Behavior Coding of October 2018 Agricultural Labor Survey

Joseph B. Rodhouse
Heather Ridolfo
Emilola J. Abayomi
David Biagas Jr.

United States Department of Agricultural

National Agricultural Statistics Service

Research and
Development Division
Washington, DC 20250

RDD Research Report
Number RDD-21-02

April 2021

The findings and conclusions in this report are those of the author and should not be construed to represent any official USDA or U.S. Government determination or policy.

EXECUTIVE SUMMARY

This report examines interviewer and respondent performance in the 2018 October Agricultural Labor Survey and the effort to increase interviewer standardization in question administration by reducing interviewer deviations from the interviewing script. Behavior coding studies of the 2017 October and 2018 April Agricultural Labor Surveys found serious and widespread issues with both interviewer and respondent performance in the computer-assisted telephone interview (CATI) mode of data collection. This report investigates whether the standardized interviewing intervention encouraging interviewers to read survey questions on the October version of the 2018 Agricultural Labor Survey exactly as worded had an impact in three areas: 1) the proportion of questions that were read exactly as worded, 2) the proportion of first exchanges between the interviewer and respondent resulting in a response that satisfies the intent of the question, and 3) the proportion of final exchanges with a satisfactory response per the question's intent.

Additionally, several factors that could impact interviewer and respondent performance in the three areas described above are considered. These include the Data Collection Center (DCC) where interviewers conducted survey interviews, the version of the questionnaire being administered (the original questionnaire or the experimental questionnaire), the reference period being asked about (e.g., last week vs. a week from three months ago), and the specific questions or question themes being administered. Finally, the report examines whether additional interviewer training before the October 2018 data collection that encouraged interviewers to read questions exactly as they are worded improved standardized data collection compared to the April 2018 data collection. The April 2018 data also contained an experimental questionnaire testing new questions on wages.

In general, the results show that standardization of question administration improved from about 12 percent to 53 percent between the 2018 April and October versions of the survey after encouraging interviewer adherence to the interviewing script. Interviewer standardization behaviors did not significantly differ between questionnaire versions. However, respondent behaviors improved in the experimental (revised) version. Both interviewer performance and respondent performance were significantly worse in the battery of questions for the second reference period (the set of questions for July) compared to the first reference period (the set of questions for October). Interviewer standardization behaviors also varied significantly by question and question theme. Respondent behaviors were similarly affected. More instances of interviewer standardization and code-able answers from respondents occurred for questions/question themes that appeared in the beginning and end of the survey. Desirable interviewer and respondent behaviors varied significantly among DCCs and suggests that some DCCs are performing standardized interviewers at much higher rates than other DCCs.

RECOMMENDATIONS

1. Dedicate additional resources for interviewer training on conducting standardized interviews.

2. Dedicate additional resources for CATI instrument and questionnaire improvements.
    a. Decrease cognitive burden of the labor questions over the phone.
    b. Increase the usability design (UX) of the Blaise CATI instrument for this survey.

3. Change from a biannual survey to a quarterly survey.

4. Increase buy-in for standardization from all the DCC's. All DCC's should be relatively equal in interviewer standardization rates.

TABLE OF CONTENTS

Behavior Coding of the October 2018 Agricultural Labor Survey

Joseph Rodhouse[1]
Heather Ridolfo
Emilola J. Abayomi
David Biagas, Jr.

Abstract

Cognitive testing on the Agricultural Labor Survey in 2016 revealed that respondents were having difficulty mapping their responses to questions. This led to further inquiries into the data collection process for this survey in the computer-assisted telephone interview (CATI) mode, which is the primary source of completed interviews for this survey. Behavior coding was conducted on recorded CATI interviews and found widespread problems with both interviewer and respondent behaviors. Further efforts were undertaken to improve both the survey's CATI questionnaire and interviewer performance in administering the survey, and new training and instructions were provided for interviewers on how to conduct standardized interviews. This report examines the extent to which standardized interviews were achieved after making these changes in time for the October 2018 Agricultural Labor Survey data collection. Several factors that impact standardized performance are examined, including characteristics of the questionnaire and questions, and the interviewer's Data Collection Center. The findings demonstrate that efforts to increase interviewer standardization in question administration improved significantly compared to the previous iteration of the survey in April 2018. However, only a slight majority of all the questions administered by interviewers are read exactly as worded. Thus, more training is still needed for interviewers to continue improving the rate at which standardized interviews are conducted. The results highlight the need for questionnaire designers and Blaise designers to work together to make it easier for interviewers to conduct standardized interviews.

Key Words: Behavior Coding, Interviewer-Respondent Interaction, Data Quality

1. INTRODUCTION

This report marks the third installment of a series of behavior coding studies analyzing interviewer and respondent behaviors in the Agricultural Labor Survey conducted by the USDA's National Agricultural Statistics Service (NASS). In 2016, cognitive testing of the questionnaire found numerous issues with respondents' abilities to accurately map their responses to the questions in the intended format (Sloan 2017). In response, NASS developed an experimental questionnaire that had new survey questions. A field test compared the experimental questionnaire to the original questionnaire during the 2017 October Agricultural Labor Survey as detailed in Biagas et al. (2019). After comparing the field test results, the

[1] Authors are researchers at the U.S. Department of Agriculture's (USDA) Research and Development Division (RDD) in the Survey Methodology and Technology Section (SMTS).

experimental questionnaire was further revised. A second field test during the 2018 April Agricultural Labor Survey compared the revised experimental and original questionnaires. The 2018 April Agricultural Labor Survey field test evaluated the impact of questions on base, bonus, and overtime wages on overall data quality. For the April 2018 field test, the original and experimental versions of the survey were exactly the same with the exception of the inclusion of base, bonus, and overtime wages in the experimental version.

Behavior coding studies of the 2017 October and 2018 April Agricultural Labor Surveys found serious and widespread issues with both interviewer and respondent performance in the computer-assisted telephone interview (CATI) mode of data collection (Ridolfo et al. 2020, 2021). Principally, interviewers were more often than not deviating from the interviewing script, and thus negatively impacting the quality of the data collected. To address the issues discovered, problematic questions were redesigned to make it easier for interviewers to collect the desired information, resulting in a revised 2018 October Agricultural Labor Survey questionnaire. Also, interviewers received additional training and instructions for the new versions of the survey.

This report examines interviewer and respondent performance in the 2018 October Agricultural Labor Survey and the effort to increase interviewer standardization in question administration by reducing interviewer deviations from the interviewing script. Evidence-based research in the survey methodology literature suggests that as interviewer standardization increases (i.e., following the interviewing script exactly as worded) the quality of the data collected also increases (see Groves 1989, for an overview). Many characteristics of the data collection process can impact the ability of interviewers and respondents to engage in a standardized fashion during question administration, and several of these specific to the Agricultural Labor Survey will be used to put context around interviewer and respondent exchanges. For example, question-level characteristics, survey-level characteristics, and location-level characteristics (i.e., where interviewers are located) are used to examine how each impacts standardization in the Agricultural Labor Survey. The results inform where in the data collection process standardization is going well and where areas of improvement are still needed.

2. BACKGROUND

Two recent studies at NASS examined interviewer and respondent behaviors in the Agricultural Labor Survey. In the first study (Biagas et al. 2019), it was found that interviewers were largely employing flexible (e.g., conversational) interviewing styles to administer the survey. Interviewers deviated from the interviewing script often and read a low proportion of questions exactly as worded. The second study measured how these deviations might be impacting overall data quality. Specifically, were these deviations helping or hurting the quality of responses?

The Agricultural Labor Survey can be a complicated form for many respondents. The length of the survey (i.e., completion time) can increase for establishments with many and varying types of hired workers. Administration of the survey over the phone can present additional challenges, and interviewer experience with the agricultural population may prime them to anticipate the burden imposed on the respondents. Interviewers may be adopting flexible interviewing styles in an attempt to reduce respondent burden and increase the respondents' comprehension of the questions and to motivate respondents to provide responses when they are fatigued later in the survey. The conversational style could result in better data quality in the aggregate compared to strict standardization.

However, the results of the second NASS study proved otherwise when analyzing CATI data in the 2018 April Agricultural Labor Survey. The proportion of data unlikely to contain measurement errors significantly decreased when interviewers deviated from strict standardization (Rodhouse et al. 2019). Furthermore, the authors found that quality control mechanisms in the data review and editing stage did not adequately identify and correct responses containing measurement errors. For example, only 9 percent of the data identified as containing measurement errors were corrected to something more accurate during the review and editing stage (Rodhouse et al. 2019). Additionally, approximately 7 percent of the data identified by the authors as being accurate were changed to something less accurate during this stage. The authors concluded that response quality decreased as a result of interviewers deviating from strict standardization in the interviewing script. In addition, efforts should be made to increase standardization in this survey because the data review and editing stage was not reliably able to correct measurement errors on the backend.

One aim of this report is to investigate whether the standardized interviewing intervention encouraging interviewers to read survey questions on the October version of the 2018 Agricultural Labor Survey exactly as worded had an impact in three areas: 1) the proportion of questions that were read exactly as worded, 2) the proportion of first exchanges between the interviewer and respondent resulting in a response that satisfies the intent of the question, and 3) the proportion of final exchanges with this outcome. Based on the results from Rodhouse et al. (2019), increases in each of these three areas is likely to be correlated with better data quality overall. To examine this, the results are compared to the April 2018 version of the Agricultural Labor Survey.

The call for more standardized interviewing behaviors is not new. In fact, a synthesis of research on the topic showed that the proportion of survey questions read exactly as worded by interviewers varied widely in the survey methodology literature (Groves, 1989). Early studies on interviewer behaviors found that the proportion of survey questions read exactly as worded could be as low as 30 percent (Oksenberg 1981) and as high as 96 percent (Mathiowetz and Cannell 1980). The literature has also shown that even slight wording changes can have major effects on aggregate data distributions (Willis 2005; Bradburn and Sudman 1991; Schuman and Presser 1981; Sudman and Bradburn 1982). Other research has shown that interviewers who deviate from the instructions given for administering a questionnaire may increase response errors in various ways. For instance, non-standardized reading of the questions may lead to increases in the number of non-substantive responses in the first exchange. Non-substantive responses in the first exchange may require additional probing. If interviewer behaviors, such as question reading and probing, vary across interviewers, the intra-interviewer correlation could increase, thus increasing the variance of descriptive estimates and reducing the effective sample size of the survey (West and Blom 2017; Groves 2004).

The ability to produce an answer that satisfies the meaning of the survey question (i.e., a code-able answer) is subject to both respondent characteristics and interviewer effects (Groves 1989). Although respondent sex (Groves and Magilavy 1986) and education level (Fowler and

Mangione 1985) seem unrelated to interviewer effects in some ways, the age of the respondent appears to be related. Older respondents tend to exhibit more nonresponse (Groves and Kahn 1979) and, if they respond, to exhibit greater response errors on questions requiring recall of factual material (Sudman and Bradburn 1973).

The Agricultural Labor Survey is a good candidate to assess whether encouraging adherence to the interviewing script can reduce the likelihood of these errors for a couple of reasons. First, the agricultural population tends to skew older. According to the 2017 Census of Agriculture, the average age of a farmer is 57.5. Second, the Agricultural Labor Survey requires recall of factual material for two different reference periods. Therefore, because this survey population tends to skew older and the questionnaire has questions that require a lot of recall, interviewer variability in question administration could exacerbate response errors and increase the variance of descriptive estimates.

One challenge in trying to encourage more adherence to the interviewing script in the Agricultural Labor Survey is buy-in from data collection centers and interviewers that have a lot of experience administering surveys to the agricultural population. In their experience, this population tends to be more receptive to conversational interviewing and rapport building, and these interviewing styles result in better outcomes (e.g., greater response rates). One way NASS survey designers tried to address this in the 2018 October survey was by including a sentence for interviewers to read to respondents in the introduction of the survey that informed them that all questions would be read exactly as worded, and that even though it may seem repetitive at some points, that was what the interviewer was tasked to do. This type of forecasting is akin to what Fowler and Mangione (1990) recommended for setting the stage for standardized interviews. Their argument was that if interviewers explain to respondents why questions are going to be asked exactly as worded, then respondents will do a better job in a standardized interview interaction. Interviewer training on this new questionnaire script also reinforced the importance of interviewers adhering to the interviewing script during question administration.

3. METHODOLOGY

Behavior coding was conducted on a subsample of the original and experimental versions of the survey. Behavior coding is an objective, quantitative method for studying the interviewer and respondent interaction (Fowler and Cannell 1996). Part of the goal of this behavior coding effort was to evaluate interviewer and respondent behavior across the two versions of the 2018 October Agricultural Labor Survey. Particularly, the research focused on determining whether there were differences between the two versions in terms of the interviewers' ability to administer the questions in a standardized manner and respondents' abilities to provide adequate responses.

In behavior coding, each turn in the interview can be coded. A turn begins when the first person begins speaking and ends when the second person begins speaking. A pair of turns is referred to as an exchange (Ongena and Dijkstra 2006). During the administration of a single question, a number of exchanges may occur before a final answer to the question is given. In general, the ideal scenario in an interviewer-administered survey entails one exchange (or turn) per question. That is, the interviewer asks the survey question exactly as worded one time, and the respondent subsequently (in that same exchange) provides an answer that satisfies the intent of the question.

However, deviations from this ideal scenario often occur, leading to multiple exchanges. For example, the respondent may need the question to be repeated, ask for clarification, or provide an inadequate response that requires follow-up probes from the interviewer. As a result, assigning behavior codes to each exchange that may occur for a particular question can be time consuming. Research has found there to be diminishing returns to coding all exchanges for a single question (Oksenberg et al. 1991). Therefore, only the first exchange for each question and the final response given by the respondent (which may have occurred in the first exchange, second exchange, third exchange, etc.) were coded.

For the purposes of this study, the behaviors for both the interviewer and respondent were coded. Although interviewer behavior was evaluated, it is important to remember the goal of the study is not to rate their performance, but rather to identify patterns in the data that provide insight into whether any systematic problems occur during data collection in the aggregate. The results highlight questions that may be difficult for interviewers to administer in a standardized way and problems with the data collection instrument itself, as much as it does interviewer performance. Accordingly, it is incumbent upon NASS researchers and survey designers to use the results from this report to ultimately make the job of the interviewer easier by fixing problematic question designs and CATI functionalities.

The codes used to assess the behavior of interviewers are summarized in Table 1 below.

Table 1. Behavior Codes for Interviewer Behavior

Code	Description
ES	Exact wording
ESOP	Exact wording + optional text
VER	Verified response
MC	Major change
SC	Shortcutting (Falsifying or failing to verify a response)
OTH	Other

If the interviewer read the question as worded with only slight or minimal changes, the ES code was applied. Questions were coded as MC (major change) if interviewers read the questions in a manner that substantially altered the question meaning. Verification (VER) occurred when interviewers verified a response that respondents preemptively provided in previous questions or made an assumption about the response. Finally, questions were coded as shortcutting (SC) if interviewers failed to read the question entirely or failed to verify a response. Major changes and shortcutting are considered to be problematic behavior. When these codes are applied to a question at least 15 percent of the time, it is an indication that there is a problem with the survey question (Fowler 2011).

The codes used to assess the response behavior of the respondents are summarized in Table 2 below. If the respondent gave an answer that satisfied the intent of the question, the behavior was coded as CA (codable answer) if it occurred in the first exchange. It was coded as CAFR (codable answer final response) if the behavior was the final result before the interviewer moved on to the next question. Ideally, in a survey interview, every question has CAFR for the respondent behavior before the interviewer proceeded throughout the survey. In general, more instances of CA and CAFR in the results are likely correlated with better data quality. Areas where instances of these codes are low, and instead have one of the other codes from Table 2, are likely correlated with worse data quality.

Table 2. Behavior Codes for Respondent Behaviors in the 1st and Final Exchanges

Respondent Behavior	Code for 1st Exchange	Code for Final Exchange
Provided a response/answer that satisfied the intent/meaning of the question	CA	CAFR
Asked for clarification	CLAR	-
Said they don't know	DK	DKFR
No response - the interviewer did not read the question to the respondent	SC	SCFR
Interrupted the interviewer during the reading of the question	INTERRPT	-
Answered the intro text instead of waiting to hear the question	INTRO	INTROFR
Gave a “qualified” response, expressing doubt or confusion about the answer	QA	QAFR
Refused to provide an answer	REF	REFFR
Corrected interviewer when interviewer verified information rather than reading the question	VERCORR	VERCORRF
Gave no response/was silent when interviewer verified information rather than reading the question	VERNORES	VERNORESF
Other	OTHR	OTHFR

Three researchers trained in behavior coding coded the interviews. Before coding began, Cohen’s kappa was calculated to ensure consistency across coding. Four interviews were selected at random for the kappa calculation. Two of these interviews were conducted using the original instrument and two were conducted using the experimental instrument. Each researcher coded the four interviews independently. Cohen's kappa (Cohen 1960) was calculated for all possible coder pairs. The overall average of these kappa combinations was 0.70, indicating substantial agreement among the three coders (Landis and Koch 1977). It is reasonable to conclude that the results of the behavior coding research reflect real and identifiable patterns in the data and are not a result of random chance alone.

4. RESULTS

The results follow a format designed to highlight how certain characteristics of the Agricultural Labor Survey may impact standardized interview administration. In the aggregate, the findings should help identify areas where standardization can be improved. To start, in section 4.1, the report examines standardization broadly by comparing the difference in overall standardization between the 2018 October survey and the 2018 April survey. Following this, the results drill down into more specific areas. The rate of standardization is compared by the questionnaire version (original/control vs. revised/experimental) in section 4.2 and by the reference period (i.e., the battery of questions about a week in October vs. the battery of questions about a week in July) and question order in section 4.3. In section 4.4, standardization by question theme (e.g., Number in Worker Category vs. hours worked vs. wages paid, etc.) is analyzed. After this, the role of data collection centers (DCC's) in producing standardized interviews is critiqued in section 4.5 (for example, it could be that some DCC's have higher standardization rates than others, or that certain question themes see better standardization rates than others). Finally, within each of these sections, results for the number of exchanges that occurred between the interviewer and respondent and the distributions of respondent behaviors by the first exchange and final exchange are displayed.

4.1 Data Collection Period (October vs. April)

The final dataset consisted of a total of 3,682 behavior-coded items for analysis, with 1,622 items from the April data collection and 2,060 from the October data collection. The behavior coding results comparing April to October are compelling and show marked improvement in standardized administration between the two data collections.

Table 4.1a shows the distributions of interviewer behaviors for the two data collections. Comparing the two data collections to each other, the overall Chi-square of 1046.68 ($p < 0.0001$) demonstrates that the interviewer behaviors exhibited are significantly associated with the data collection iteration. In the April 2018 iteration, the proportion of questions read exactly as worded (ES) by interviewers was 11.1 percent. After changes were made to encourage interviewers to read the questions exactly as worded, the proportion jumped to 52.77 percent in the October 2018 iteration (chi-square = 698.15, $p < 0.0001$). Shortcutting (SC) (falsifying or streamlining) dropped from 49.01 percent in April to 11.55 percent in October (chi-square = 630.90, $p < 0.0001$), and major changes (MC) fell from 29.28 percent to 23.20 percent (chi-square = 17.49, $p < 0.0001$). Although major changes fell, the number is still higher than the 15 percent threshold defining a systematic problem (Fowler 2011).

Table 4.1a: Interviewer Behaviors Overall: Comparison between April and October 2018

Data Collection Iteration	*April 2018*	*October 2018*	*Test of Independence*	
Interviewer Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
ES	11.10	52.77	698.15	<.0001
ESOP	1.29	1.41	0.09	0.77
SC	49.01	11.55	630.90	<.0001
MC	29.28	23.20	17.49	<.0001
VER	9.25	11.02	3.10	0.08
OTH	0.06	0.05	0.03	0.87

Notes: Overall Chi-square of Interviewer Behavior (April 2018 vs. October 2018) = 1046.68, $p < 0.0001$.

The behavior coding results comparing respondent behaviors in the first exchange between April and October are also compelling and show marked improvement between the two data collections. In Table 4.1b, the distributions of respondent behaviors for the two data collections are shown. The overall Chi-square of the table is 673.24 ($p < 0.0001$), meaning that respondent behaviors are significantly associated with data collection iteration. It is important to note that for almost all the respondent codes besides “provided a response/answer that satisfied the intent/meaning of the question” it is difficult to determine whether the observed changes are due to the differences in the number of times interviewers skipped the questions. As a result, the focus here is on the proportion of times respondents provided a response that satisfied the intent of the question and the proportion of times respondents were unable to respond because interviewers skipped the question.

In the April 2018 iteration, the proportion of first exchanges in which the respondent was able to provide a response/answer that satisfied the intent or meaning of the question (CA) was 34.27 percent. After changes were made to encourage interviewers to read the questions exactly as worded, the proportion jumped to 61.47 percent in the October 2018 iteration (chi-square = 163.91, $p < 0.0001$). The proportion of times the respondent was unable to answer because the interviewer skipped the question (SC) fell from 49.01 percent in April 2018 to 12.62 percent in October 2018 (chi-square = 595.55, $p < 0.0001$).

Table 4.1b Respondent Behaviors Overall in the 1st Exchange: By Data Collection Iteration

Data Collection Iteration	*April 2018*	*October 2018*	*Test of Independence*	
Respondent Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
CA	34.27	61.47	163.91	<.0001
CLAR	4.04	5.90	3.32	0.07
DK	0.68	1.14	1.23	0.27
SC	49.01	12.62	595.55	<.0001
INTERRPT	0.93	0.81	0.043	0.51
INTRO	0.37	0.22	1.03	0.31
QA	3.48	7.36	18.21	<.0001
REF	0.06	0.16	0.59	0.44
VERCORR	1.00	0.32	7.38	0.01
VERNORES	2.24	4.71	11.28	0.001
OTHR	3.48	4.06	0.09	0.76

Notes: Overall Chi-Square of Respondent Behavior (April 2018 vs. October 2018) = 673.24, $p < 0.0001$.

Table 4.1c shows the distributions of respondent behaviors in the final exchange. Respondent behaviors in the final exchange were significantly associated with data collection iteration (overall Chi-square = 668.18, $p < 0.0001$). In the April 2018 iteration, the proportion of final exchanges in which the respondent was able to provide a response/answer that satisfied the intent or meaning of the question (CAFR) was 41.18 percent. After changes were made to encourage interviewers to read the questions exactly as worded, the proportion increased to 71.65 percent in the October 2018 iteration (chi-square = 193.14, $p < 0.0001$). The proportion of times the respondent was unable to answer because the interviewer skipped the question (SCFR) fell from 49.01 percent in April 2018 to 14.09 percent in October 2018 (chi-square = 587.90, $p < 0.0001$).

Table 4.1c Respondent Behaviors Overall in the Final Exchange: By Data Collection Iteration

Data Collection Iteration	*April 2018*	*October 2018*	*Test of Independence*	
Respondent Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
CAFR	41.65	71.65	193.14	<.0001
DKFR	0.31	0.81	2.96	0.09
SCFR	49.01	14.09	587.90	<.0001
INTROFR	0.44	0.22	1.72	0.19
QAFR	2.24	4.77	11.75	0.001
REFFR	0.06	0.16	0.59	0.44
VERCORRF	0.87	0.24	6.80	0.01
VERNORESF	2.18	4.32	13.04	0.0003
OTHFR	2.68	3.20	0.15	0.70

Notes: Overall Chi-Square of Final Exchange (April 2018 vs. October 2018) = 668.18, $p <$ 0.0001.

In Table 4.1d, the distributions of the number of exchanges needed to arrive at the final answer are shown. The overall Chi-square of 456.70 ($p < 0.0001$) indicates that the number of exchanges is significantly associated with the data collection iteration. In the April 2018 survey, the proportion of the time it took just one exchange between interviewers and respondents to arrive at an answer that satisfied the intent of the question was 35.66 percent. After changes were made to encourage interviewers to read the questions exactly as worded, the proportion jumped to 65.95 percent in the October 2018 iteration (chi-square = 344.36, $p < 0.0001$). No exchanges largely occurred when the interviewer failed to provide the respondent an opportunity to respond as a result of falsifying or streamlining through the question. This type of interviewer behavior, which leads to no exchanges, fell from 51.01 percent in April 2018 to 18.22 percent in October 2018 (chi-square = 421.64, $p < 0.0001$).

Table 4.1d: Number of Exchanges between the Interviewer and Respondent Needed to Arrive at the Final Answer: By Data Collection Iteration

Data Collection Iteration	*April 2018*	*October 2018*	*Test of Independence*	
Number of Exchanges	**Percent**	**Percent**	**Chi-Square**	**P-value**
One exchange	35.66	65.95	344.36	<.0001
Two exchanges	7.67	9.74	5.44	0.02
Three or more exchanges	5.66	6.09	0.45	0.50
Not applicable (because interviewer behavior = "SC")	51.01	18.22	421.64	<.0001

Notes: Overall Chi-Square of Number of Exchanges (April 2018 vs. October 2018) = 456.70, *p* < 0.0001.

4.2 Questionnaire Version

The final dataset consisted of 2,060 behavior coded items from the October data collection, with 1,001 for the original (control) version of the questionnaire and 1,059 from the experimental version of the questionnaire. The behavior coding results comparing the two versions show few significant differences, meaning that the proportion of questions administered by interviewers in a standardized fashion were roughly the same.

In Table 4.2a, the distributions of interviewer behaviors for the two questionnaire versions are shown. The overall Chi-square of 11.21 (*p* = 0.047) indicates that interviewer behaviors were significantly associated with questionnaire version. However, as Table 4.2a suggests, this association seems to be driven by interviewers' tendencies to exhibit verification (VER) behaviors more so in the control version than the experimental version (Chi-square = 6.93, *p* = 0.01). None of the other interviewer behaviors showed significant associations with questionnaire version. For instance, in the control version of the questionnaire, interviewers read the questions exactly as worded (ES) 53.25 percent of the time. In the experimental version of the questionnaire, the proportion was 52.31 percent (chi-square = 0.18, *p* = 0.67). The proportion of times interviewers made major changes to the question wording or meaning (MC) was marginally significantly higher in the experimental questionnaire compared to the control, with 24.74 percent of questions falsified or streamlined compared to 21.58 percent, respectively (chi-square = 2.89, *p* < 0.09).

Table 4.2a: Interviewer Behaviors Overall: Comparisons between Questionnaire Versions

Version of the Questionnaire	*Control*	*Experimental*	*Test of Independence*	
Interviewer Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
ES	53.25	52.31	0.18	0.67
ESOP	1.10	1.70	1.34	0.24
SC	11.09	11.99	0.41	0.52
MC	21.58	24.74	2.89	0.09
VER	12.89	9.25	6.93	0.01
OTH	0.10	0.00	1.06	0.30

Notes: Overall Chi-square of Interviewer Behavior by Questionnaire Version = 11.21, *p* = 0.047

Next, the distributions of respondent exchanges by questionnaire version were analyzed and presented in Table 4.2b. The overall Chi-square of 24.17 (*p* = 0.007) indicates that respondent behaviors were significantly associated with the version of the questionnaire being administered. The proportion of first exchanges that resulted in a codable answer (CA) was marginally higher in the experimental version compared to the control version (Chi-square = 2.42, *p* = 0.12). However, comparing individual respondent behaviors between the two questionnaire versions yielded mostly no significant associations or only marginal associations, except for instances where there were verification behaviors by the interviewer and no response from the respondent to the verification statement. Instances where there was verification and no response (VERNORES) was significantly associated with questionnaire version (Chi-square = 6.11, *p* = 0.01), meaning the experimental version of the questionnaire resulted in fewer instances of this undesired exchange.

Table 4.2b: Respondent Behaviors in the 1st Exchange: By Questionnaire Version

Version of the Questionnaire	*Control*	*Experimental*	*Test of Independence*	
Respondent Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
CA	59.57	63.10	2.42	0.12
CLAR	5.07	6.65	2.09	0.15
DK	0.68	1.56	3.21	0.07
SC	15.09	12.68	2.25	0.13
INTERRPT	0.90	0.73	0.17	0.68
INTRO	0.45	0.00	NA	NA
QA	7.77	6.96	0.44	0.51
REF	0.34	0.00	NA	NA
VERCORR	0.45	0.21	0.84	0.36
VERNORES	5.97	3.53	6.11	0.01
OTHR	3.49	4.57	1.39	0.24

Notes: Overall Chi-square of Respondent Behaviors in the First Exchange by Questionnaire Version = 24.17, *p* = 0.007.

Table 4.2c below presents the distributions of respondent behaviors in the final exchange by questionnaire version. The overall Chi-square of 27.08 (*p* = 0.001) indicates that respondent behaviors in the final exchanges were significantly associated with questionnaire version. In the control version of the questionnaire, the proportion of final exchanges in which the respondent was able to provide a response/answer that satisfied the intent or meaning of the question (CAFR) was 69.54 percent. In the experimental version, the proportion was 73.60 percent (chi-square = 4.18, *p* = 0.04). The significant chi-square statistic suggests that the ability to provide substantive response in the final exchange is associated with the questionnaire version. As noted above, the experimental version also yielded marginally more substantive responses in the first exchange than the control version. Given that interviewers were no different in standardized administration between the two versions, the evidence suggests that the experimental version of the questionnaire is superior at prompting respondents to produce a substantive response by the final exchange.

Table 4.2c: Respondent Behaviors in the Final Exchange: By Questionnaire Version

Version of the Questionnaire	*Control*	*Experimental*	*Test of Independence*	
Respondent Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
CAFR	69.54	73.60	4.18	0.04
DKFR	0.45	1.14	2.74	0.10
SCFR	15.58	12.68	3.19	0.07
INTROFR	0.45	0.00	NA	NA
QAFR	4.63	4.89	0.07	0.79
REFFR	0.34	0.00	NA	NA
VERCORRF	0.34	0.21	NA	NA
VERNORESF	6.21	3.53	7.19	0.01
OTHFR	2.37	3.95	3.72	0.05

Notes: Overall Chi-square of Respondent Behaviors in the Final Exchange by Questionnaire Version = 27.08 (*p* = 0.001).

Table 4.2d shows the distributions of the total number of exchanges needed to arrive at the final answer. The overall Chi-square of 9.74 (*p* = 0.021) means that there is a significant association between the version of the questionnaire and the number of exchanges it takes to arrive at a final answer. In the control version of the questionnaire, the proportion of the time it took just one exchange between interviewers and respondents to arrive at an answer that satisfied the intent of the question was 65.93 percent. In the experimental version, the proportion was 65.53 percent (chi-square = 0.04, *p* = 0.85). The two questionnaire versions were also no different from each other in the prevalence of 2 exchanges (Chi-square = 1.49, *p* = 0.22); however, the prevalence of 3 or more exchanges was significantly associated with the questionnaire version (Chi-square = 5.53, *p* = 0.02), suggesting that the experimental version of the questionnaire may yield more exchanges to result in a final answer, overall.

Table 4.2d: Total Number of Exchanges between the Interviewer and Respondent: By Version of the Questionnaire

Version of the Questionnaire	*Control*	*Experimental*	*Test of Independence*	
Number of Exchanges	**Percent**	**Percent**	**Chi-Square**	**P-value**
One exchange	65.93	65.53	0.04	0.85
Two exchanges	8.89	10.48	1.49	0.22
Three or more exchanges	4.80	7.27	5.53	0.02
Not applicable (because interviewer behavior = "SC")	19.88	16.53	3.90	0.05

Notes: Overall Chi-Square of Total Exchanges by Questionnaire Version = 9.74 (p = 0.021).

4.3 Reference Week

The final dataset consisted of 2,060 behavior coded items from the October data collection, with 1,068 (51.84 percent) for the battery of questions about the first reference period (October 7-13, 2018) and 992 items (48.16 percent) for the battery of questions about the second reference period (July 8-14, 2018). For context, fewer items were coded for the second reference period because a few respondents indicated they did not have any hired labor for that reference period. The behavior coding results comparing the two reference periods show several significant differences, meaning that the proportion of questions administered by interviewers in a standardized fashion was associated with the reference period.

In Table 4.3a, the distributions of interviewer behaviors for the two reference periods are shown. The overall Chi-square of 98.71 ($p < 0.0001$) indicates that interviewer behaviors are significantly associated with reference period. In the battery of questions for the week in October, the proportion of first exchanges in which the respondent was able to provide a response/answer that satisfied the intent or meaning of the question (ES) was 58.05 percent. In the battery of questions for the week in July, the proportion was 47.08 percent (Chi-square = 24.86, $p < 0.0001$). The proportion of times the respondent was unable to answer because the interviewer skipped the question (SC) was significantly higher in the battery of questions for the week in July compared to the week in October, with 16.03 percent of questions falsified or streamlined compared to 7.40 percent, respectively (Chi-square = 37.5, $p < 0.0001$). Similarly, the proportion of times the interviewer verified (VER) previous information rather than reading the question as intended or worded was significantly higher in the July portion of the survey compared to the October portion of the survey, with 15.93 percent of questions being "verified" in July and 6.46 percent of questions being "verified" in October (Chi-square = 47.01, $p <$ 0.0001). The proportion of questions where interviewers made major changes (MC) to the wording or meaning of the question was 26.5 percent in the October battery of questions and 19.66 percent in the July battery of questions (Chi-square = 47.01, $p <.0001$). The significant Chi-square indicates there is an association between reference period and interviewers making major changes to questions when they are reading them to respondents. The fact that it is lower in the July reference period could be due to interviewers' higher proclivity to use the information from the first reference period to either skip reading the question altogether in the July portion or to "verify" the previous information rather than reading the July questions as instructed.

Table 4.3a Interviewer Behaviors Overall: Comparisons between Reference Weeks

Battery of Labor Questions	*Week in October*	*Week in July*	*Test of Independence*	
Interviewer Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
ES	58.05	47.08	24.86	<.0001
ESOP	1.59	1.21	0.54	0.46
SC	7.40	16.03	37.50	<.0001
MC	26.5	19.66	13.51	0.0002
VER	6.46	15.93	47.01	<.0001
OTH	0.00	0.10	1.08	0.30

Notes: Overall Chi-Square of Interviewer Behaviors by Reference Week = 98.71, $p < 0.0001$.

Respondent behaviors in the first exchange with the interviewer by the question reference period are shown below in Table 4.3b. The overall Chi-square of 99.24 ($p < 0.0001$) indicates that respondent behaviors in the first exchange are significantly associated with reference period. In the battery of questions for October, the proportion of first exchanges in which the respondent was able to provide a response/answer that satisfied the intent or meaning of the question (CA) was 63.02 percent. In the battery of questions for July, the proportion was 59.94 percent (Chi-square = 1.69, $p = 0.190$). The Chi-square statistic suggests that the ability to provide a substantive response in the first exchange is not associated with reference period. Asking for clarification (CLAR) (Chi-square = 9.44, $p = 0.002$), saying "don't know" (DK) (Chi-square = 5.85, $p = 0.016$), and giving a qualified response (QA) (Chi-square = 19.96, $p < 0.0001$) are associated with the battery of questions for each reference period, and suggests that these undesirable outcomes in the first exchange are higher for the first battery of questions. Respondents giving no response when the interviewer verified previous information rather than reading the question was also associated with reference period (VERNORES) (Chi-square = 28.85, $p < 0.0001$), but suggests this undesirable outcome is associated with the battery of questions for the second reference period in the survey. These results make sense, as respondents are more likely to have questions or be confused at the beginning of surveys in the first round of questions rather than in the second round of questions where the same questions are repeated for a different reference period. Lastly, interviewer shortcutting (SC) was significantly associated with reference period (Chi-square = 27.34, $p < 0.0001$), indicating that interviewers exhibited more of this behavior when administering the second battery of questions.

Table 4.3b: Respondent Behaviors in the 1st Exchange: By Reference Period

Battery of Labor Questions	*Week in October*	*Week in July*	*Test of Independence*	
Respondent Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
CA	63.02	59.94	1.69	0.190
CLAR	7.59	4.21	9.44	0.002
DK	1.74	0.54	5.85	0.016
SC	9.65	18.03	27.34	<0.0001
INTERRPT	0.54	1.08	1.67	0.196
INTRO	0.00	0.43	NA	NA
QA	10.09	4.64	19.96	<0.0001
REF	0.33	0.00	NA	NA
VERCORR	0.11	0.54	NA	NA
VERNORES	2.06	7.34	28.85	<0.0001
OTHR	4.88	3.24	3.16	0.075

Notes: Overall Chi-Square of Respondent Behaviors in 1st Exchange by Reference Period = 99.24 (p < 0.0001).

Respondent behaviors in the final exchange with the interviewer by the question reference period are shown below (Table 4.3c). The overall Chi-square of 84.67 (p < 0.0001) indicates that respondent behaviors in the final exchange are significantly associated with reference period. In the battery of questions for October, the proportion of final exchanges in which the respondent was able to provide a response/answer that satisfied the intent or meaning of the question (CAFR) was 76.17 percent, and 67.17 percent in July (Chi-square = 17.38, p < 0.0001). The Chi-square statistic suggests that the ability to provide a substantive response in the final exchange is significantly associated with reference period. Saying “don’t know” was significantly associated with reference period in the first exchange but not in the final exchange (DK) (Chi-square = 0.62, p = 0.430). Consistent with the first exchange, in the final exchange, a qualified response (QAFR) is associated with the battery of questions for each reference period (Chi-square = 17.40, p < 0.0001) and suggests that these undesirable outcomes in the final exchange are higher for the first battery of questions. Also consistent with the first exchange, respondents giving no response (VERNORESF) when the interviewer verified previous information rather than reading the question was also associated with reference period in the final exchange (Chi-square = 30.47, p < 0.0001), and suggests this undesirable outcome is associated with the battery of questions for the second reference period in the survey. These results point to respondents likely being more confused and having more questions at the beginning of the survey in the first round of questions rather than in the second round of questions where the same questions are repeated for a different reference period, and interviewers are more likely to verify answers given for the first reference period rather than ask the same questions for the second reference period.

Table 4.3c: Respondent Behaviors Overall in the Final Exchange: By Reference Period

Battery of Labor Questions	*Week in October*	*Week in July*	*Test of Independence*	
Respondent Behavior	**Percent**	**Percent**	**Chi-Square**	**P-value**
CAFR	76.17	67.17	17.38	<0.0001
DKFR	0.98	0.65	0.62	0.430
SCFR	9.90	18.25	26.84	<0.0001
INTROFR	0.00	0.43	NA	NA
QAFR	6.86	2.70	17.40	<0.0001
REFFR	0.33	0.00	NA	NA
VERCORRF	0.11	0.43	NA	NA
VERNORESF	1.96	6.48	30.47	<0.0001
OTHFR	3.59	2.81	0.88	0.346

Notes: Overall Chi-Square of Respondent Behaviors in Final Exchange by Reference Period = 84.60 (p < 0.0001).

Table 4.3d shows the distribution of the number of exchanges needed to arrive at the final answer. The overall Chi-square of the table is 37.45 (p < 0.0001), suggesting that number of exchanges was significantly associated with reference period. In the battery of questions for the first reference period (the week in October), the proportion of the time it took just one exchange between interviewers and respondents to arrive at an answer that satisfied the intent of the question was 67.70 percent. In the battery of questions for the second reference period (the week in July), the proportion was 63.61 percent. The Chi-square statistic (Chi-square = 3.82, p = 0.051) suggests that reference period is marginally associated with resulting in only one exchange needed to arrive at a final answer that satisfies the intent of the question. The same conclusion can be drawn for the two reference periods in the prevalence of two exchanges (Chi-square = 3.36, p = 0.07). However, the prevalence of three or more exchanges was significantly associated with reference period (Chi-square = 6.87, p = 0.01), suggesting that the battery of questions for the first reference period may yield more exchanges to result in a final answer. The type of interviewer behavior that leads to no exchanges increased from 13.48 percent in the battery of questions for the first reference period to 23.19 percent in the second reference period. The significant Chi-square statistic (Chi-square = 421.64, p <.0001) suggests that reference period is associated with the prevalence of no exchanges between interviewers and respondents. This finding is mostly due to interviewers’ tendency to exhibit shortcutting behaviors more in the second reference period than the first.

Table 4.3d: Number of Exchanges between the Interviewer and Respondent: By Reference Period

Battery of Labor Questions	*Week in October*	*Week in July*	*Test of Independence*	
Number of Exchanges	**Percent**	**Percent**	**Chi-Square**	**P-value**
One exchange	67.70	63.61	3.82	0.051
Two exchanges	10.86	8.47	3.36	0.07
Three or more exchanges	7.40	4.64	6.87	0.01
Not applicable (because interviewer behavior = "SC")	13.48	23.19	32.58	< 0.0001

Notes: Overall Chi-Square of Number of Exchanges by Reference Period = 37.45 ($p < 0.0001$).

4.4 Question Theme

The final dataset consisted of 2,060 behavior coded items from the 2018 October Agricultural Labor Survey. The behavior coding results comparing question themes show several significant differences, indicating that the proportion of questions administered by interviewers in a standardized fashion was associated with the theme of the question being asked. Table 4.4a illustrates the survey questions that comprise the question theme. In Table 4.4b, the distribution of interviewer behaviors by question theme is shown; in Table 4.4c, the distribution of respondent behaviors in the first exchange by question theme is shown; in Table 4.4d, the distribution of respondent behaviors in the final exchange is shown; and in Table 4.4e, the distribution of the number of exchanges needed to arrive at the final answer by question theme is shown.

Table 4.4a: Survey Questions Comprising the Question Themes

Question Theme	Question Text
Intro Text	The agricultural labor survey is the only survey that provides employment and wage estimates for all workers directly hired by farms and ranches in the United States. I will be asking you questions about agricultural workers that you had on your payroll in October and July. Some of these questions may seem repetitive, but I need to ask them as worded.
	Let's categorize the [n] worker(s) based on the type of work they were HIRED TO DO. I will go through the five categories now. Report each worker under ONE CATEGORY.
	Let's separate those [n] [worker category] into more specific categories. I have [n] categories to choose from.

	Now I have some questions about the [n] [worker category] workers you had during that week. Let's separate those [n] workers into specific categories. I have [n] categories to choose from.
Screeners	Did this operation have anyone on payroll to do agricultural work the week of Sunday, October 7th THROUGH Saturday, October 13th, 2018?
	Now I will ask about a week in July. Did this operation have anyone on the payroll to do agricultural work the week of Sunday, July 8th THROUGH Saturday, July 14th, 2018?
Number of Workers	How many workers did you have on the payroll to do agricultural work that week?
Includes/Excludes	Did you have any part-time workers, paid family members, or hired managers on the payroll to do agricultural work that week in [October/July]?
	Were any of these workers the following types: Contract workers, Custom workers, Retail workers, Value added workers?
	How many CONTRACT, CUSTOM, RETAIL, or VALUE ADDED WORKERS did you have?
	This survey does not include these types of workers. If I exclude them, you had [n] workers that week. Is that correct?
Number in Worker Category	During the week of October 7th THROUGH October 13th, 2018, how many of the [n] paid workers were hired to be [worker category]?
	During the week of July 8th THROUGH July 14th, 2018, how many of the [n] paid workers were hired to be [worker category]?
Worker Category Verification	I want to verify that I have your worker(s) categorized correctly. The other possible types of workers are: [worker categories]. Do any of these categories better describe what the workers were hired to do that week?
Hours Worked	Let’s talk about the [n] [worker category]. How many TOTAL HOURS did these [n] [worker category] work that week?
Gross Wages	What were the total gross wages for these [n] [worker category] that week?
Base Wages	How much of the [$$$] gross wages paid that week were BASE wages? Base wages include the minimum amount paid regardless of method of pay (salaried, hourly, piece rate etc.) but exclude overtime and bonus pay.
Wage Unit	Is that a Total Amount Paid, Hourly Wage, Average Weekly Salary PER WORKER, Average Monthly Salary PER WORKER or Average Annual Salary PER WORKER?
Gross Wages Confirmation	The total gross wages for those [n] [worker category] that week was ($$$). Is that correct?
Bonus or Overtime Wages (asked separately)	How much of the [$$$] gross wages paid that week were BONUS wages?
	How much of the [$$$] gross wages paid that week were OVERTIME wages?

Bonus or Overtime Wages (asked together)	Did you pay any BONUS or OVERTIME pay?
Number working 150 days or more	In 2018, how many of these [n] workers will be paid by the operation for 150 days or more?
Peak Number on Payroll	During 2018, what was or will be the largest number of hired workers on the payroll on any one day?
Any H-2A	During 2018, did or will this operation have any H-2A Temporary Agricultural Workers on the payroll?

Interviewer behaviors were significantly associated with question theme (overall chi-square for Table 4.4b = 291.50, $p < 0.0001$). Given the significance of the overall chi-squared statistics, potential differences in interviewers' performance for specific behavior across question themes was explored. Interviewers' performance in reading questions exactly as worded (interviewer behavior = "ES") was significantly associated with the theme of the question being read to the respondent (chi-square = 117.07, p < 0.0001). In other words, interviewers were better at reading some types of questions exactly as worded than others. Some question themes had a high proportion of being read exactly as worded (see "Screeners," "Any H-2A," and "Peak Number on Payroll"), while some had a very low proportion of being read exactly as worded (see "Bonus Wages," "Base Wages," and "Wage Unit"). The rest of the question themes were read exactly as worded only about half the time.

Table 4.4b: Interviewer Behaviors by Question Theme

Question Theme	*Interviewer Behavior*					
	ES	*ESOP*	*SC*	*MC*	*VER*	*OTHR*
Intro Text	66.29	0.00	13.71	20.00	0.00	0.00
Screeners	57.04	0.00	9.15	33.80	0.00	0.00
Number of Workers	64.29	0.00	5.71	17.14	12.86	0.00
Includes/Excludes	70.37	2.47	7.41	12.35	7.41	0.00
Number in Worker Category	50.30	2.54	11.49	23.73	11.94	0.00
Worker Category Verification	57.61	0.00	13.04	20.65	8.70	0.00
Hours Worked	55.06	3.93	6.18	26.40	8.43	0.00
Gross Wages	52.25	0.00	8.43	26.97	11.80	0.56
Base Wages	22.89	0.00	18.07	34.94	24.10	0.00
Wage Unit	22.83	0.00	23.91	17.39	35.87	0.00
Gross Wages Confirmation	51.90	0.00	20.25	16.46	11.39	0.00
Bonus or Overtime Wages (asked separately)	40.63	0.00	25.00	21.88	12.50	0.00
Bonus or Overtime Wages (asked together)	44.00	0.00	16.00	14.00	26.00	0.00
Number working 150 days or more	51.52	0.00	9.09	30.30	9.09	0.00
Peak Number on Payroll	80.56	2.78	0.00	11.11	5.56	0.00
Any H2A	77.78	5.56	2.78	11.11	2.78	0.00

Notes: Numbers shown are percentages. Overall Chi-square of Question Theme by Interviewer Behavior = 291.50 (p < 0.0001). The Chi-square of Question Theme by Interviewer Behavior code ES = 117.07 (p < 0.0001).

Figure 4.4b below more clearly demonstrates interviewers' abilities in administering certain question themes in a standardized manner. The chart is displayed in ascending order, where the question themes that were read exactly as worded, the lowest proportion of the time appear at the top, and those with the highest proportion appear at the bottom. What is interesting is that the questions that are best read by interviewers are the very first question of the survey ("Screeners") and the final two questions of the survey ("Any H-2A" and "Peak Number on Payroll"). Interviewers did not perform as well on question themes that fall in between the first question and the last two questions (likely as a result of increased question complexity in the middle of the survey, but there could be an order effect, as well). Interviewers have significant problems administering wage questions in a standardized fashion when it comes to base wages, bonus wages, overtime wages, and categorizing the wage unit.

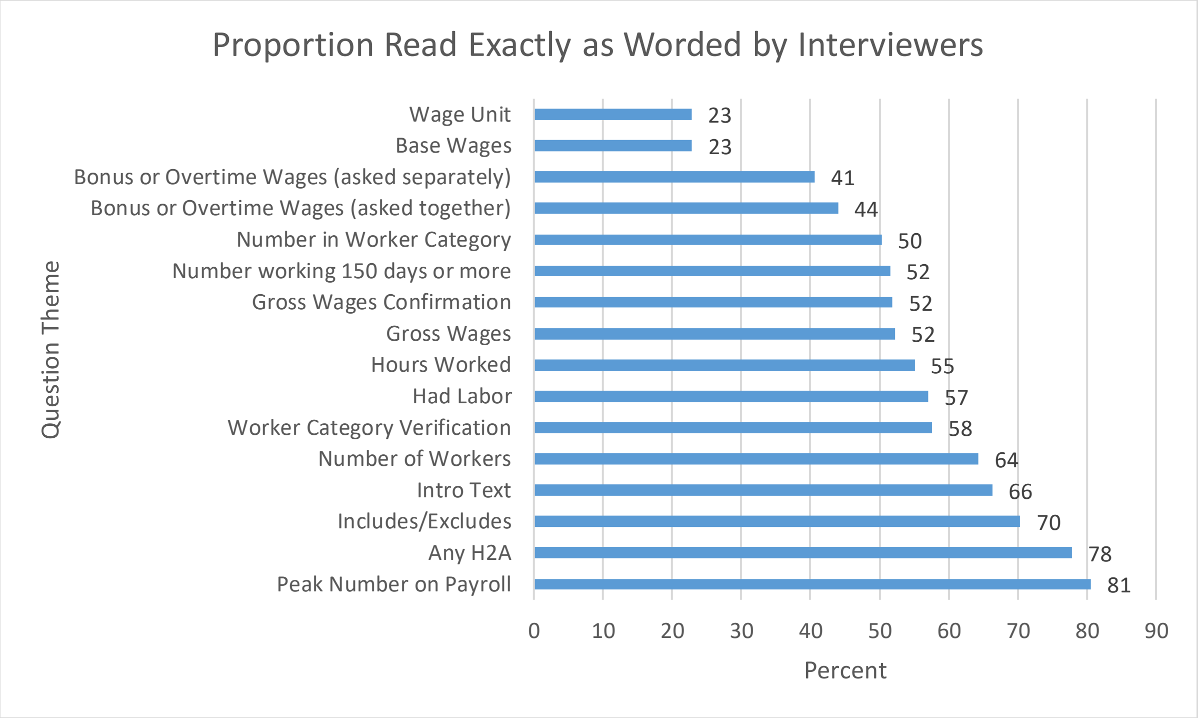

Figure 4.4b Proportion of Question Themes Read Exactly as Worded by Interviewers

Respondent behaviors in the first exchange are significantly associated with question theme (overall chi-square of the table 4.4c = 1921.27, $p < 0.0001$). More granularly, respondents' ability to provide codable responses (CA) in the first exchange were significantly associated with the question theme being asked by the interviewer (chi-square = 61.05, $p < .0001$). This means that respondents were significantly better able to provide an adequate answer for some questions in the first exchange than others. Some question themes had a relatively high proportion of first exchanges result in a codable answer (see "Bonus or Overtime Wages," "Peak Number on Payroll," and "Any H2A"), while some had a relatively low proportion (see "Base Wages," "Wage Unit," "Includes/Excludes," and "Gross Wages"). For those with a relatively low proportion, the result is likely due to the fact that interviewers falsified or streamlined on those question themes more than on other question themes, thereby not allowing respondents as many opportunities to provide a response at all on those questions.

Table 4.4c: Respondent Behaviors in the 1st Exchange; By Question Theme

	Respondent Behavior										
Question Theme	**CA**	**CL**	**DK**	**SC**	**INTRRPT**	**INTRO**	**OTH**	**QA**	**RF**	**VC**	**VNR**
Screeners	54.65	12.79	1.16	13.95	1.16	2.33	6.98	6.98	0.00	0.00	0.00
Number of Workers	65.71	8.57	1.43	5.71	0.00	0.00	1.43	10.00	0.00	0.00	7.14
Includes/Excludes	69.14	4.94	0.00	7.41	3.70	0.00	0.00	12.35	0.00	0.00	2.47
Number in Worker Category	66.47	3.44	0.30	14.82	0.15	0.15	2.84	6.44	0.00	0.75	4.64
Worker Category Verification	57.61	9.78	0.00	15.22	5.43	0.00	3.26	7.61	0.00	0.00	1.09
Hours Worked	59.55	7.30	2.25	6.18	0.56	0.00	6.74	13.48	0.00	0.00	3.93
Gross Wages	54.49	5.62	6.18	8.43	0.00	0.00	8.43	10.67	1.69	0.56	3.93
Base Wages	50.60	10.84	0.00	18.07	1.20	0.00	6.02	8.43	0.00	0.00	4.82
Wage Unit	51.09	2.17	0.00	23.91	1.09	0.00	1.09	1.09	0.00	0.00	19.57
Gross Wages Confirmation	54.43	1.27	1.27	21.52	0.00	0.00	13.92	0.00	0.00	0.00	7.59
Bonus or Overtime Wages (asked separately)	65.63	3.13	0.00	25.00	0.00	0.00	0.00	3.13	0.00	0.00	3.13
Bonus or Overtime Wages (asked together)	72.00	2.00	0.00	16.00	0.00	0.00	0.00	2.00	0.00	0.00	8.00
Number working 150 days or more	68.18	12.12	0.00	10.61	0.00	0.00	3.03	4.55	0.00	0.00	1.52
Peak Number on Payroll	72.22	11.11	2.78	0.00	0.00	0.00	0.00	13.89	0.00	0.00	0.00
Any H2A	72.22	19.44	0.00	2.78	0.00	0.00	0.00	5.56	0.00	0.00	0.00

Notes: Values displayed are percentages. The overall Chi-square of Question Theme by Respondent Behavior in the 1st Exchange = 1921.27 ($p < 0.0001$). The Chi-square for Question Theme by Respondent Behavior code CA = 61.05 ($p < 0.0001$).

Figure 4.4c below more clearly demonstrates respondents' abilities in responding to certain question themes in the first exchange. The chart is displayed in ascending order; the question themes that had the lowest proportion of codable responses in the first exchange appear at the top, and those with the highest proportion appear at the bottom. It is interesting, but perhaps expected, that the questions that require a one-word answer, such as 'yes,' 'no,' or for example, 'twenty', experienced the highest proportion of responses in the first exchange (see "Bonus or Overtime Wages," "Any H2A," "Peak Number on Payroll," and "Number Working 150 Days or More"). Respondents were least able to provide codable responses in the first exchange question themes regarding base wages, wage units, includes/excludes, gross wages, worker category verification questions, and questions about hours worked. A large part of these results could be due to interviewers' relative inabilities to read the questions exactly as worded to the respondents.

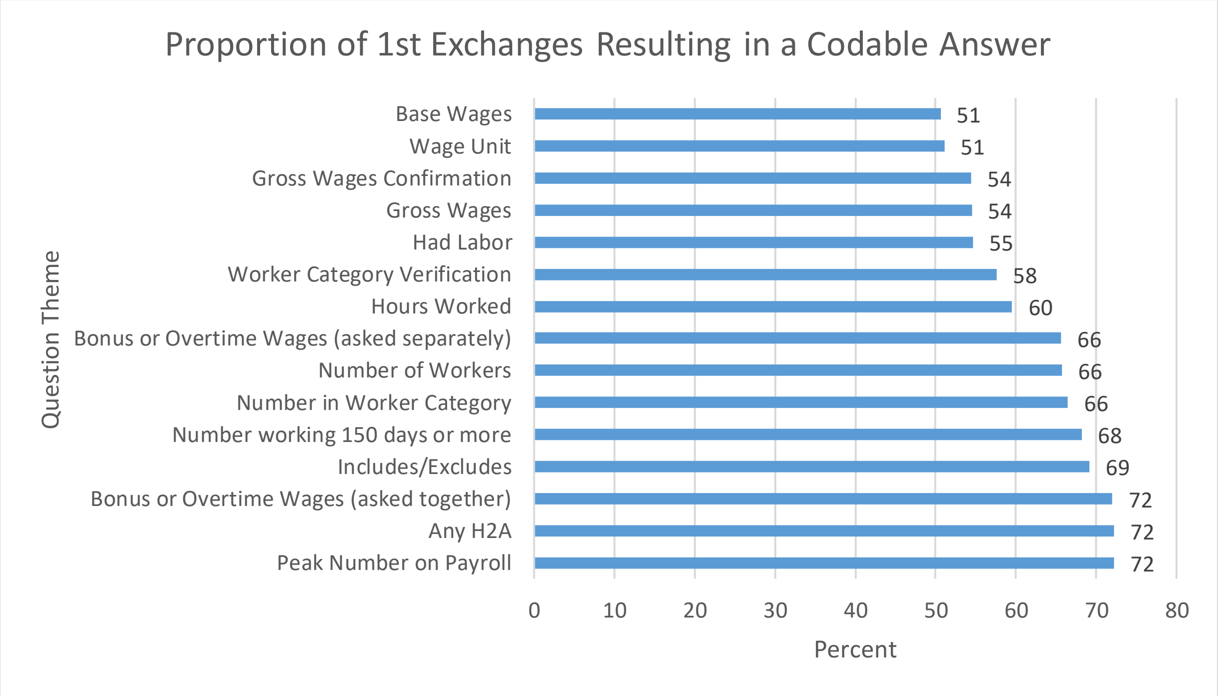

Figure 4.4c Proportion of First Exchanges Resulting in a Codable Answer: By Question Theme

Respondent behaviors in the final exchange are significantly associated with question theme (overall chi-square of Table 4.4d = 1829.03, p < 0.0001). More granularly, respondents' ability to provide codable responses (CAFR) in the final exchange was significantly associated with the question theme being asked by the interviewer (Chi-square = 95.23 *p* <.0001). This means that respondents were significantly better able to provide an adequate final answer for some questions in the final exchange than others. Some question themes had a relatively high proportion of final exchanges result in a codable answer (see "Bonus or Overtime Wages," "Peak Number on Payroll," and "Any H2A"), while some had a relatively low proportion (see "Base Wages," "Wage Unit," "Includes/Excludes," and "Gross Wages"). For those with a relatively low proportion, the result is likely due to the fact that interviewers falsified or streamlined on those question themes more so than on other question themes, thereby not allowing respondents as many opportunities to provide a response at all on those questions. Secondarily, these questions were perhaps more difficult (or burdensome) for respondents to answer.

Table 4.4d: Respondent Behaviors in the Final Exchange: By Question Theme

	Respondent Behavior								
Question Theme	CAFR	DK -FR	SCFR	INTRO -FR	QA-FR	RF-FR	VC-FR	VN-RFR	OTH-FR
Screeners	71	1	14	2	6	0	0	0	0
Number of Workers	77	3	6	0	7	0	0	7	0
Includes/Excludes	84	0	7	0	5	0	0	2	0
Number in Worker Category	73	0	15	0	4	0	1	4	1
Worker Category Verification	75	0	16	0	3	0	0	1	0
Hours Worked	75	1	6	0	10	0	0	4	0
Gross Wages	70	3	9	0	7	2	1	4	1
Base Wages	61	0	18	0	8	0	0	5	0
Wage Unit	54	0	24	0	1	0	0	14	5
Gross Wages Confirmation	57	0	22	0	0	0	0	6	1
Bonus or Overtime Wages (Separately)	69	0	25	0	3	0	0	3	0
Bonus or Overtime Wages (Together)	74	2	16	0	0	0	0	8	0
Number working 150 days or more	82	2	11	0	5	0	0	2	0
Peak Number on Payroll	86	3	0	0	11	0	0	0	0
Any H2A	97	0	3	0	0	0	0	0	0

Notes: Values displayed are percentages rounded to the nearest whole number. The overall Chi-square of Respondent Behaviors in the Final Exchange by Question Theme = 1829.03 ($p < 0.0001$). The Chi-square for Respondent Behaviors in the Final Exchange code CAFR by Question Theme = 95.23 ($p < 0.0001$).

Figure 4.4d below more clearly demonstrates respondents' abilities in responding to certain question themes in the final exchange. The chart is displayed in ascending order, where the question themes that had the lowest proportion of codable responses in the final exchange appear at the top, and those with the highest proportion appear at the bottom. Similar to responses in the final exchange, the questions that require a one word answer, such as 'yes,' 'no,' or for example, 'twenty', experienced the highest proportion of responses in the final exchange (see "Bonus or

Overtime Wages," "Any H2A," "Peak Number on Payroll," and "Number Working 150 Days or More"). Respondents were least able to provide codable responses in the final exchange question themes regarding wage units, base wages, includes/excludes, bonus wages, and overtime wages. A large part of these results could be due to interviewers' relative inabilities to read the questions exactly as worded to the respondents, which could have confused respondents. Secondarily, these questions were perhaps more difficult (burdensome) for respondents to answer, thereby resulting in a lower proportion of codable responses in the final exchange.

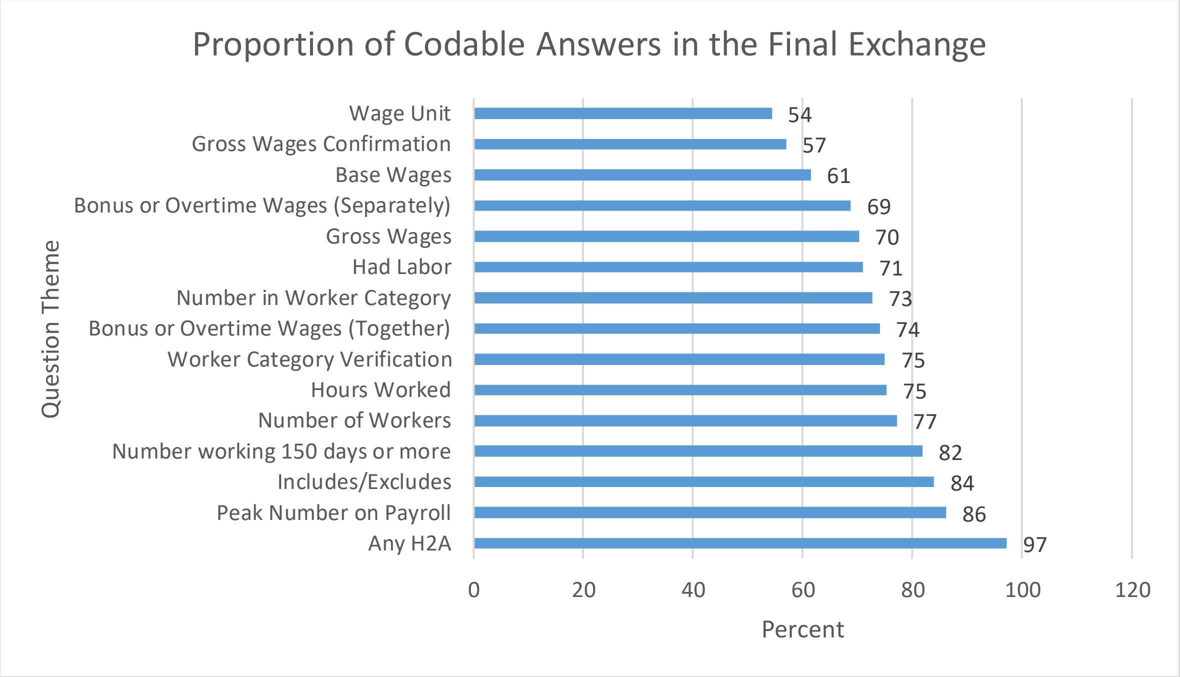

Figure 4.4d Proportion of Codable Answers in the Final Exchange: By Question Theme

The number of exchanges between the interviewer and respondent needed to arrive at a final answer was significantly associated with question theme (overall chi-square of Table 4.4e = 313.57, $p < 0.0001$). More specifically, the proportion of the time only one exchange was needed between the interviewer and respondent to arrive at an answer that satisfied the intent of the question was significantly associated with question theme (chi-square = 62.34 p <.0001). Question themes that most often only required one exchange between the interviewer and respondent were "Bonus Wages," "Bonus or Overtime Wages," "Screeners," and "Totals" questions. Each of these question themes achieved this 75 percent of the time or more. The question themes that were least successful at this data quality measure were "Gross Wages," "Includes/Excludes," and "Worker Category Verification" questions. Each of these question themes resulted in one exchange less than 60 percent of the time.

Table 4.4e: Number of Exchanges between the Interviewer and Respondent: By Question Theme

	Number of Exchanges			
Question Theme	**One exchange**	**Two exchanges**	**Three or more exchanges**	**Not Applicable (because interviewer behavior = SC)**
Screeners	61.27	9.86	2.11	26.76
Number of Workers	75.71	4.29	11.43	8.57
Includes/Excludes	70.37	11.11	8.64	9.88
Number in Worker Category	73.27	6.61	5.56	14.56
Worker Category Verification	58.70	19.57	6.52	15.22
Hours Worked	61.58	18.08	11.86	8.47
Gross Wages	55.93	19.77	12.99	11.30
Base Wages	68.67	9.64	4.82	16.87
Wage Unit	60.87	6.52	3.26	29.35
Gross Wages Confirmation	68.35	3.80	0.00	27.85
Bonus or Overtime Wages (Separately)	71.88	3.13	0.00	25.00
Bonus or Overtime Wages (Together)	72.00	12.00	0.00	16.00
Number working 150 days or more	74.24	6.06	7.58	12.12
Peak Number on Payroll	61.11	25.00	13.89	0.00
Any H2A	72.22	16.67	8.33	2.78

Notes: The overall Chi-square of Number of Exchanges by Question Theme = 313.57 (*p* < 0.0001). The Chi-square for Question Theme*Number of Exchanges (where Number of Exchanges = "One Exchange") = 62.34, *p* <.0001.

Figure 4.4e below more clearly demonstrates interviewers' and respondents' abilities to keep the interactions for each question theme to one exchange. The chart is displayed in ascending order, where the question themes that had the lowest proportion of interactions resulting in one exchange appear at the top, and those with the highest proportion appear at the bottom. A large part of these results could be due to interviewers' relative inabilities to read the questions exactly as worded to the respondents, which could have confused respondents. Secondarily, questions regarding worker pay, and categorizing workers (what category and whether their workers qualify to be included in the responses) were perhaps more difficult (burdensome) for respondents to answer, thereby resulting in a lower proportion of interactions between the interviewer and respondent that only needed one exchange.

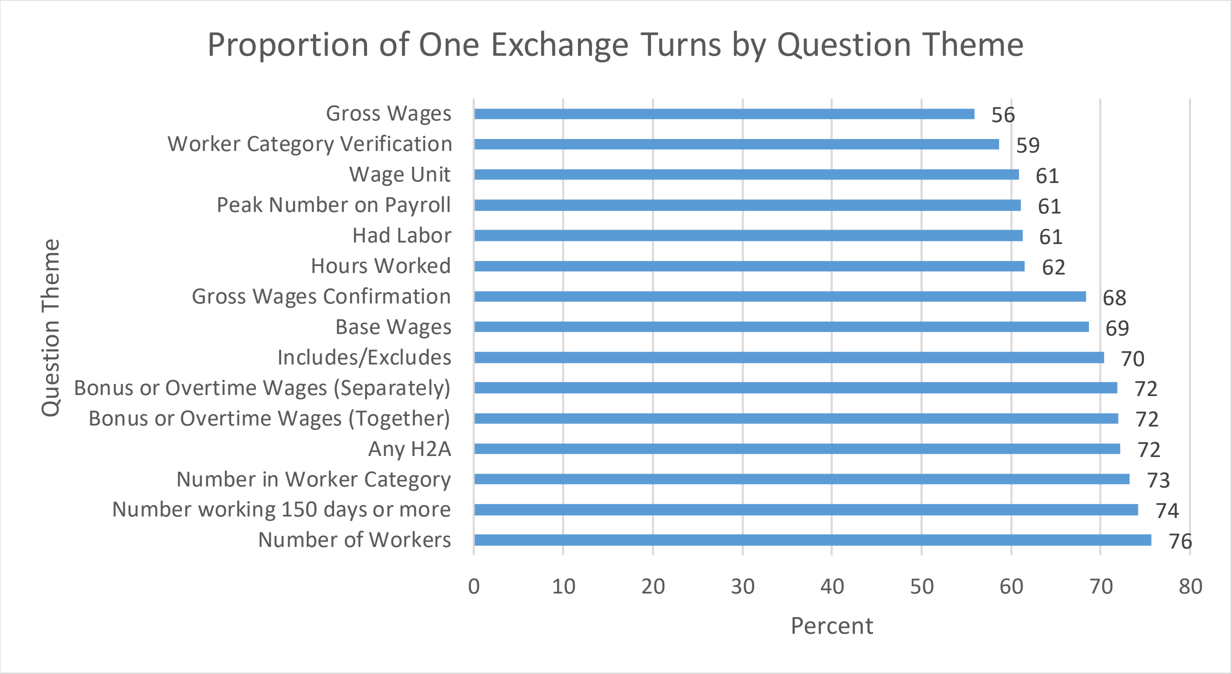

Figure 4.4e Proportion of One Exchange Turns between the Interviewer and Respondent: By Question theme.

4.5 Data Collection Center

The final dataset consisted of 2,060 behavior coded items from the 2018 October Agricultural Labor Survey, with 322 items from the Arkansas (AR) DCC, 430 items from the Montana (MT) DCC, 382 items from the National Operations Center (NOD) DCC, 504 items from the Oklahoma (OK) DCC, and 422 items from the Wyoming (WY) DCC. The behavior coding results comparing the DCCs show several significant differences, meaning that the proportion of questions administered by interviewers in a standardized fashion was associated with the locations where interviewers were conducting interviews. In Table 4.4a, the distribution of interviewer behaviors between the two reference periods is shown; in Table 4.4b, the distribution of respondent behaviors in the first exchange is shown; in Table 4.4c, the distribution of respondent behaviors in the final exchange is shown; and in Table 4.4d, the distribution of the number of exchanges needed to arrive at the final answer is shown.

The distributions of interviewer behaviors differed among DCCs (overall Chi-square of Table 4.5a = 221.05, *p* < 0.0001) suggests that interviewer behavior was significantly associated with DCC. The proportion of questions interviewers read exactly as scripted (ES) was significantly differed among DCCs where interviewers were housed (chi-square = 146.59, *p* < 0.0001). The DCCs with the highest proportion of standardized interviewing occurred within NOD (63.61 percent), MT (62.56 percent), and AR (60.87 percent). At the WY DCC, the proportion of questions read as scripted was 53.79 percent and 30.16 percent at the OK DCC. Interviewer behaviors such as falsifying or streamlining (SC) (chi-square = 70.29, *p* < 0.0001), making major changes (MC) to question wording or meaning (chi-square = 73.67, *p* < 0.0001), and verifying previous information (VER) while not reading the question as instructed (chi-square = 14.94, *p* = 0.005) all were significantly associated with interviewers' DCCs. The OK DCC was the only DCC to have the proportion of falsifying or streamlining be above the 15 percent threshold constituting a systematic problem, with 20.63 percent of questions being coded as such. Three DCCs were over the 15 percent threshold defining a systematic problem for making major changes to question wording or meaning: OK (35.52 percent), WY (24.41 percent), and MT (21.86 percent). The other two DCCs were very close to the threshold: NOD (14.66 percent) and AR (14.29 percent). Verification behaviors were also significantly associated with DCC location, but none of them reached the 15 percent threshold.

Table 4.5a Interviewer Behaviors Overall: Comparisons between Data Collection Centers

DCC	*AR*	*MT*	*NOD*	*OK*	*WY*	*Test of Independence*	
Interviewer Behavior	**Percent**	**Percent**	**Percent**	**Percent**	**Percent**	**Chi-Square**	**P-value**
ES	60.87	62.56	63.61	30.16	53.79	146.59	<.0001
ESOP	1.86	0.47	3.4	0.79	0.95	16.21	0.003
SC	9.63	6.74	4.71	20.63	13.27	70.29	<.0001
MC	14.29	21.86	14.66	35.52	24.41	73.67	<.0001
VER	13.35	8.14	13.61	12.9	7.58	14.94	0.005
OTH	0	0.23	0	0	0	3.79	0.43

Notes: The overall Chi-square of Interviewer Behaviors by Data Collection Centers (DCC) = 221.05 (*p* < 0.0001).

The distributions of respondent behaviors in the first exchange differed significantly with DCC (overall chi-square of the Table 4.5b = 134.18, *p* < 0.0001). In particular, providing a response that satisfied the meaning of the question (CA) (chi-square = 25.94, *p* <.0001) was associated with the DCC where interviewers were administering surveys. The MT DCC had the highest proportion of first exchanges that resulted in a codable answer (62.56 percent), followed by NOD (58.90 percent), AR (54.66 percent), and WY (54.50 percent). The OK DCC had the lowest proportion of first exchanges result in a codable answer (46.83 percent).

Table 4.5b: Respondent Behaviors Overall in the 1st Exchange: By DCC

DCC	*AR*	*MT*	*NOD*	*OK*	*WY*
Respondent Behavior	**Percent**	**Percent**	**Percent**	**Percent**	**Percent**
CA	54.66	62.56	58.90	46.83	54.50
CLAR	6.83	5.35	4.71	4.17	5.92
DK	2.80	0.70	0.26	0.79	0.95
SC	11.18	7.67	5.50	20.04	15.40
INTERRPT	1.24	0.47	0.26	0.60	1.18
INTRO	0	0	0.26	0	0.71
QA	5.28	6.28	6.54	8.33	5.92
REF	0	0	0.79	0	0
VERCORR	0	0	1.05	0.40	0
VERNORES	4.97	3.02	5.24	5.56	2.37
OTHR	3.11	2.56	4.45	4.76	3.08

Notes: The overall Chi-square of Respondent Behaviors in the 1st Exchange by DCC = 134.18 ($p < 0.0001$). The Chi-square for DCC*1st Exchange (where 1st Exchange = “CA”) = 25.94 ($p < 0.0001$).

The distribution of respondent behaviors in the final exchange were significantly different among DCCs (overall chi-square of Table 4.5c = 153.82, $p < 0.0001$). Providing a response that satisfied the meaning of the question (CAFR) in the final exchange (chi-square = 35.77, $p < 0.0001$) was significantly associated with the DCC where interviewers were administering surveys. The MT DCC had the highest proportion of final exchanges that resulted in a codable answer (72.33 percent), followed by NOD (67.80 percent), AR (64.60 percent), and WY (63.98 percent). The OK DCC had the lowest proportion of final exchanges result in a codable answer (54.37 percent).

Table 4.5c: Respondent Behaviors Overall in the Final Exchange: By DCC

DCC	*AR*	*MT*	*NOD*	*OK*	*WY*
Respondent Behavior	**Percent**	**Percent**	**Percent**	**Percent**	**Percent**
CAFR	64.60	72.33	67.80	54.37	63.98
DKFR	2.80	0.23	0	0.40	0.71
SCFR	11.18	8.14	5.76	20.24	15.40
INTROFR	0	0	0.26	0	0.71
QAFR	3.73	2.79	3.14	6.94	4.03
REFFR	0	0	0.79	0	0
VERCORRF	0	0	1.05	0.20	0
VERNORESF	5.28	3.02	5.24	5.56	2.61
OTHFR	2.48	2.09	3.93	3.17	2.61

Notes: The overall Chi-square of Respondent Behaviors in the Final Exchange by DCC = 153.82 (p < 0.0001). The Chi-square for DCC*Final Exchange (where Final Exchange = "CAFR") = 35.77, p < 0.0001.

The distributions of the number of exchanges between the interviewer and respondent needed to arrive at a final answer differed significantly among DCCs (overall chi-square of Table 4.5d = 65.34, p < 0.0001). Furthermore, the proportion of the time only one exchange was needed between the interviewer and respondent to arrive at an answer that satisfied the intent of the question was significantly associated with the DCC where interviewers were performing data collection (chi-square = 26.66, p < 0.0001). The DCCs most successful at this measure of data quality were the MT DCC (70.23 percent), NOD (69.37 percent), WY (68.72 percent), and AR (65.84 percent).

Table 4.5d: Number of Exchanges between the Interviewer and Respondent: By DCC

DCC	*AR*	*MT*	*NOD*	*OK*	*WY*
Number of Exchanges	**Percent**	**Percent**	**Percent**	**Percent**	**Percent**
One exchange	66.04	70.23	69.55	57.00	68.88
Two exchanges	12.42	9.07	8.90	10.71	7.82
Three or more exchanges	8.70	4.42	8.38	3.97	6.16
Not applicable (because interviewer behavior = "SC")	12.73	16.28	13.09	27.98	17.06

Notes: The overall Chi-square of Number of Exchanges by DCC = 65.34 (p < 0.0001). The Chi-square for DCC*Number of Exchanges (where Number of Exchanges = "One Exchange") = 26.66, p < 0.0001.

5. Conclusion

This report focused on several factors that could have major impacts on standardized data collection in the CATI mode of the Agricultural Labor Survey: 1) the DCCs where interviews are being conducted; 2) varying versions of the questionnaire (one with experimental questions and a control using the original questionnaire); 3) questions battery (series of questions for October vs. series of questions for July); 4) the specific questions or question themes in the survey; and 5) encouraging adherence to the interviewing script between the April-October iterations of data collection. Using behavior coded data of recorded CATI interviews conducted across various DCCs, many significant results highlighting the impact to standardization elicit both areas of encouragement and areas for improvement in the data collection process for this survey. In general, standardization of question administration improved from 12 percent to 53 percent between the 2018 April and October versions of the survey after encouraging interviewer adherence to the interviewing script. While this is encouraging, the results show there is still room for improvement in standardized administration overall. Early studies on interviewer behaviors found that the proportion of questions read exactly as worded in surveys could be as low as 30 percent (Oksenberg 1981) and as high as 96 percent (Mathiowetz and Cannell 1980). The overall standardized rate of 53 percent in this report puts the 2018 October Agricultural Labor Survey at the lower end of this distribution.

As seen in Table 4.2a in the results section, standardization behaviors were significantly associated with the version of the questionnaire being administered, but that seemed to be driven mostly by the difference in verification behaviors between the two versions. For instance, reading questions exactly as worded was not significantly associated with questionnaire version, suggesting interviewers performed equally well on this metric between the two versions. However, respondents' abilities to provide responses that satisfied the intent of the questions in the first exchange significantly improved in the experimental version of the survey compared to the control. The battery of questions for each quarter also mattered. Both interviewer performance and respondent performance were significantly worse in the battery of questions for the second quarter that was asked (the set of questions for July) compared to the first quarter (the set of questions for October). This suggests that burden may be a factor here, since data quality is worse when the second quarter of questions are introduced. This may point to the need to shorten the survey and only ask about the most recent quarter.

Interviewer standardization behaviors also varied significantly by question and question theme. Respondent behaviors were similarly affected. More instances of interviewer standardization and codable answers from respondents occurred for questions that appeared in the beginning and end of the survey. One explanation for this could be that the questions in the middle are more burdensome for interviewers to read and for respondents to process cognitively and answer than the questions at either end of the survey. Meaning, it behooves the survey and questionnaire designers to re-examine these questions to try to come up with alternative designs that are less burdensome. Doing so may help to increase interviewer adherence to the survey script and respondents' abilities to provide answers that satisfy the intent of the question in the first exchange.

The DCC where interviewers were conducting surveys also mattered. Desirable interviewer and respondent behaviors were significantly associated with the DCCs and suggest some DCCs are performing standardized interviewers at much higher rates than other DCCs. However, as all the DCCs still have room for improvement when it comes to standardization, more interviewer training focusing on the importance of standardization would still be prudent and necessary for all DCCs. One challenge in trying to encourage more adherence to the interviewing script in this survey is buy-in from DCCs and interviewers that have a lot of experience administering surveys to the agricultural population. In their experience, this population tends to be more receptive to conversational interviewing and rapport building, and these interviewing styles result in better outcomes (e.g., greater response rates). One way to address this would be to incentivize DCCs and interviewers for data quality, and not just response rates.

One way NASS survey designers tried to address this in the 2018 October survey was by including a sentence for interviewers to read to respondents in the introduction of the survey that informed them that all questions would read exactly as worded, and that even though it may seem repetitive at some points, that was what the interviewer was tasked to do. This type of forecasting is akin to what Fowler and Mangione (1990) recommended for setting the stage for standardized interviews. Their argument was that if interviewers explain to respondents why questions are going to be asked exactly as worded, then respondents will do a better job in a standardized interview interaction. The results in this report seem to support this assertion, as instances of respondents providing codable answers in the first and final exchanges improved significantly in the October 2018 data collection after including the standardized forecasting text in the introduction of the survey and after interviewers had been retrained on the importance of reading questions as worded.

The call for more standardized interviewing behaviors is not new. In fact, a synthesis of research on the topic showed that the proportion of survey questions read exactly as worded by interviewers varied widely in the survey methodology literature (Groves 1989). The literature has also shown that even slight wording changes can have major effects on aggregate data distributions (Willis 2005; Bradburn and Sudman, 1991; Schuman and Presser 1981; Sudman and Bradburn, 1982). Other research has shown that interviewers who deviate from the instructions given for administering a questionnaire may increase response errors in various ways. For instance, non-standardized reading of the questions may lead to increases in the number of non-substantive responses in the first exchange. Non-substantive responses in the first exchange may require additional probing. If interviewer behaviors, like question reading and probing, vary across interviewers, the intra-interviewer correlation could increase, thus increasing the variance of descriptive estimates and reducing the effective sample size of the survey (West and Blom 2017; Groves 2004). For these well-documented reasons, it is imperative that NASS continue to dedicate resources toward improving standardization in interviewer-administered surveys.

5.1 Recommendations

The findings in this report highlight the need for NASS to follow several recommendations:

1. Dedicate additional resources for interviewer training on conducting standardized interviews.

2. Dedicate additional resources for CATI instrument and questionnaire improvements.
    a. Decrease cognitive burden of the labor questions over the phone.
    b. Increase the usability design (UX) of the Blaise CATI instrument for this survey.

3. Change from a biannual survey to a quarterly survey.

4. Increase buy-in for standardization from all the Data Collection Centers (DCCs).
    a. All DCCs should be relatively equal in interviewer standardization rates.

6. References

Biagas, D., Abayomi, E., Rodhouse, J., and Ridolfo, H. (2019), "Examining Interviewer Effects on the Agricultural Labor Survey: A Mixed Methods Approach," Poster presented at the *Interviewers and Their Effects from a Total Survey Error Perspective Workshop*.

Bradburn, N. M., and Sudman, S. (1991), "The Current Status of Questionnaire Research," in Measurement Errors in Surveys, eds. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, New York: John Wiley & Sons, 29-40.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20, 37-46.

Fowler, F. J., Jr., and Mangione, T. W. (1990), *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*, Newbury Park, CA: Sage.

Fowler, F. J., Jr., & Cannell, C. F. (1996), "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions," in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, eds. N. Schwarz and S. Sudman, San Francisco, CA: Jossey-Bass, 15-36,

Fowler, F. J., Jr. (2011), "Coding the Behavior of Interviewers and Respondents to Evaluate Questions," in *Question Evaluation Methods: Contributing to the Science of Data Quality,* eds. J. Madans, K. Miller, A. Maitland, and G. Willis, Hoboken, NJ: John Wiley & Sons, 5-21.

Groves, R. M. (1989), *Survey Errors and Survey Costs,* Hoboken, NJ: John Wiley & Sons.

Groves, R. M. (2004), "The Interviewer as a Source of Survey Measurement Error," in *Survey Errors and Survey Costs* (2nd ed.), New York: Wiley-Interscience.

Groves, R. M., and Kahn, R. L. (1979), *Surveys by Telephone*, New York, Academic Press.

Groves, R. M., and Magilavy, L. J. (1986), "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys," *Public Opinion Quarterly*, 50, 251-266.

Landis, J., & Koch, G. (1977), "The Measurement of Observer Agreement for Categorical Data," *Biometrics,* 33(1), 159-174.

Mathiowetz, N., and Cannell, C. (1980), "Coding Interviewer Behavior as a Method of Evaluating Performance," *Proceedings of the American Statistical Association*, pp. 525-528.

Oksenberg, L. (1981), *Analysis of Monitored Telephone Interviews*, Report to the U.S. Bureau of the Census for JSA 80-83, Ann Arbor, Survey Research Center, The University of Michigan, May, 1981.

Oksenberg, L., Cannell, C., and Kalton, G. (1991), "New Strategies for pretesting survey questions," *Journal of Official Statistics*, 7(3), 349-365.

Ongena, Y. P., & Dijkstra, W. (2006), "Methods of behavior coding of survey interviews," *Journal of Official Statistics*, *22*(3), 1-34.

Rodhouse, J., Ridolfo, H., Abayomi, E. J. (2019), "Does Encouraging Adherence to the Interviewing Script Improves Estimates in a Complex Survey?" Poster presented at the Interviewers and Their Effects from a Total Survey Error Perspective Workshop.

Ridolfo, H., Biagas, D., Abayomi, E. J., and Rodhouse, J. (2020), "Behavior Coding of the October 2017 Agricultural Labor Survey," Washington, DC: National Agricultural Statistics Service.

Ridolfo, H., Biagas, D., Abayomi, E. J., and Rodhouse, J. (2021), "Behavior Coding of the April 2018 Agricultural Labor Survey," Washington, DC: National Agricultural Statistics Service

Schuman, H., and Presser, S. (1981), *Questions and Answers in Attitude Surveys*, New York, NY: Academic Press.

Sloan, R. (2017), "Agricultural Labor Survey Cognitive Interview Report," Washington, DC: National Agricultural Statistics Service.

Sudman, S., and Bradburn, N. (1973), "Effects of Time and Memory Factors on Response in Surveys," *Journal of the American Statistical Association*, 68(344), 805-815.

Sudman, S., and Bradburn, N. (1982), *Asking Questions. A Practical Guide to Questionnaire Design*. San Francisco, CA: Jossey-Bass.

West, B. T., and Blom, A. G. (2017), "Explaining Interviewer Effects: A Research Synthesis," *Journal of Survey Statistics and Methodology*, 5, 175-211.

Willis, G. B. (2005), *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA.: Sage Publications.