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

September 7, 2017

2017 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES # ACS17-RER-05

MEMORANDUM FOR	Victoria Velkoff Chief, American Community Survey Office
From:	David Waddington Chief, Social, Economic, and Housing Statistics Division (SEHSD)
Prepared by:	Brian McKenzie Social, Economic, and Housing Statistics Division (SEHSD)
Subject:	2016 American Community Survey Content Test Evaluation Report: Journey to Work – Travel Mode of Commute and Time of Departure for Work

Attached is the final report for the 2016 American Community Survey (ACS) Content Test for Journey to Work. This report describes the results of the test for revised versions of the Mode of Commute and Time of Departure for Work questions.

If you have any questions about this report, please contact Alison Fields at 301-763-2456 or Brian McKenzie at 301-763-6532.

Attachment

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

Intentionally Blank

2016 American Community Survey Content Test Evaluation Report: Journey to Work – Travel Mode of Commute and Time of Departure for Work

FINAL REPORT

Brian McKenzie and Alison Fields, Social, Economic, and Housing Statistics Division
Michael Risley, Decennial Statistical Studies Division
R. Chase Sawyer, American Community Survey Office

Intentionally Blank

Table of Contents

List of Tables

List of Figures

EXECUTIVE SUMMARY

Overview

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

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

Journey to Work (Commuting)

This report discusses two journey to work, or commuting, questions that appear on the ACS: Commute Mode and Time of Departure for Work.

For the commuting questions on the ACS, this iteration of content testing is an attempt to clarify the meaning of existing question wording and maximize response. Both questions provide crucial information for transportation planning. Improving accuracy and decreasing missing data rates are important to maintaining their utility for the transportation planning community.

The proposed changes to Commute Mode are motivated by changes in public transportation infrastructure across the United States, particularly the increased prevalence of light rail systems and the need to update and clarify the terminology used to refer to commute modes that already appear as categories on the ACS. For this test, commute mode categories were modified to reflect the nation's actual public transportation options and the language used to describe them. For example, we added "light rail" as a commute mode category. See Section 1.3 for a comparison of the test and control questions.

The question about Time of Departure has historically raised concerns about privacy. An alternative treatment of this question was tested with the objective of phrasing the question in a

less intrusive way. While the control version of the question asks people when they leave home to go to work, the test version asks what time the person's trip to work began and does not include the word "home."

Research Questions and Results

Missing Data: Some research questions apply to both ACS commuting questions, while others are more specific to one of the two commuting questions. For both Commute Mode and Time of Departure, we tested whether the rate of missing data is lower for the control treatment than the test treatment. This test resulted in a failure to conclude that the control version had a statistically lower missing data rate than the test version for each of the commuting questions.

Response Distributions: We compared the response distributions of the test and control treatments of each commuting question based on categories used in published tables. There were no significant differences between the test and control treatments for the modified travel categories tested for Commute Mode. We also combined the rail-related categories into one category, finding no significant difference between the test and control treatments for this experimental combined category. The Commute Mode distribution was also examined for a small subset of metropolitan areas with high rates of transit usage. For this subsample of workers, as with the overall national sample, the distributions showed no significant difference between test and control versions.

For Time of Departure, responses were grouped into time intervals corresponding to those in published tables for easier comparison. Results show no statistically significant difference in Time of Departure distribution.

Response Reliability: We are interested in understanding how reliable a respondent's answer to each commuting question is, as measured by asking respondents the same question at two points in time. That is, we are interested in the respondent's likelihood of giving the same response for both the original interview and a follow-up interview. Of particular interest is whether reliability for the test treatment is higher than the control treatment. There was insufficient evidence to conclude that response reliability in the test treatment was higher than the control treatment for Commute Mode or Time of Departure.

Other Question-Specific Analyses: For Commute Mode, we compared the extent to which respondents incorrectly chose multiple commute modes on the paper questionnaire, rather than choosing the one for which they traveled the longest distance. There were no significant differences in incidences of multiple commute modes between test and control treatments. For Time of Departure, respondents often round their reported departure times to numbers ending in 0 or 5. We tested the rate at which this occurred for test and control treatments and found no significant difference between them.

Conclusions

The final wording in the test versions of the commuting questions is the product of consultation with industry experts and extensive cognitive testing. This new version of each ACS question is viewed as preferable to the current version. Among the various metrics used to answer our research questions, none revealed statistically different results between the test version and control version of each question. For both commuting questions tested, response distributions did not differ between test and control versions, which is consistent with the expectations. This suggests continuity in the meaning and interpretation of the control and test versions of each question. We recommend moving forward with the implementation of the new "test" version of each question, Commute Mode and Time of Departure. This recommendation is motivated by the goal of completeness in *existing* commute mode categories, particularly addressing the current absence of light rail, and reducing respondent burden. Cognitive testing also produced positive feedback for both of the commuting questions tested. Finally, consultation with industry experts representing the field of transportation planning and research resulted in unequivocal support for these category changes.

Intentionally Blank

1. BACKGROUND

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

The 2016 ACS Content Test included the following topics:

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

This report discusses Journey to Work: Commute Mode and Time of Departure for Work. For brevity, these commuting questions are referred to as Commute Mode and Time of Departure in this report.

1.1. Justification for Inclusion of Journey to Work in the Content Test

Commute Mode

A question collecting details of a person's mode of transportation to work was first introduced in the 1960 Census. The question wording and the transportation modes have changed over time to accommodate evolving transportation options and travel behavior. The 1960 version of Commute Mode included response options for automobile, bus, subway, walked, worked at home, and other means. Since then, several categories have been added or modified. For example, "Bicycle" was added as a separate category in 1980. "Streetcar" first appeared in 1970 and "Streetcar or trolley car" was presented for the first time in 1990. The current version of the question has been used since 1990.

Light rail is a transit mode that exists in over 30 metropolitan areas in the United States (American Public Transportation Association, 2014), and the transportation planning community

has argued that the current ACS rail-related questions should explicitly include light rail in the list of options. The Federal Transit Administration (FTA) funds most of the nation's fixed rail projects, but currently cannot directly measure commuting rates for light rail projects using ACS data in cities where light rail competes with other rail modes (such as subway). The addition of light rail to the categories will provide a crucial metric by which local and federal transit agencies can assess ridership of light rail systems as distinct from other modes.

Among national surveys that ask about means of transportation to work, no ongoing national survey explicitly includes light rail as a category and provides estimates for small areas. The American Housing Survey (AHS) asks how people get to work and school, including the type of public transportation used. Light rail is listed among the public transportation options, but the AHS is not an annual survey and does not provide estimates for areas smaller than large metropolitan areas. The National Household Travel Survey (NHTS) is a survey conducted by the U.S. Department of Transportation that also collects a person's mode of transportation to work. The last survey was conducted in 2009 and offered 24 distinct categories of transportation modes for the journey to work, plus an 'other' category (Federal Highway Administration, U.S. Department of Transportation, 2009). The rail-related categories available on that survey also did not include light rail. Those categories are Amtrak/intercity, commuter train, subway/elevated, and streetcar/trolley.

In addition to adding light rail to one of the existing commute mode categories, five of the existing twelve categories were modified in order to more closely reflect today's commute modes and how people refer to them. Details of question development and content are discussed in subsequent sections.

Time of Departure

A question collecting details of a person's time of departure from home to go to work was first introduced in the 1990 Census. The question has not changed since then. The ACS Content Review conducted in 2014 reported Time of Departure from Home to be one of the top questions that caused respondents discomfort and reluctance to answer, according to interviewers (Chappell & Obenski, 2014). However, the question is crucial for transportation planning efforts. The initial content testing goal for this question was to shift the focus of the question away from when a person leaves their home toward when they arrive at work. This new version was expected to alleviate privacy concerns for some respondents while still providing transportation planners with essential information about when commuters are on the road.

There are surveys other than the ACS that ask about time of departure from home for work, but none of these surveys are nationally representative or occur regularly. For example, the NHTS, which was last conducted in 2009, asked how many minutes it took to get from home to work and what time a person usually arrived at work (Federal Highway Administration, U.S. Department of Transportation, 2009). The NHTS is only conducted about every seven years and does not provide estimates at geographic levels smaller than the metro area. The Survey of Income and Program Participation (SIPP), a national survey, also collects such information, but has a much smaller sample size and does not provide estimates at small geographies.

1.2. Question Development

Initial versions of the new and revised questions were proposed by federal agencies participating in the U.S. Office of Management and Budget (OMB) Interagency Committee for the ACS. The initial proposals contained a justification for each change and described previous testing of the question wording, the expected impact of revisions to the time series and the single-year as well as five-year estimates, and the estimated net impact on respondent burden for the proposed revision.[1] For proposed new questions, the justification also described the need for the new data, whether federal law or regulation required the data for small areas or small population groups, if other data sources were currently available to provide the information (and why any alternate sources were insufficient), how policy needs or emerging data needs would be addressed through the new question, an explanation of why the data were needed with the geographic precision and frequency provided by the ACS, and whether other testing or production surveys had evaluated the use of the proposed questions.

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

The Census Bureau contracted with Westat to conduct three rounds of cognitive testing. The results of the first two rounds of cognitive testing informed decisions on specific revisions to the proposed content for the stateside Content Test (Stapleton and Steiger, 2015). In the first round, 208 cognitive interviews were conducted in English and Spanish and in two modes (self-administered on paper and interviewer-administered on paper). In the second round of testing, 120 cognitive interviews were conducted for one version of each of the tested questions, in English and Spanish, using the same modes as in the first round.

A third round of cognitive testing involved only the Puerto Rico Community Survey (PRCS) and Group Quarters (GQ) versions of the questionnaire (Steiger, Anderson, Folz, Leonard, & Stapleton, 2015). Cognitive interviews in Puerto Rico were conducted in Spanish; GQ cognitive interviews were conducted in English. The third round of cognitive testing was carried out to assess the revised versions of the questions in Spanish and identify any issues with questionnaire wording unique to Puerto Rico and GQ populations.[2] The proposed changes identified through cognitive testing for each question topic were reviewed by the Census Bureau, the corresponding topical subcommittee, and the Interagency Council on Statistical Policy Subcommittee for the ACS. The OMB then provided final overall approval of the proposed wording for field testing.[3]

---

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

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

[3] A cohabitation question and domestic partnership question were included in cognitive testing but ultimately we decided not to move forward with field testing these questions.

Commute Mode

The initial proposal for Commute Mode included a request to capture multiple commute modes, not just the one for which the respondent traveled the longest distance. For example, if a commuter traveled to work by driving their car to a train station, then taking a commuter train for the remainder of their trip, they would be able to select more than one commute mode.

The proposal to measure multiple commuting modes was rejected because it was viewed as an additional question, rather than a modification to an existing question. While the information would be valuable, execution of such an effort would present considerable operational burdens and conceptual challenges. For example, the addition of walking as a second mode presents challenges related to what defines a walking segment of a trip (across a parking lot, two blocks to a transit stop, etc.). Such departures from the straightforward 'longest distance' question format could present ambiguities for respondents similar to those currently suspected to affect the set of rail options.

The final question proposal included a single set of modified commute mode categories; all related to public transportation modes. The category "Streetcar or trolley car" was changed to "Light rail, street car, or trolley;' "Subway or elevated" was changed to "Subway or Elevated Rail;" "Railroad" was changed to "Long-distance train or commuter rail." These three rail-related categories were also slightly reordered so that "Subway or elevated rail," the most prevalent rail mode, is listed first. Finally, for these three rail-related categories, the subcommittee discussed including the word "Rail" at the beginning of each (see below). After considerable discussion, this idea was rejected.

__ Rail: light rail, streetcar, or trolley

__ Rail: subway or elevated

__ Rail: commuter or long-distance railroad

The subcommittee discussed moving the "Worked at home" category to the beginning of the list so that workers who work at home could immediately skip to the next set of questions rather than read the entire list of commute modes. The subcommittee eventually decided against this because it only affects about 4 percent of workers.

The first round of cognitive testing resulted in three additional changes to the commute mode categories (Stapleton & Steiger, 2015). The phrase "trolley bus" was dropped from the test version and the phrase "worked at home" was changed to "worked from home." The category "Commuter or long distance railroad" was changed to "Commuter rail or long distance train" to add clarity. The subject matter group unanimously agreed to make these changes. The second round of testing resulted in a minor change of one category. The category "Commuter rail or long-distance train" was changed to "Long-distance train or commuter rail." The subheading of instructions was also modified:

From:

How did this person usually get to work LAST WEEK? If this person usually used more than one method of transportation during the trip, mark (X) the box of the one used for most of the distance.

To:

How did this person usually get to work LAST WEEK? Mark ONE box for the method of transportation used for most of the distance.

This change was made to simplify instructions and remove any ambiguity associated with the (X). It is crucial that respondents choose only one commute mode because choosing more than one results in the case being allocated.

Time of Departure

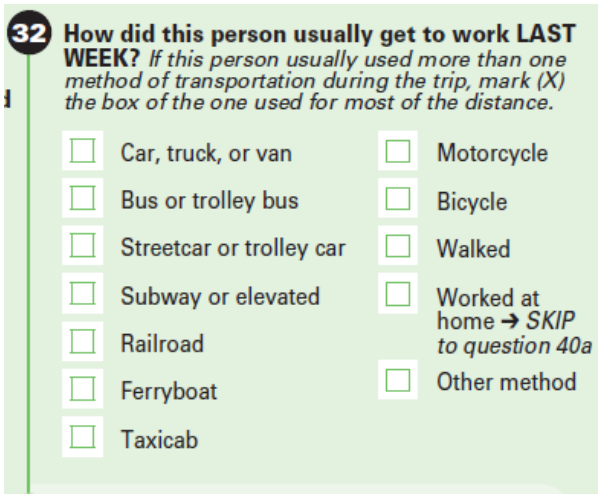
ACS respondents are currently asked, "What time did this person leave home to go to work last week?" For the first round of testing, the subcommittee modified the question to focus on the end of the commuter's trip, when they arrived at work, rather than the beginning, which presumes an initial departure from home. Respondents involved in cognitive testing reported no difference in sensitivity between the test and control versions of this question. Responses to the test question were less accurate and highly rounded compared with the control version. Respondents had difficultly estimating the time of arrival at work for other members of the household. Respondents have a reasonable approximation for when other workers within the household departed for work, but were more likely to give highly rounded arrival times. The time of arrival question also resulted in confusion among respondents about the exact point at which they "arrive" at work. For example, they wondered whether the question referred to when they enter the premises or are situated at their workstation.

For the second round of testing, the subcommittee revisited the approach of retaining a focus on the beginning of the commuter's trip, but removed the word "home" to generalize the question. The subcommittee finally decided on asking, "Last week, what time did this person's trip to work usually begin?" This took the emphasis off the sensitive word "home," while still gathering information on the beginning of the work trip. Cognitive testing found that respondents answered this version of the question more accurately than the one that focused on time of arrival to work, especially for other members of the household. The "heaping" (clustering around numbers ending in 0 and 5) associated with this version was comparable to the control version.

1.3. Question Content

Figure 1. Control and Test Versions of Commute Mode Question

Control Version



Test Version

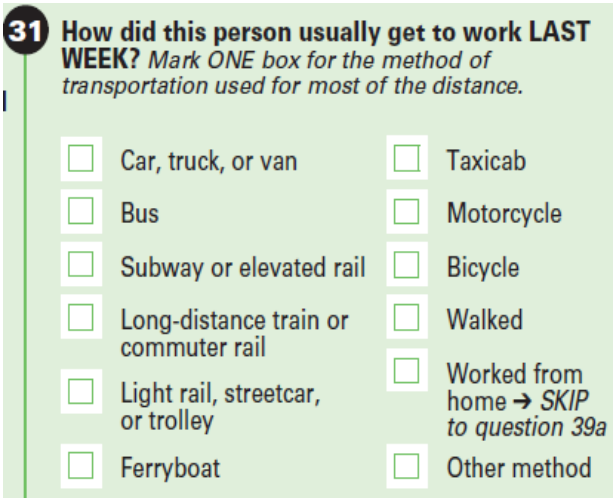
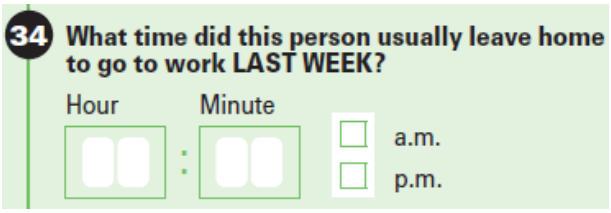
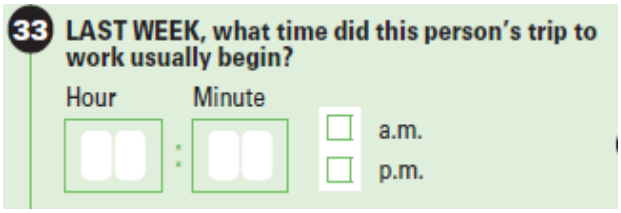


Figure 2. Control and Test Versions of Time of Departure Question

Control Version



Test Version



1.4. Research Questions

The following research questions were formulated to guide the analyses of the Commute Mode and Time of Departure for Work questions. The analyses assessed how the test versions of the questions performed compared to the control versions in the following ways: how often the respondents answered the questions, the consistency and accuracy of the responses, and how the responses affected the resulting estimates.

Commute Mode

1. *Is the missing data rate the same or lower for the test treatment than for the control treatment?*

2. *How do the test and control response distributions compare at the national level? This will be compared first using all 12 categories on the ACS questionnaire. These 12 categories are included in American Factfinder (AFF) Table B08301. The distributions*

from the ten categories in AFF Table B08006 and six categories in AFF ACS Table S0801 will then be shown.

3. *How does the proportion of respondents marking one of the three rail categories compare between test and control versions?*

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

5. *For the paper questionnaire, is the proportion of person records for which respondents incorrectly marked multiple modes of transportation comparable between control and test versions? When multiple modes are marked, if the sample size is large enough, which combinations are most common in each version? Note that respondents are instructed to mark only one commute mode.*

6. *How do the test and control response distributions compare in metro areas with high levels of light rail usage?*

7. *How do the test and control response distributions compare when the sample is restricted to only metro areas with high levels of overall rail usage?*

Time of Departure

8. *Is the missing data rate the same or lower for the test treatment than for the control treatment?*

9. *Using the categories defined in AFF, ACS Table B08302, are the distributions comparable between the test and control questionnaires?*

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

11. *Is the proportion of respondents who leave home at a time that ends in 0 or 5 comparable between test and control versions?*

2. METHODOLOGY

2.1. Sample Design

The 2016 ACS Content Test consisted of a nationally representative sample of 70,000 residential addresses in the United States, independent of the production ACS sample. The Content Test sample universe did not include GQs, nor did it include housing units in Alaska, Hawaii, or Puerto Rico.[4] The sample design for the Content Test was largely based on the ACS production

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

sample design with some modifications to better meet the test objectives.[5] The modifications included adding an additional level of stratification by stratifying addresses into high and low self-response areas, oversampling addresses from low self-response areas to ensure equal response from both strata, and sampling units as pairs.[6] The high and low self-response strata were defined based on ACS self-response rates at the tract level. Sampled pairs were formed by first systematically sampling an address within the defined sampling stratum and then pairing that address with the address listed next in the geographically sorted list. Note that the pair was likely not neighboring addresses. One member of the pair was randomly assigned to receive the control version of the question and the other member was assigned to receive the test version of the question, thus resulting in a sample of 35,000 control cases and 35,000 test cases.

As in the production ACS, if efforts to obtain a response by mail or telephone were unsuccessful, attempts were made to interview in person a sample of the remaining nonresponding addresses (see Section 2.2 Data Collection for more details). Addresses were sampled at a rate of 1-in-3, with some exceptions that were sampled at a higher rate.[7] For the Content Test, the development of workload estimates for the Computer-Assisted Telephone Interviews (CATI) and Computer-Assisted Personal Interviews (CAPI) did not take into account the oversampling of low response areas. This oversampling resulted in a higher than expected workload for CATI and CAPI and therefore required more budget than was allocated. To address this issue, the CAPI sampling rate for the Content Test was adjusted to meet the budget constraint.

2.2. Data Collection

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

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

---

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

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

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

The ACS Content Test's data collection protocol differed from the production ACS in a few significant ways. The Content Test analysis did not include data collected via the Telephone Questionnaire Assistance (TQA) program since those who responded via TQA used the ACS production TQA instrument. The Content Test excluded the telephone Failed Edit Follow-Up (FEFU) operation.[8] Furthermore, the Content Test had an additional telephone reinterview operation used to measure response reliability. We refer to this telephone reinterview component as the Content Follow-Up, or CFU. The CFU is described in more detail in Section 2.3.

ACS production provides Spanish-language versions of the internet, CATI, and CAPI instruments, and callers to the TQA number can request to respond in Spanish, Russian, Vietnamese, Korean, or Chinese. The Content Test had Spanish-language automated instruments; however, there were no paper versions of the Content Test questionnaires in Spanish.[9] Any case in the Content Test sample that completed a Spanish-language internet, CATI, or CAPI response was included in analysis. However, if a case sampled for the Content Test called TQA to complete an interview in Spanish or any other language, the production interview was conducted and the response was excluded from the Content Test analysis. This was due to the low volume of non-English language cases and the operational complexity of translating and implementing several language instruments for the Content Test. CFU interviews for the Content Test were conducted in either Spanish or English. The practical need to limit the language response options for Content Test respondents is a limitation to the research, as some respondents self-selected out of the test.

2.3. Content Follow-Up

For housing units that completed the original interview, a CFU telephone reinterview was also conducted to measure response error.[10] A comparison of the original interview responses and the CFU reinterview responses was used to answer research questions about response error and response reliability.

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

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

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

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

the household member who originally responded, but if that was not possible, the CFU reinterview was conducted with any other eligible household member (15 years or older).

The CFU asked basic demographic questions and a subset of housing and detailed person questions that included all of the topics being tested, with the exception of Telephone Service, and any questions necessary for context and interview flow to set up the questions being tested.[11] All CFU questions were asked in the reinterview, regardless of whether or not a particular question was answered in the original interview. Because the CFU interview was conducted via telephone, the wording of the questions in CFU followed the same format as the CATI nonresponse interviews. Housing units assigned to the control version of the questions in the original interview were asked the control version of the questions in CFU; housing units assigned to the test version of the questions in the original interview were asked the test version of the questions in CFU. The only exception was for retirement, survivor, and disability income, for which a different set of questions was asked in CFU.[12]

2.4. Analysis Metrics

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

All Content Test data were analyzed without imputation due to our interest in how question changes or differences between versions of new questions affected "raw" responses, not the final edited variables. Some editing of responses was done for analysis purposes, such as collapsing response categories or modes together or calculating a person's age based on his or her date of birth.

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

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

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

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

2.4.1. Unit Response Rates and Demographic Profile of Responding Households

The unit response rate is generally defined as the proportion of sample addresses eligible to respond that provided a complete or sufficient partial response.[13] Unit response rates from the original interview are an important measure to look at when considering the analyses in this report that compare responses between the control and test versions of the survey questionnaire. High unit response rates are important in mitigating potential nonresponse bias.

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

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

We also calculated the CFU interview unit response rate overall and by mode of data collection of the original interview and compared the control and test treatments because response error analysis (discussed in Section 2.4.5) relies upon CFU interview data. Statistical differences between CFU response rates for control and test treatments will not be taken as evidence that one version is better than the other. For the CFU response rates, the universe for each mode consists of housing units that responded to the original questionnaire in the given mode (internet, mail, CATI, or CAPI) and were eligible for the CFU interview. We expected the response rates to be similar between treatments; however, we calculated the rates to verify that assumption.
Another important measure to look at in comparing experimental treatments is the demographic profile of the responding households in each treatment. The Content Test sample was designed with the intention of having respondents in both control and test treatments exhibit similar distributions of socioeconomic and demographic characteristics. Similar distributions allow us to compare the treatments and conclude that any differences are due to the experimental treatment instead of underlying demographic differences. Thus, we analyzed distributions for data from the

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

following response categories: *age, sex, educational attainment,* and *tenure.* The topics of *race, Hispanic origin,* and *relationship* are also typically used for demographic analysis; however, those questions were modified as part of the Content Test, so we could not include them in the demographic profile. Additionally, we calculated *average household size* and the *language of response* for the original interview.[14]

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

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

2.4.2. Item Missing Data Rates

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

Commute Mode

The percent of eligible persons who did not provide a response to this question in the control treatment is compared to the corresponding percent from the test treatment. Statistical significance between versions is determined using a one-tailed t-test. Note that for the purposes of this analysis, we count mail mode responses where multiple (two or more) categories are selected (checked) as missing responses. A research question specifically addressing multiple-category responses in the Mail mode is included in the Other Analysis Section 2.4.6.
We expected that the test treatment would not have a missing data rate that is the same as or lower than the control treatment. While the categories that a respondent chooses may vary across test and control versions of the survey, each survey includes an "Other" category that should be a last resort for a respondent who is confused about which commute mode category to choose.

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

Time of Departure

The percent of eligible persons who did not provide a response to this question in the control treatment is compared to the corresponding percent from the test treatment. Statistical significance between versions is determined using a one-tailed t-test. Note that for the purposes of this analysis, we count mail or internet mode responses as missing when any of the three parts (hour, minute, am/pm) are missing.

2.4.3. Response Distributions

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

Proportion estimates for *Commute Mode* were calculated as:

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

For the *Time of Departure* question, we defined ranges of valid responses, which are described below.

Commute Mode

For Commute Mode, distributions are compared first using all 12 categories on the questionnaire, then using a 10-category collapse from American Factfinder (AFF) Table B08006, and finally using six categories, as found in AFF Table S0801 (see Figure 1. in Section 1.3 for the categories). The most detailed category schema involves 12 commute mode categories. The control version of these categories is shown below. One anticipated finding was a smaller proportion of cases in the "Other" category in the test version because respondents might now find clarity for commute modes that were previously unclear. Individual rail-related commute modes may also show small differences due to increased clarity in category names. For example, a respondent who commuted by light rail may have previously chosen Subway or Elevated in the absence of a category that specifically includes light rail. We compare each pair of distributions (control versus test) using a Chi-squared test. Several combinations of collapsed commute mode categories are tested. A t-test is also used to test the proportion of the three rail-related commuting categories combined. Figure 3 shows the commute mode categories included in each distribution.

Figure 3. Commute Mode Categories (control version) Used in Analysis

12 Categories	10 Categories	6 Categories
Bicycle	Bicycle	Bicycle
Bus	Bus	Car, truck or van
Car, truck or van	Car, truck or van	Public Transportation
Ferryboat	Ferryboat	Taxi, Motorcycle or Other Method
Motorcycle	Railroad	Walked
Other Method	Streetcar	Worked at Home
Railroad	Subway	
Streetcar	Taxi, Motorcycle or Other Method	
Subway	Walked	
Taxicab	Worked at Home	
Walked		
Worked at Home		

Time of Departure

Since this question is presented in an open-ended write-in format, answers are grouped into intervals for easier comparison. We compare each pair of distributions (control versus test) using a Chi-squared test. The time of departure intervals are:

12:00 a.m. to 4:59 a.m.	8:00 a.m. to 8:29 a.m.
5:00 a.m. to 5:29 a.m.	8:30 a.m. to 8:59 a.m.
5:30 a.m. to 5:59 a.m.	9:00 a.m. to 9:59 a.m.
6:00 a.m. to 6:29 a.m.	10:00 a.m. to 10:59 a.m.
6:30 a.m. to 6:59 a.m.	11:00 a.m. to 11:59 a.m.
7:00 a.m. to 7:29 a.m.	12:00 p.m. to 3:59 p.m.
7:30 a.m. to 7:59 a.m.	4:00 p.m. to 11:59 p.m.

2.4.4. Benchmarks

No other surveys collect directly comparable data to use as a benchmark in this analysis. The National Household Travel Survey (NHTS), a survey conducted by the U.S. Department of Transportation, uses a travel diary method to collect information about travel patterns in the United States. This survey generally serves as a useful comparison for ACS commuting estimates such as travel mode distribution and travel time. For the purpose of this content test,

the most appropriate benchmark is to compare responses from the test treatment to the current production questions, as done in the comparisons provided throughout section 5.

2.4.5. Response Error

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

We measured simple response variance by comparing valid responses to the CFU reinterview with valid responses to the corresponding original interview.[15] The Census Bureau has frequently used content reinterview surveys to measure simple response variance for large demographic data collection efforts, including the 2010 ACS Content Test, and the 1990, 2000, and 2010 decennial censuses (Dusch & Meier, 2012).

The following measures were used to evaluate consistency:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The measures of response error assume that those characteristics in question did not change between the original interview and the CFU interview. To the extent that this assumption is incorrect, we assume that it is incorrect at similar rates between the control and test treatments.

2.4.6. Other Analysis Methodology Specific to Commuting Questions

Commute Mode

We are especially interested in analysis for cities with a diverse set of transit options, specifically rail options. It is in these places that we expect to see meaningful differences in the three rail categories. We looked at metro areas with high rates of rail usage, as defined by the American Public Transportation Association 2014 Transit Ridership Report (American Public Transportation Association, 2014). Test responses from these metro areas are combined and compared against the combined control responses from these areas.

Figure 4. Selected Cities Included in Targeted Rail Metro Analyses

Cities with High Levels of Overall Rail Ridership	Cities with High Levels of Light Rail Usage
New York, NY	Boston, MA
Washington, DC	Los Angeles, CA
Chicago, IL	San Francisco, CA
Boston, MA	San Diego, CA
San Francisco, CA	Portland, OR
Philadelphia, PA	Philadelphia, PA
Atlanta, GA	Dallas, TX
Los Angeles, CA	Denver, CO
Miami, FL	Salt Lake City, UT
Baltimore, MD	St. Louis, MO

An additional test for Commute Mode focused on the extent to which respondents who received a paper questionnaire incorrectly marked more than one commute mode. The question instructs

respondents to choose the single commute mode for which the longest distance was traveled, but a small percentage of respondents invariably choose more than one. For standard ACS processing, these cases are ultimately allocated. The frequency of choosing multiple modes was compared for the test and control treatments using different travel mode category pairs.

Time of Departure

There is a tendency for respondents to answer this question with a time that is rounded to a time ending in 0 or 5 (Stapleton & Steiger, 2015). If the test version of this question produces fewer responses 'heaped' on times ending in 0 or 5, the test version might be providing estimates that are more precise. This analysis is conducted using a two-tailed t-test.

2.4.7. Standard Error Calculations

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

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

where:

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

3. DECISION CRITERIA

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

Table 2. Decision Criteria for Commute Mode

Research Questions	Decision Criteria, in order of priority
3	The test version should have the same or higher rate of responses commuting by rail than the control version.
2, 6 and 7	Differences in the distribution of commute mode categories should be minimal between test and control versions.
1	The item missing data rates for the test version should be the same or lower than the control version.
4	The reliability for the test version should be higher than the control version.
5	The proportion of person records that mark multiple modes should be comparable between the control and test versions.

Table 3. Decision Criteria for Time of Departure

Research Questions	Decision Criteria, in order of priority
8	The item missing data rates for the test version should be the same or lower than the control version.
11	The proportion of responses in the test version that appear to be rounded should be the same or lower than in the control version.
10	The reliability for the test version should be higher than the control version.
9	The distributions between the control and test versions should have minimal to no differences.

4. LIMITATIONS

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

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

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

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

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

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

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

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

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

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

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

For Commute Mode, testing categories involving rail is challenging because rail-related travel infrastructure only exists in a small percentage of U.S. cities. Testing a commuting category that only applies to a small sample of the working population requires a relatively large sample size to obtain margins of error large enough to produce significant differences between test and control treatments. Small proportional differences also require relatively large samples in order to register as significantly different.

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

5. RESEARCH QUESTIONS AND RESULTS

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

5.1. Unit Response Rates and Demographic Profile of Responding Households

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

5.1.1. Unit Response Rates for the Original Content Test Interview

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

Table 4 shows the unit response rates for the original interview for each mode of data collection (internet, mail, CATI, and CAPI), all modes combined, and both self-response modes (internet and mail combined) for the control and test treatments. When looking at the overall unit response rate (all modes combined) the difference between control (93.5 percent) and test (93.5 percent) is less than 0.1 percentage points and is not statistically significant.

Table 4. Original Interview Unit Response Rates for Control and Test Treatments, Overall and by Mode

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

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

When analyzing the unit response rates by mode of data collection, the only modal comparison that shows a statistically significant difference is the mail response rate. The control treatment

had a higher mail response (19.6 percent) than the test treatment (18.4 percent) by 1.2 percentage points. As a result of this difference, we looked at how mail responses differed in the high and low response areas. Table 5 shows the mail response rates for both treatments in high and low response areas.[17] The difference in mail response rates appears to be driven by the difference of rates in the high response areas.

It is possible that the difference in the mail response rates between control and test is related to the content changes made to the test questions. There are some test questions that could be perceived as being too sensitive by some respondents (such as the test question relating to same-sex relationships) and some test questions that could be perceived to be too burdensome by some respondents (such as the new race questions with added race categories). In the automated modes (internet, CATI, and CAPI) there is a higher likelihood of obtaining a sufficient partial response (obtaining enough information to be deemed a response for calculations before the respondent stops answering questions) than in the mail mode. If a respondent is offended by the questionnaire or feels that the questions are too burdensome they may just throw the questionnaire away, and not respond by mail. This could be a possible explanation for the unit response rate being lower for test than control in the mail mode.

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

Table 5. Mail Response Rates by Designated High (HRA) and Low (LRA) Response Areas

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

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

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

Table 6 shows the unit response rates for the CFU interview by mode of data collection of the original interview and for all modes combined, for control and test treatments. Overall, the differences in CFU response rates between the treatments are not statistically significant. The rate at which CAPI respondents from the original interview responded to the CFU interview is lower for test (34.8 percent) than for control (37.7 percent) by 2.9 percentage points. While the protocols for conducting CAPI and CFU were the same between the test and control treatments, we could not account for personal interactions that occur in these modes between the respondent

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

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

Table 6. Content Follow-Up Interview Unit Response Rates for Control and Test Treatments, Overall and by Mode of Original Interview

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

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

5.1.3. Demographic and Socioeconomic Profile of Responding Households

One of the underlying assumptions of our analyses in this report is that the sample for the Content Test was selected in such a way that responses from both treatments would be comparable. We did not expect the demographics of the responding households for control and test treatments to differ. To test this assumption, we calculated distributions for respondent data for the following response categories: *age, sex, educational attainment,* and *tenure.*[18] The response distribution calculations can be found in Table 7. Items with missing data were not included in the calculations. After adjusting for multiple comparisons, none of the differences in the categorical response distributions shown below is statistically significant.

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

Table 7. Response Distributions – Test versus Control Treatment

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

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

We also analyzed two other demographic characteristics shown by the responses from the survey: *average household size* and *language of response.* The results for the remaining demographic analyses can be found in Table 8 and Table 9.

Table 8. Comparison of Average Household Size

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

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

Table 9. Comparison of Language of Response

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

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

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

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

5.2. Item Missing Data Rates

This section addresses research question number 1: *Is the missing data rate the same or lower for the test treatment than for the control treatment?*

Table 10 shows the item missing data rates for the control and test versions of each of the two commuting questions. The universe for Commute Mode is all workers aged 16 and older who were in the workforce during the reference week. The universe for the time leaving home question is the same except that it does not include workers who worked at home.

For Commute Mode, the p-value was not significant, indicating that there was insufficient evidence to conclude that the response item-missing data rate for the test treatment is higher than that of the control treatment. This is consistent with expectations. While the categories that a respondent chooses may vary across test and control versions of the survey, each survey includes an "Other" category that should be a last resort for a respondent who is confused about which commute mode category to choose.

The item missing data rate for the test treatment of Time of Departure is not statistically higher than that of the control version, at 10.8 and 10.9 percent, respectively. A lower missing data rate for the test version would be a more favorable outcome, but evidence that the item missing data is not higher for test is consistent with expectations. For this question, the number of people who have expressed concern about privacy is small, so a larger sample may be needed to yield significant differences in response rates.

Table 10. Item Missing Data Rates for Control and Test Treatments – Commute Mode and Time of Departure

Item	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	P-Value (one-tailed)
Commute Mode	17,739	1.4 (0.1)	17,951	1.5 (0.1)	>-0.1 (0.2)	0.69
Time Leaving Home	16,631	10.8 (0.4)	16,820	10.9 (0.4)	-0.1 (0.6)	0.59

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

5.3. Response Distributions

This section addresses research questions number 2 and number 9: *How do the test and control response distributions compare?* Research question number 3 is also addressed: *How does the proportion of respondents marking one of the three rail categories compare between test and control versions when all three categories are combined?*

Commute Mode

For Commute Mode, we compared each pair of distributions (control versus test) using a Chi-squared test. Table 11 shows the results of the Rao-Scott Chi-squared statistic for each. The results revealed no statistically significant difference between the test and control treatments for any of the standard travel mode category distributions (12 categories, 10 categories, and 6 categories). This result was not surprising given that our goal was limited to refining and clarifying existing categories to prevent ambiguity and keep up with the changing public transportation landscape. Additionally, the number of commuters who are expected to choose a public transportation mode is relatively small. The lack of significant differences served as an indicator that our refined and more inclusive wording for public transportation categories would not undermine comparability across years.

Table 11. Commute Mode: Chi-Square Statistic Comparing Control and Test Treatment

Category	Rao-Scott Chi-Square Statistic	P-value
12 Category Distribution	8.5	0.67
10 Category Distribution	7.0	0.64
6 Category Distribution	4.2	0.52

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: Significance testing was done at the α=0.1 level based on a chi-square test.

Table 12 shows the distribution of the 12-category version of Commute Mode. The control and test distributions were not significantly different from one another. While we did not anticipate differences in distributions, our expectation was that any potential difference between test and control versions would involve one of the public transportation categories or a reduction in the number of people who choose the "Other" category in the test version, due to increased clarity of public transportation categories. For example, a respondent who commuted by light rail may

have previously chosen "Subway" or "Elevated" in the absence of a category that specifically includes light rail. No such differences were detected.

Table 12. Response Distribution for Control and Test Treatment for Commute Mode

12 Category Distribution (simplified category names)	Test Percent (n=17,429)	Control Percent (n = 17,604)
Bicycle	0.6 (0.1)	0.5 (0.1)
Bus	2.0 (0.1)	1.9 (0.1)
Car, truck or van	86.2 (0.4)	86.5 (0.4)
Ferryboat	<0.1 (<0.1)	<0.1 (<0.1)
Motorcycle	0.1 (<0.1)	0.1 (<0.1)
Other Method	1.0 (0.1)	0.7 (0.1)
Railroad	0.7 (0.1)	0.7 (0.1)
Streetcar	0.1 (<0.1)	0.1 (<0.1)
Subway	1.7 (0.2)	1.5 (0.1)
Taxicab	0.1 (<0.1)	0.2 (<0.1)
Walked	2.3 (0.1)	2.6 (0.2)
Worked at Home	5.2 (0.2)	5.1 (0.3)
Total	100.0	100.0

Source: U.S. Census Bureau, 2016 American Community Survey Content Test
Note: $\chi 2 = 8.5$, p-value=0.67. Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance testing was done at the α=0.1 level, based on a chi-square test.

To identify overall differences in the public transportation categories, we focus on the three rail-related categories, which were modified to improve clarity and reduce redundancy of all public transportation categories. For testing purposes only, we created a combined category of all three rail-related commute modes and then assessed their prevalence between the test and control treatments. The combined rail categories were not significantly different across treatments (Table 13). This is consistent with the expectation that there will be little to no difference between commute mode distributions across treatments.

Table 13. Proportion of Three Rail-Related Commute Mode Categories Combined

Category	Test Sample Size	Test Percent	Control Sample Size	Control Percent	Test minus Control	P-Value
Combined Rail	17,429	2.5 (0.2)	17,604	2.3 (0.2)	-0.3 (0.2)	0.28

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

Responses for Commute Mode were broken down by survey mode (internet, mail, or "interview-assisted modes" which includes CATI and CAPI). Among the three survey methods, test and control distributions were compared for the six commute mode categories using a Chi-square test. The collapsed six-category distribution is used in this test in order to obtain a sufficiently

large sample for each survey mode. For mail and interviewed modes, the distributions of the test and control treatments were not statistically different from one another, but the test and control distributions for internet mode were statistically different with a p-value of <0.10 for the overall distribution. Table 14 shows the difference between control and test in the distribution of individual commute modes for respondents who responded by internet.

Table 14. Proportion of Commute Mode for Test and Control Treatments – Internet Response Mode

Commute Mode	Test Percent	Control Percent	Test minus Control	Adjusted P-Value
Bicycle	0.6 (0.1)	0.6 (0.1)	<0.01 (0.1)	0.96
Car, truck or van	85.9 (0.4)	87.2 (0.4)	-1.4 (0.7)	0.23
Public Transportation	4.8 (0.3)	4.3 (0.3)	0.4 (0.4)	0.88
Taxi, Motorcycle or Other Method	1.1 (0.2)	0.7 (0.1)	0.3 (0.2)	0.37
Walked	2.1 (0.2)	2.3 (0.2)	-0.2 (0.2)	0.89
Worked at Home	5.6 (0.3)	4.8 (0.3)	0.8 (0.4)	0.37

Source: U.S. Census Bureau, 2016 American Community Survey Content Test. Significance was tested based on a two-tailed t-test at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. $\chi 2 = 9.4$, p-value<0.10.

Time of Departure

Table 15 shows the distributions for the test and control versions of Time of Departure. The Rao-Scott Chi-square test found no statistical difference between the two distributions of Time of Departure. The primary goal is to reduce the sensitivity of the question, increase response rates, and retain the current distribution. This is an acceptable outcome based on expectations and decision criteria. Note that since there was no significant difference in the distributions, tests were not done on the individual time of departure categories.

Table 15. Response Distribution for Control and Test Treatment for Time of Departure

Departure Time Categories	Test Percent (N=14,729)	Control Percent (N=14,973)
12:00 am to 4:59 am	4.4 (0.2)	4.4 (0.3)
5:00 am to 5:29 am	3.7 (0.2)	3.5 (0.2)
5:30 am to 5:59 am	4.6 (0.3)	4.4 (0.2)
6:00 am to 6:29 am	9.0 (0.3)	8.3 (0.4)
6:30 am to 6:59 am	10.0 (0.4)	10.0 (0.4)
7:00 am to 7:29 am	14.6 (0.3)	15.2 (0.4)
7:30 am to 7:59 am	12.7 (0.4)	12.4 (0.4)
8:00 am to 8:29 am	11.7 (0.4)	12.0 (0.4)
8:30 am to 8:59 am	5.6 (0.2)	5.7 (0.3)
9:00 am to 9:59 am	6.2 (0.3)	6.5 (0.2)
10:00 am to 10:59 am	2.8 (0.2)	2.8 (0.2)
11:00 am to 11:59 am	1.4 (0.2)	1.3 (0.1)
12:00 pm to 3:59 pm	6.9 (0.3)	6.6 (0.3)
4:00 pm to 11:59 pm	6.5 (0.3)	6.9 (0.3)

Source: U.S. Census Bureau, 2016 American Community Survey Content Test.
Note: $\chi 2 = 6.7$, p-value=0.91. Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance testing was done at the α=0.1 level using a chi-square test.

Responses for Time of Departure were broken down by survey mode (internet, mail, or "interviewed," which includes CATI and CAPI). Among the three modes, test and control distributions were compared across Time of departure categories using a Chi-square test. The distributions for mail and interviewed modes were not statistically different, but the test and control distributions for internet mode were statistically different from one another (Table 16). Among individual Time of Departure categories, the 4:00 p.m. to 11:59 p.m. category for the control version was 1.9 percentage points higher than that of the test version.

Table 16. Time of Departure Distribution for Internet Response Mode

Time of Departure	Test Percent	Control Percent	Test Minus Control	Adjusted P-Value
12:00 am to 4:59am	3.0 (0.2)	2.7 (0.2)	0.3 (0.3)	1.00
5:00 am to 5:29 am	2.8 (0.2)	2.9 (0.2)	-0.1 (0.3)	1.00
5:30 am to 5:59 am	4.4 (0.3)	4.3 (0.3)	0.1 (0.4)	1.00
6:00 am to 6:29 am	7.8 (0.4)	7.3 (0.4)	0.5 (0.6)	1.00
6:30 am to 6:59 am	10.6 (0.4)	10.9 (0.5)	-0.3 (0.6)	1.00
7:00 am to 7:29 am	16.5 (0.5)	15.2 (0.5)	1.3 (0.8)	1.00
7:30 am to 7:59 am	15.2 (0.5)	14.8 (0.5)	0.5 (0.8)	1.00
8:00 am to 8:29 am	11.8 (0.4)	12.2 (0.4)	-0.4 (0.6)	1.00
8:30 am to 8:59 am	6.8 (0.3)	6.8 (0.3)	-0.1 (0.4)	1.00
9:00 am to 9:59 am	6.6 (0.4)	6.9 (0.3)	-0.4 (0.4)	1.00
10:00 am to 10:59 am	2.5 (0.2)	2.5 (0.2)	-0.1 (0.3)	1.00
11:00 am to 11:59 am	1.3 (0.2)	1.1 (0.1)	0.2 (0.2)	1.00
12:00 pm to 3:59 pm	6.1 (0.3)	5.8 (0.3)	0.2 (0.5)	1.00
4:00 pm to 11:59 pm	4.8 (0.3)	6.6 (0.3)	-1.9 (0.4)	<0.01*

Source: U.S. Census Bureau, 2016 American Community Survey Content Test.
Note: $\chi 2 = 23.6$, p-value=0.03. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. P-values with an asterisk (*) indicate a significant difference at the α=0.1 level.

5.4. Benchmarks

No other surveys collect directly comparable data to use as a benchmark in this analysis. The National Household Travel Survey (NHTS), a survey conducted by the U.S. Department of Transportation, includes information about travel mode, including public transportation. While differences in sample size and universe, collection years, methodology, and question wording preclude the possibility of a direct comparison, the 2009 NHTS, the most recent available, shows a public transportation rate of 5.1 percent among commuters, similar to that of the ACS in 2015 (5.1 percent of workers) and other recent years. No statistical testing was conducted for this comparison, but the NHTS provides a useful approximation for public transportation commuting rates and other ACS commuting estimates. For the purpose of this content test, the most appropriate benchmark is to compare responses from the test treatment to the current production questions, as done in the comparisons provided throughout section 5.

5.5. Response Error

This section addresses research questions number 4 and number 10: *Are the measures of response reliability (gross difference rate and index of inconsistency) better for the Test treatment than for the control treatment?*

To test this, a portion of the original sample population was reinterviewed, and the answers for their responses between the first and second interviews were compared.

Commute Mode

The hypothesis is that the increased clarity of the rail categories will lead to more consistent responses over time. Statistical significance between the GDR and IOI will be determined using a one-tailed t-test.

One limitation to assessing the reliability of this question is that the reference period of the question is "last week." The reference period for the original response will therefore always be different from the time frame for the CFU response. This could reasonably lead to a different answer between responses. We assume, however, that any inconsistency in responses due to this would occur at the same rate in the control version as in the test version. The GDR test shown in Table 17 indicates that there is variation in the degree of reliability across commute modes, but this variation is generally consistent between test and control versions. No travel mode category in the test treatment was statistically more reliable than their control treatment counterpart. The category "Subway or Elevated Rail" shows a comparatively low adjusted p-value that rounds to 0.10, but is still greater than 0.10.

Table 17. Difference in Gross Difference Rates (GDR) between Test Percent and Control Percent – Commute Mode

Response Category	Test GDR Percent	Control GDR Percent	Test Minus Control	Adjusted P-value
Car, truck, or van	4.7 (-0.4)	4.6 (0.4)	0.1 (0.5)	1.00
Bus	0.9 (0.1)	0.9 (0.1)	0.0 (0.2)	1.00
Subway or Elevated Rail	0.6 (0.1)	1.0 (0.2)	-0.4 (0.2)	0.10
Long-distance train or commuter rail	0.7 (0.1)	0.5 (0.1)	0.1 (0.2)	1.00
Light rail, streetcar, or trolley	-	-	-	-
Ferryboat	-	-	-	-
Taxicab	-	-	-	-
Motorcycle	-	-	-	-
Bicycle	0.2 (0.1)	0.3 (0.1)	-0.1 (0.1)	1.00
Walked	1.7 (0.3)	1.2 (0.2)	0.5 (0.3)	1.00
Worked from home	2.8 (0.3)	3.1 (0.3)	-0.3 (0.4)	1.00
Other method	1.1 (0.2)	1.0 (0.2)	0.1 (0.3)	1.00

Source: U.S. Census Bureau, 2016 American Community Survey Content Test.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a one-tailed t-test (test ≥ control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The '-' entry in a cell indicates that either no sample observations or too few sample observations were available to compute an estimate or standard error.

For Commute Mode, the IOI test results in Table 18 show a pattern similar to that of the GDR in that the degree of consistency in responses between original interviews and reinterviews were similar for test and control treatments. One rail-related category stands out as having comparatively low p-values, "Subway or elevated rail," but the test was not statistically lower

than the control. While the relative differences are instructive, the small sample size for these categories may limit the potential for statistically different results between treatments.

Table 18. Index of Inconsistency between Control and Test Treatments – Commute Mode

Response Category	Test IOI Percent	Control IOI Percent	Test Minus Control	Adjusted P-value
Car, truck, or van	19.4 (1.8)	19.6 (1.5)	-0.2 (2.2)	1.00
Bus	23.6 (3.7)	26.1 (3.8)	-2.5 (5.3)	1.00
Subway or Elevated Rail	22.2 (4.1)	35.2 (4.8)	-13.0 (5.9)	0.11
Long-distance train or commuter rail	38.6 (6.9)	39.4 (7.5)	-0.8 (10.5)	1.00
Light rail, streetcar, or trolley	—	—	—	—
Ferryboat	—	—	—	—
Taxicab	—	—	—	—
Motorcycle	—	—	—	—
Bicycle	18.6 (8.3)	21.6 (8.3)	-3.0 (11.8)	1.00
Walked	32.0 (4.4)	22.5 (4.5)	9.5 (6.4)	1.00
Worked from home	27.0 (2.9)	30.0 (2.5)	-3.0 (3.6)	1.00
Other method	78.0 (11.5)	84.9 (4.9)	-7.0 (11.6)	1.00

Source: U.S. Census Bureau, 2016 American Community Survey Content Test.
Note: Standard errors are shown in parentheses. Minor additive discrepancies are due to rounding. Significance was tested based on a one-tailed t-test (test ≥ control) at the α=0.1 level. P-values have been adjusted for multiple comparisons using the Holm-Bonferroni method. The '-' entry in a cell indicates that either no sample observations or too few sample observations were available to compute an estimate or standard error.

Time of Departure

A different approach was taken to test reliability for Time of Departure. We compared the proportion of follow-up responses that fell within a difference five minutes or less to their corresponding responses for the original interview (Table 19). For both the test and control treatments, about half of the response pairs (original and follow-up interviews) fell within five minutes of one another. While the control treatment showed a higher rate of response pairs within five minutes, the test rate was not significantly smaller.

Table 19. Persons Reporting a Difference of Five Minutes or Less for Time of Departure

Test Rate (%)	Control Rate (%)	Test Minus Control	P-Value
49.5 (1.2)	51.4 (1.0)	-1.9 (1.6)	0.12

Source: U.S. Census Bureau, 2016 American Community Survey Content Test.
Note: Standard errors are shown in parentheses. Significant at α=0.1 level based on a one-tailed t-test (test ≥ control).

5.6. Results for Analysis Specific to Journey to Work

Commute Mode

Public transportation systems are geographically concentrated within large cities and metro areas. Research question number 7 asks: *How do the test and control response distributions compare when the sample is restricted to only metro areas with high levels of overall rail usage?*

To improve our understanding of the commute mode distribution within areas where public transportation categories are most relevant, we combined the sample for metros with high rates of public transportation usage (see Section 2.4.6. for a list of metro areas). These are also metro areas with a diverse set of transportation options. For this group of metro areas, no significant difference was found between the commute mode distributions of test and control treatments (Rao-Scott Chi-Square = 7.8 and p-value=0.73).

Research question number 6 asks*: How do the test and control response distributions compare in metro areas with high levels of light rail usage?* To answer this, with a focus on areas where light rail is most relevant, a separate comparison combined metro areas with the 10 largest light rail systems. This includes the set of metro areas listed in Section 2.4.6. The results show no statistical differences between test and control treatments for any commute mode category (Rao-Scott Chi-Square = 8.1.

Research question number 5 asks: *For the paper questionnaire, is the proportion of person records that respondents incorrectly marked multiple modes of transportation comparable between control and test versions? When multiple modes are marked, if the sample size is large enough, which combinations are most common in each version?*

A final analysis specific to Commute Mode assessed the prevalence of respondents incorrectly marking two or more travel modes. When this occurs, the commute mode is allocated. The control treatment showed 59 unweighted incidences of respondents marking multiple commute modes, whereas the test treatment showed only 33 such incidences. For both treatments, the most prevalent combination of modes was a combination of bus and long-distance rail or bus and subway. Still, with such a small sample, the comparison two-tailed t-test shows no statistically significant differences between unweighted test and control treatments (p-value=.15).

Time of Departure

Respondents tend to answer this question with a time that is rounded, particularly on numbers ending in “0” and ”5” (Stapleton & Steiger, 2015). Research question number 11 explores the rate at which such rounding occurs between the test and control treatments of Time of Departure. We anticipated that the test version of Time of Departure would produce as many or fewer instances of this type of heaping. A two-tailed t–test was used to compare the percentage of responses that end in “0” or “5” for the test and control versions. For the test version, 98.7 percent of respondents heaped on a time ending in “0” or “5,” compared with 98.0 percent for the control version, but these rates were not statistically different from one another (p-value=.12). Still, this analysis was instructive in that it showed that a high percentage of

respondents round their Time of Departure answer to a "0" or "5," regardless of how the question is asked.

6. CONCLUSIONS AND RECOMMENDATIONS

This report discusses findings from the 2016 ACS Content Test for two questions related to commuting, Commute Mode and Time of Departure for Work. The motivation for modifying each question differed. For Commute Mode, the original set of categories reflected travel modes and terminology of the 1950s, when the question was developed. We modified commute mode categories to more accurately reflect the nation's public transportation options and the current terminology used to describe them.

Time of Departure has long been considered a sensitive question because it specifically asks respondents when they leave their home to go to work. Our aim is to develop a question that captures crucial information about when our nation's roads and transit systems are used throughout the day, while reducing the respondents' sensitivity to the question. We tested a new version of the question asking people what time their trip to work began, with the aim of asking the question in a way that seems less intrusive and does not include the word "home."

Among the various metrics used to answer our research questions, none revealed statistically different results between the test version and control version of each question. For both commuting variables, neither the distributions of the test version nor the control version were statistically different from one another. This is consistent with the expectation that the distribution of departure times would not differ between test and control versions. For Commute Mode, the distribution of rail-related categories did not differ between test and control treatments, which is a satisfactory outcome given that the goal was to ensure clarity among commute mode categories, not to change the distribution. This applies to individual rail-related modes as well as a special combined category (including the three rail-related categories). For both commuting questions, item response rates for the test treatment was not lower than that of the control treatment. Reliability metrics for both Commute Mode and Time of Departure did not show that the test version performed better than the control.

The final wording in the test versions of the commuting questions is the product of consultation with industry experts and extensive cognitive testing. This new wording is preferred to the control version of the ACS questions. Overall, the results of the various comparisons between test and control versions of each test showed surprising similarities between the two. The lack of significant differences between distributions suggests continuity in the meaning of the control and test versions of each question, which is an acceptable outcome. The overarching goal is to improve and clarify the wording of the question, not to alter the distribution. While smaller item missing data rates for test versions would be a favorable outcome, the findings of no significant difference in item missing data rates is also acceptable. The test versions of the Commute Mode and Time of Departure questions are preferred over the current version, therefore we recommend implementing the new test version of each question.

While transportation technology and travel behavior have changed rapidly in recent years, this iteration of ACS question modification changes has taken a conservative approach to modifying the Commute Mode questions by only refining and clarifying terminology for existing categories

rather than adding new categories. We will strongly consider the possibility of testing additional transportation categories that correspond with emerging travel trends in future ACS content test iterations.

7. ACKNOWLEDGEMENTS

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

- Census Bureau staff in the American Community Survey Office, Application Development and Services Division, Decennial Information Technology Division, Decennial Statistical Studies Division, Field Division, National Processing Center, Population Division, and Social, Economic, and Housing Statistics Division.
- Representatives from other agencies in the Federal statistical system serving on the Office of Management and Budget's Interagency Working Group for the ACS and the Topical Subcommittees formed by the Working Group for each topic tested on the 2016 ACS Content Test.
- Staff in the Office of Management and Budget's Statistical and Science Policy Office.

The authors would like to thank the following individuals for their contributions to the analysis and review of this report: Elizabeth Poehler, Nicole Scanniello, and Jennifer Ortman.

8. REFERENCES

American Public Transportation Association. (October 2014). *Light Rail & Streetcar Systems: How They Differ; How They Overlap.* Retrieved June 23, 2015 from American Public Transportation Association: http://www.apta.com/resources/reportsandpublications/Documents/APTA Light Rail-Streetcars-How They Differ-How They Overlap Oct 14.pdf

Chappell, G., & Obenski, S. (November 2014). *ACS Fiscal Year 2014 Content Review Results.* Washington, D.C.: U.S. Census Bureau. Retrieved June 24, 2015 from http://www.census.gov/acs/www/Downloads/operations_admin/2014_content_review/Methods%20and%20Results%20Report/2014_ACS_Content_Review_Final_Documentation.pdf

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

Federal Highway Administration, U.S. Department of Transportation. (2009). *Introduction to the 2009 NHTS*. Retrieved June 24, 2015 from National Household Travel Survey: Our Nation's Travel: http://nhts.ornl.gov/introduction.shtml

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

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

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

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

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

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

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

APPENDIX A. Supplemental Table for Unit Response Rates

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

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

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