the way that other methods capture similar data (e.g., filling out relevant tax forms) allows a respondent to use that previous recall exercise to improve their responses.

The Census Bureau is currently conducting research to determine the feasibility of using administrative sources to validate survey responses and to possibly serve as a replacement or supplement for income questions in its surveys (Bee and Rothbaum, 2019). Possible administrative data sources include the Longitudinal Employer-Household Dynamics (LEHD), the Payment History Update System (PHUS), and the Supplemental Security Records (SSR). The LEHD provides quarterly information on earnings that comes from each state's Unemployment Insurance System. The PHUS and SSR provide monthly information on Social Security Income and Supplemental Security Income that comes from the Social Security Administration. The Census Bureau has established data sharing agreements with states and the Social Security Administration to provide this data.

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, Robins, and Stapleton, 2017). Using the recommendations from the Westat report, RTI International further cognitively tested versions of the questions specifically for this Content Test (RTI International, 2021). The improvements made to the questions, as a result of cognitive testing, are outlined in Section 13.2 below.

13.2 QUESTION CONTENT

To be able to analyze the effect of changing the reference period and the effect of the other modifications made to the questions, we are testing two versions of the Income series of questions. Version 1 includes a change to the reference period, along with all of the modifications made to the questions as a result of cognitive testing (listed below). Version 2 is identical to production, except for a change to the reference period.

Aside from a new reference period, Version 1 has the following modifications:

(1) Pre-amble instructions on the paper questionnaire:
    a. Added the text, "Report all types of income received, taxable and non-taxable" before the new instructions for the reference period.
    b. Added break-even net income instructions.
    c. Capitalized the words "income received jointly" so that only those to whom the instructions apply will be alerted to read the instructions and others can skip them.
    d. Removed net loss of income instructions from the pre-amble and put them after the questions for self-employment and rental income.

(2) Modified question wording for the following sources of income:

    a. Self-employment (all modes): Added "including work paid for in cash" and put "farm or non-farm" in parentheses.
    b. Public assistance (all modes): Changed "any public assistance or welfare payments" to "any financial assistance or payments."
    c. Total income (all modes): Added "Including all types of income" to the beginning of the question.

(3) Separated rental income from the question about interest, dividends, royalty, estates, and trusts (paper and internet) [CAPI mode already has the question asked separately.]

(4) Modified instructions:

    a. Public Assistance (all modes): Added instructions about what types of income to exclude.

    b. Retirement income (paper and internet): Moved the instruction "Do NOT include Social Security" up; it now appears right after the question.

Figure 34. Control Version of the Income Questions (Paper)

Figure 35. Test Version 1 of the Income Questions (Paper)

Figure 36. Test Version 2 of the Income Questions (Paper)

INCOME IN 2021

Mark (X) the "Yes" box for each type of income this person received from January 1, 2021 to December 31, 2021. Give your best estimate of the TOTAL AMOUNT.

Mark (X) the "No" box to show types of income NOT received.

If net income was a loss, mark the "Loss" box to the right of the dollar amount.

For income received jointly, report the appropriate share for each person – or, if that's not possible, report the whole amount for only one person and mark the "No" box for the other person.

a. Wages, salary, commissions, bonuses, or tips from all jobs. *Report amount before deductions for taxes, bonds, dues, or other items.*

☐ Yes → $ ▯▯,▯▯▯,▯▯▯.00
☐ No   TOTAL AMOUNT for 2021

b. Self-employment income from own nonfarm businesses or farm businesses, including proprietorships and partnerships. *Report NET income after business expenses.*

☐ Yes → $ ▯,▯▯▯,▯▯▯.00   ☐
☐ No   TOTAL AMOUNT for 2021   Loss

c. Interest, dividends, net rental income, royalty income, or income from estates and trusts. *Report even small amounts credited to an account.*

☐ Yes → $ ▯,▯▯▯,▯▯▯.00   ☐
☐ No   TOTAL AMOUNT for 2021   Loss

d. Social Security or Railroad Retirement.

☐ Yes → $ ▯▯,▯▯▯.00
☐ No   TOTAL AMOUNT for 2021

e. Supplemental Security Income (SSI).

☐ Yes → $ ▯▯,▯▯▯.00
☐ No   TOTAL AMOUNT for 2021

f. Any public assistance or welfare payments from the state or local welfare office.

☐ Yes → $ ▯▯,▯▯▯.00
☐ No   TOTAL AMOUNT for 2021

g. Retirement income, pensions, survivor or disability income. *Include income from a previous employer or union, or any regular withdrawals or distributions from IRA, Roth IRA, 401(k), 403(b), or other accounts specifically designed for retirement. Do not include Social Security.*

☐ Yes → $ ▯▯▯,▯▯▯.00
☐ No   TOTAL AMOUNT for 2021

h. Any other sources of income received regularly such as Veterans' (VA) payments, unemployment compensation, child support or alimony. *Do NOT include lump sum payments such as money from an inheritance or the sale of a home.*

☐ Yes → $ ▯▯▯,▯▯▯.00
☐ No   TOTAL AMOUNT for 2021

What was this person's total income in 2021?
Add entries in questions 44a to 44h; subtract any losses. If net income was a loss, enter the amount and mark (X) the "Loss" box next to the dollar amount.

☐ OR $ ▯▯,▯▯▯,▯▯▯.00   ☐
None   TOTAL AMOUNT for 2021   Loss

13.3   RESEARCH QUESTIONS AND METHODOLOGY

The following section describes the specific analysis to be conducted on the Income questions in the 2022 ACS Content Test. Information in this section includes:

- Known benchmarks that the Content Test results will be compared to (Section 13.3.1).
- Specifics of the item missing data rate analysis (Section 13.3.2).
- Specifics of the response distribution analysis (Section 13.3.3).
- Specifics of the response reliability analysis (Section 13.3.4).
- Other analysis planned for this topic (Section 13.3.5).

13.3.1   Benchmarks for Income

We will compare proportions in each treatment with proportions from the 2021 Current Population Survey Annual Social and Economic Supplement (CPS ASEC). These comparisons will only be able to see whether Content Test estimates are reasonably close to the benchmark estimates; we will not do significance testing of differences.

Research Question 1

For each treatment, how do the proportions of persons (in the universe) who received self-employment, interest/dividends/royalty/estates/trusts, rental income, and public assistance compare with published CPS/ASEC data?

13.3.2   Item Missing Data Rates for Income

We will evaluate item nonresponse by calculating the item missing data rates for each mode of response separately and for all modes combined. The item missing data rate is the percent of all persons eligible to respond who did not provide a valid response to the question.

13.3.2.1   Self-Employment Income

Research Question 2

Version 1: For self-employment income, are the item missing data rates for recipiency or amount different for Version 1 than for Version 2?

Research Question 3

Version 2: For self-employment income, are the item missing data rates for recipiency or amount different for Version 2 than for the Control version?

The universe for self-employment income recipiency is all persons aged 16 or older that did not reply "No" to the question, "In 2021, did you work for pay, even for a few days?"

For mail and internet mode responses, if an amount is entered, we will consider the response to the recipiency question to be "Yes" (even if "No" is checked or if neither box is checked). Otherwise, a valid response will be "Yes", "No", or "Loss."

To calculate item missing data rates for the amount, the universe is all persons aged 16 or older for whom there is a "Yes" response to the recipiency question (or an implied "Yes" for a mail response). A valid amount response is zero, a positive amount, or a negative amount. "Don't Know" and "Refused" CAPI responses will considered nonresponses.

We will compare the recipiency and amount item missing data rates between treatments (Version 1 vs Version 2 and Version 2 vs. Control) using two-tailed t-tests to determine statistically significant differences.

13.3.2.2   Interest, Dividends, Royalty Income, Rental Income, and Income from Estates and Trusts

<u>Research Question 4</u>

Version 1: For interest, dividends, royalty income, rental income, and income from estates and trusts, are the item missing data rates for recipiency or amount different for Version 1 than for Version 2? (We will combine the two questions to compare Version 1 with Version 2.)

<u>Research Question 5</u>

Version 2: For interest, dividends, royalty income, rental income, and income from estate and trusts, are the item missing data rates for recipiency or amount different for Version 2 than for the Control version?

The universe for income recipiency from interest, dividends, royalty income, rental income, and income from estates and trusts is persons aged 15 or older. For mail and internet mode responses, if an amount is entered, we will consider the response to the recipiency question to be "Yes" (even if "No" was checked or if neither box is checked). Otherwise, a valid response will be "Yes", "No", or "Loss."

For question Version 1, since rental income is a separate question for mail and internet, a "Yes" to the recipiency question for either rental income OR interest, dividends, royalty income, and income from estates and trusts income will be counted as a "Yes" for aggregate recipiency.

To calculate item missing data rates for the amount, the universe is all persons aged 15 or older for whom there is a "Yes" response to the recipiency question (or an implied "Yes" for a mail response). A valid amount response is zero, a positive amount, or a negative amount (rental income). "Don't Know" and "Refused" CAPI responses will be considered nonresponses.

Attachment C

We will compare the recipiency and amount item missing data rates between treatments (Version 1 vs Version 2 and Version 2 vs. Control) using two-tailed t-tests to determine statistically significant differences.

We will also calculate and report the item missing data rates for rental income and interest, dividends, royalty income, and income from estates and trusts separately for Version 1 mail and internet response. This will be informational purposes only.

13.3.2.3   Public Assistance Income

<u>Research Question 6</u>

Version 1: For public assistance, are the item missing data rates for recipiency or amount different for Version 1 than for Version 2?

<u>Research Question 7</u>

Version 2: For public assistance, are the item missing data rates for recipiency or amount different for Version 2 than for the Control version?

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

To calculate item missing data rates for the amount, the universe is all persons aged 15 or older for whom there is a "Yes" response to the recipiency question (or an implied "Yes" for a mail response). A valid amount response is a number greater than zero. "Don't Know" and "Refused" CAPI responses will be considered nonresponses.

We will compare the recipiency and amount item missing data rates between treatments (Version 1 vs Version 2 and Version 2 vs. Control) using two-tailed t-tests to determine statistically significant differences.

13.3.2.4   Retirement and Pension Income

<u>Research Question 8</u>

Version 1: For retirement and pension, are the item missing data rates for recipiency or amount different for Version 1 than for Version 2?

For internet and CAI, there are two questions on retirement and pensions. We will combine the two questions for this analysis.

Research Question 9

Version 2: For retirement and pension, are the item missing data rates for recipiency or amount different for Version 2 than for the Control version?

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

To calculate item missing data rates for the amount, the universe is all persons aged 15 or older for whom there is a "Yes" response to the recipiency question (or an implied "Yes" for mail). A valid amount response is a number greater than zero. "Don't Know" and "Refused" CAPI responses will be considered nonresponses.

For internet and CAI, there are two questions on retirement and pensions. We will combine the two questions for this analysis.

We will compare the recipiency and amount item missing data rates between treatments (Version 1 vs Version 2 and Version 2 vs. Control) using two-tailed t-tests to determine statistically significant differences.

13.3.2.5   Total Income

Research Question 10

Version 1: For total income, are the item missing data rates different for Version 1 than for Version 2?

Research Question 11

Version 2: For total income, are the item missing data rates different for Version 2 than for the Control version?

To calculate item missing data rates for total income, the universe is all persons aged 15 or older. A nonresponse to this question is: the "None" box is not checked AND no amount is given. A valid amount response is zero, a positive amount, or a negative amount. "Don't Know" and "Refused" CAPI responses are also nonresponses.

We will compare total income item missing data rates between treatments (Version 1 vs Version 2 and Version 2 vs. Control) using two-tailed t-tests to determine statistically significant differences.

13.3.2.6   Income Series of Questions

<u>Research Question 12</u>

Version 1: Are the section missing data rates different for Version 1 than for Version 2?

<u>Research Question 13</u>

Version 2: Are the section missing data rates different for Version 2 than for the Control version?

We will calculate the rate at which respondents skip the entire section of Income questions, this includes all recipiency and amount questions and the total income question.

13.3.3   Response Distributions for Income

13.3.3.1   Self-employment Income

<u>Research Question 14</u>

Is the proportion of eligible persons reported as receiving self-employment income different for Version 1 than for Version 2?

<u>Research Question 15</u>

Is the proportion of eligible persons that reported a break-even amount of self-employment income different for Version 1 than for Version 2?

<u>Research Question 16</u>

Is the proportion of eligible persons that reported a loss for self-employment income different for Version 1 than for Version 2?

13.3.3.2   Interest, Dividends, Royalty, Rental Income, and Income from Estates and Trusts

<u>Research Question 17</u>

Is the proportion of eligible persons reported as receiving combined interest, dividends, royalty income, rental income, or income from estates and trusts different for Version 1 than for Version 2?

<u>Research Question 18</u>

Is the proportion of eligible persons that reported a break-even amount of rental income different for Version 1 than those that reported a break-even amount of combined interest, dividends, royalty income, rental income, or income from estates and trusts for Version 2?

Research Question 19

Is the proportion of eligible persons that reported a loss for rental income different for Version 1 than those that reported a loss for combined interest, dividends, royalty income, rental income, or income from estates and trusts for Version 2?

13.3.3.3   Public Assistance Income

Research Question 20

Is the proportion of eligible persons reported as receiving public assistance income different for Version 1 than for Version 2?

While we will calculate and compare the proportions of people reporting that they receive public assistance, we will not include the analysis in the decision criteria for accepting or rejecting a version of the question, since some of the changes made could increase reporting and some could decrease reporting. We have no preference or hypothesis about what will or should occur with this question.

13.3.3.4   Retirement or Pension Income

Research Question 21

Is the proportion of eligible persons reported as receiving retirement or pension income different for Version 1 than for Version 2?

13.3.4   Response Reliability for Income

Research Question 22

Is there a difference between treatments in response reliability for the following types of income recipiency: self-employment; combined interest, dividends, royalty and rental income; public assistance; and retirement income?

For self-employment, retirement, and rental income we will re-ask the original questions. For public assistance we will ask a series of questions from 2021 CPS-ASEC (questions below).

We will compare reliability metrics from Version 1 with Version 2.

Version 2 of the Income questions will not use response reliability in the decision criteria. The primary goal for Version 2 is to align the reference period of the questions more closely with administrative records. Since the only change to the questions is the reference period, the reliability of the responses will be determined by comparing Content Test data to administrative records and other known data sources.

Attachment C

13.3.5 Other Metrics for Income

Research Question 23

Is the aggregate amount of self-employment income different for Version 1 than for Version 2?

Research Question 24

Is the combined aggregate amount of interest, dividends, royalty income, rental income, and income from estates and trusts different for Version 1 than for Version 2?

Research Question 25

Is the aggregate amount of retirement and pension income different for Version 1 than for Version 2?

Research Question 26

How do the median earnings for all workers among the SOC major groups compare between treatments?

We will calculate median earnings for all types of workers by SOC major groups for each treatment. We will compare Version 1 to Version 2 and Version 2 to the control treatment. The occupation code will determine the SOC group.

Research Question 27

How do the median earnings for full-time year-round workers among the SOC major groups compare between treatments?

We will calculate median earnings for full-time year-round workers by SOC major groups for each treatment. We will compare Version 1 to Version 2 and Version 2 to the control treatment. The occupation code will determine the SOC group.

Research Question 28

How do recipiency and amounts for wages and salary income from Version 2 and control compare with Longitudinal Employer-Household Dynamics (LEHD) data?

We will compare LEHD data to responses for wages and salary earnings from Version 2 and the control treatment. This research requires:

- A link can be made between a respondent and the LEHD
- The individual's job is covered by the Unemployment Insurance system

U.S. Census Bureau

Research Question 29

How do recipiency and amounts for Social Security income from Version 2 and control compare with SSA data from the Payment History Update System (PHUS)?

We will compare PHUS data to responses for Social Security income from Version 2 and the control treatment. This research requires:

- We have access to the PHUS data in time
- A link can be made between the respondent and the PHUS

Research Question 30

How do recipiency and amounts for Supplemental Security Income (SSI) from Version 2 and control compare with SSA data from Supplemental Security Records (SSR)?

We will compare SSR data to responses for SSI income from Version 2 and the control treatment. This research requires:

- We have access to the SSR data in time
- A link can be made between the respondent and the SSR

13.4 DECISION CRITERIA

Table 26. Decision Criteria for Income: Wording Changes (Version 1 vs. Version 2)

Priority	Research Questions	Decision Criteria
1	14	Response Distributions: We hope to see no difference or an increase in the proportion of eligible persons receiving Self-Employment Income.
2	17	Response Distributions: We hope to see no difference or an increase in the proportion of eligible persons receiving combined Interest, Dividends, Royalty Income and Rental Income for the paper version.
3	21	Response Distributions: We hope to see no difference or a decrease in the proportion of eligible persons receiving Retirement, Survivor, and Disability Income.
4	2, 4, 6, 8, 10, 12	Item Missing Data Rates: We hope to see no difference or a decrease in item nonresponse for Self-Employment Income; Interest, Dividends, Royalty Income, Rental Income and Income from Estates and Trust;

		Public Assistance Income; and Retirement, Survivor, and Disability Income.
5	26,27	Other metrics: We hope to see similarities in median earnings for full-time year-round workers among the SOC major groups compared between treatments.
6	22	Response Reliability: We hope to see no difference or an increase in response reliability for Self-Employment Income, Interest, Dividends, Royalty Income, Rental Income and Income from Estates and Trust, Public Assistance Income, and Retirement, Survivor, and Disability Income.

Research questions not included in the decision criteria are for informational purposes only.

Table 27. Decision Criteria for Income: Changing the Reference Period (Version 2 vs. Control)

Priority	Research Question	Decision Criteria
1	28	We hope to see a difference in recipiency rate for wages and salary between version 2 and LEHD data that is smaller than the difference in recipiency rate for wages and salary between the control version and LEHD data.
2	29	We hope to see a difference in recipiency rate for Social Security between version 2 and SSA data that is smaller than the difference in recipiency rate for Social Security between the control version and SSA data.
3	30	We hope to see a difference in recipiency rate for SSI that is smaller between version 2 and SSA data than the difference in recipiency rate for SSI between the control version and SSA data.
4	28	We hope to see a difference in wage and salary amounts between version 2 and LEHD data that is smaller than the difference in wage and salary amounts between the control version and LEHD data.
5	29	We hope to see a difference in Social Security amounts between version 2 and SSA data that is smaller than the difference in Social Security amounts between the control version and SSA data.
6	30	We hope to see a difference in SSI amounts between version 2 and SSA data that is smaller than the difference in SSI amounts between the control version and SSA data.
7	3, 5, 7, 9, 11, 13	We hope to see no difference (or a decrease) in item missing data rates.

Research questions not included in the decision criteria are for informational purposes only.

The above criteria will be used to decide on the "best" version of the Income questions. However, before implementing any new version in production or before deciding to continue using the production version, we will also take into account the other topics that are related to this one: SNAP, Class of Worker, Industry and Occupation, and Labor Force.

13.5   REFERENCES

Bee, A., & Rothbaum, J. (2019). *The Administrative Income Statistics (AIS) Project: Research on the use of administrative records to improve income and resource estimates*. U.S. Census Bureau. Retrieved February 24, 2022, from https://www.census.gov/library/working-papers/2019/demo/SEHSD-WP2019-36.html

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

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

RTI International. (2021). Cognitive testing for the 2022 ACS Content Test: Round 1 and 2 combined briefing report.

14. ASSUMPTIONS AND LIMITATIONS

14.1   ASSUMPTIONS

1. The 2022 ACS Content Test sample is representative of the entire ACS sample frame, with respect to response rates.

2. Each treatment sample (1/3 of the full 2022 ACS Content Test sample) is representative of the entire ACS sample frame.

3. There is no difference between treatments in mail delivery timing or subsequent response time. The treatments will have the same sample size and use 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).

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

5. If there are differences in who we count (collect information on) because of changes to the Household Roster question, then we assume our proposed analytical measures will adjust for any effects seen for any of the questions on the Roster Test treatment questionnaire that would confound those questions' comparison to other treatments.

14.2   LIMITATIONS

1. GQs are not included in the sample for the test.

2. Sample housing unit addresses from Alaska, Hawaii, and Puerto Rico are not included in the sample for the test.

3. Interviews will be conducted in English and Spanish only. Respondents who need language assistance in another language will not be able to participate in the 2022 ACS Content Test. TQA responses are not included in the analysis.

4. CAPI interviewers will be 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 due to the same interviewer administering multiple versions of the same question item. Interviewers are trained to read the questions verbatim to minimize this risk, but there still exists the possibility that an

interviewer may deviate from the scripted wording of one question version to another. This could potentially mask a treatment effect from the data collected.

5. The CFU reinterview will be conducted by phone only (not in the same mode of data collection for the original interview). As a result, the data quality measures derived from the reinterview may include some bias due to the differences in mode of data collection.

15. POTENTIAL CHANGES TO THE ACS

We would expect revisions to make improvements to the data quality for the ACS. If the proposed changes are approved by OMB, major components of the ACS data collection process would require modification before production implementation (e.g., redesigned survey instruments, revised data processing specifications, and updated interviewer training materials).

16. REFERENCES

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

Dusch, G., & Meier, F. (2012). *2010 Census Content Reinterview survey evaluation report.* U.S. Census Bureau. Retrieved July 9, 2021, from https://www.census.gov/library/ publications/2012/dec/2010_cpex_206.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. Retrieved February 24, 2022, from https://www.census.gov/content/ dam/Census/library/working-papers/2016/acs/2016_Heimel_01.pdf

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

Keathley, D. (Forthcoming). *Specifications for selecting and weighting the 2022 American Community Survey Content Test sample*. U.S. Census Bureau. American Community Survey Memoradum Series.

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

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

RTI International (2021). *Cognitive Testing for the 2022 ACS Content Test: Round 2 briefing report.* Retrieved February 24, 2022, from Round Two - All Documents (census.gov) November 4, 2021.

RTI International (2022). *Cognitive Testing for the 2022 ACS Content Test: Round 3 Briefing Report,* forthcoming.

SAS Institute, Inc. (2009). *SAS/STAT(R) 9.2 User's Guide.* SAS Institute, Inc.

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 Memoradum Series.

Spiers, S. (2021b). *Coverage and nonresponse bias in the 2016 American Community Survey Content Test Follow-up Reinterview*. U.S. Census Bureau.

U.S. Census Bureau. (2013). *U.S. Census Bureau Statistical Quality Standards*." Retrieved January 27, 2022, from U.S. Census Bureau Statistical Quality Standards

U.S. Census Bureau. (2014b). *American Community Survey Design and Methodology.* Retrieved July 9, 2021, from https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html

Westfall, P., Tobias, R., & Wolfinger, R (2011). *Multiple comparisons and multiple tests using SAS® (2nd ed.)*. SAS Institute Inc.