Ridolfo, Heather, and Ashley Thompson. 2023. "Conducting Small-Scale Multimethod Questionnaire Evaluation." *Survey Practice* 16 (1). https://doi.org/10.29115/SP-2023-0017.

ARTICLES

Conducting Small-Scale Multimethod Questionnaire Evaluation

Heather Ridolfo[1], Ashley Thompson[2]

[1] U.S. Energy Information Administration, [2] National Agricultural Statistics Service

Keywords: multimethod testing, cognitive interviewing, usability testing, behavior coding, establishment survey

https://doi.org/10.29115/SP-2023-0017

Survey Practice

Vol. 16, Issue 1, 2023

When pretesting survey questionnaires, there are benefits to using a multimethod approach. However, using multiple methods can be cost prohibitive. In 2021, the National Agricultural Statistics Service (NASS) received feedback from stakeholders regarding concerns about double counting of grain stocks in the Agricultural Survey and the Grain Stocks Report. Data from these two surveys are used to produce the Grain Stocks publication, which is a principal Federal Economic Indicator publication. To address stakeholder feedback, NASS evaluated these two multimode surveys, in a short timeframe, with limited resources. Despite these limitations, NASS utilized expert review, cognitive testing, usability testing and behavior coding to evaluate these two survey questionnaires. This paper demonstrates that multimethod pretesting can be done effectively on a low-cost, small-scale basis.

INTRODUCTION

In questionnaire evaluation research, cognitive testing is often viewed as the gold standard methodology. It is a generally low-cost method that can reliably predict problems in the field (Tourangeau et al. 2019), and as a result, it is often the default method chosen when evaluating new and existing survey questions. However, there are benefits to conducting a multimethod evaluation. Multiple methods can be used to draw on the strengths of different methodologies in providing corroborating evidence of problems (Blair et al. 2007; Creswell and Poth 2018; Forsyth, Rothgeb, and Willis 2004; McCarthy et al. 2018) or in producing different, but complementary findings (Maitland and Presser 2018; Presser and Blair 1994; Tourangeau et al. 2019; Willimack et al. 2023). Selection of an evaluation method is often driven by the time and costs associated with the method (Tourangeau et al. 2019). Conducting a multimethod evaluation is typically expensive and can require multiple years of testing (McCarthy et al. 2018; Tuttle, Morrison, and Galler 2007), which can prohibit its use.

In 2020, the National Agricultural Statistics Service (NASS) received feedback from stakeholders regarding the validity of data collected for the *Grain Stocks* publication. This is a Principal Federal Economic Indicator publication of grain and oilseed stocks stored in each state and by position (on and off-farm) (National Agricultural Statistics Service 2023). Data for this publication are collected from two multimode surveys, with two different populations. At the time, NASS had limited staff and budget for testing. Additionally, travel was prohibited due to the COVID-19 pandemic. Although it was not feasible to do a large-scale multimethod evaluation, NASS still recognized the benefit of using multiple methods to address stakeholder concerns.

To determine which methods to use, NASS considered three factors: the goals of the research, survey modes, and the costs associated with different evaluation methods. Two surveys are used to produce the *Grain Stocks* publication: the Agricultural Survey and the Grain Stocks Report (GSR). The Agricultural Survey is a sample survey of farm producers and collects data on crop acreage, yield and production, and quantities of grains stored on farms. The GSR is a census of commercial facilities with rated storage capacity, which measures grains stored off-farms. Stakeholders were concerned that stored grain was being double counted in these two surveys. If double-counting were occurring, it would be more likely to be due to misreporting in the Agricultural Survey. However, NASS felt it was important to evaluate the GSR as well.

It is important to consider the survey mode(s) when evaluating questionnaires as different modes can contribute to different types of response error. Data for the Agricultural Survey are primarily collected via paper and computer-assisted telephone interviews (CATI), whereas data for the GSR are primarily collected via the web and electronic submission of files. Cognitive testing is an effective method for identifying semantic problems in survey questions in any mode, but it is not effective at identifying interviewer problems, which are more likely to be uncovered using behavior coding (Presser and Blair 1994). In web surveys, visual design and functionality (e.g., edit checks) can impact respondents' understanding of the survey questions and their ability to provide valid responses, and usability testing is an effective method for evaluating this mode (Hansen and Couper 2004; Geisen and Romano Bergstrom 2017). Combined usability and cognitive testing is an efficient way to identify issues with questionnaire wording and design (Romano Bergstrom et al. 2013).

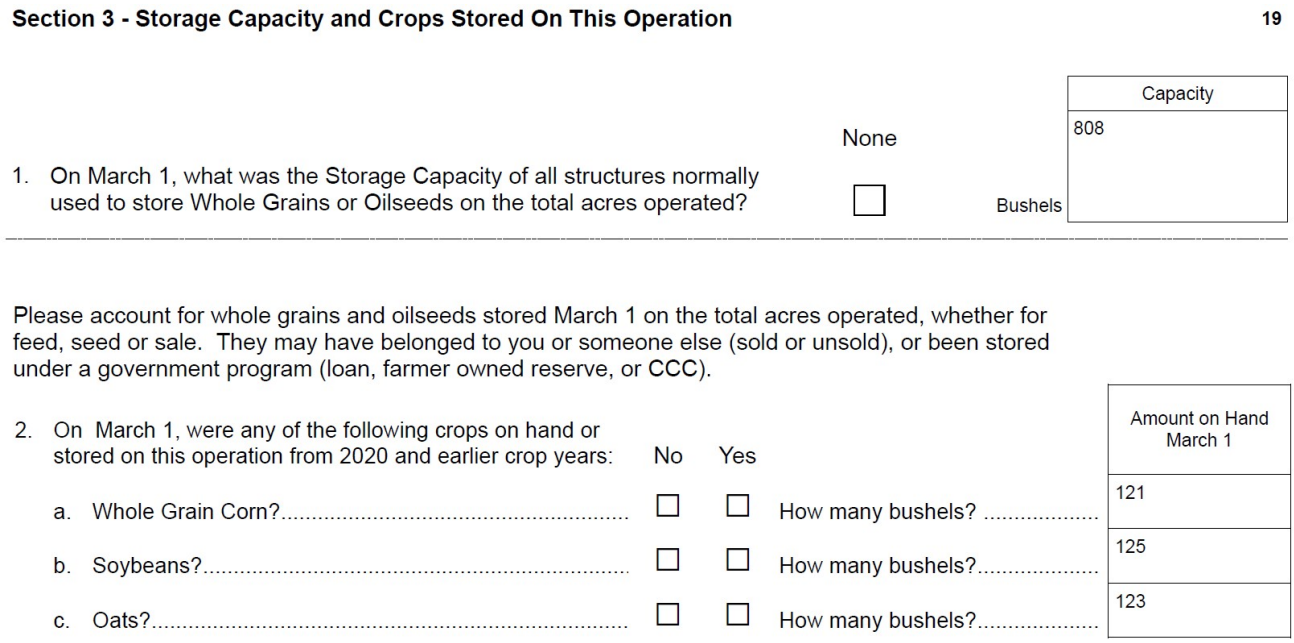
NASS had resources in place that made it cost effective to select methodologies that would permit questionnaire evaluation of different survey modes. Prior to the pandemic, NASS established procedures for conducting cognitive and usability testing remotely. Additionally, all CATI calling is recorded in NASS call centers, making behavior coding a feasible option. Four methods were selected to evaluate these two surveys: cognitive testing, usability testing, behavior coding, and expert review. The following paper will demonstrate the benefits of this small-scale multimethod survey evaluation.

METHODS

Agricultural Survey

COGNITIVE INTERVIEWING

In March 2021, 16 cognitive interviews were conducted remotely with farms that had on-farm grain storage. Testing focused on the *Storage Capacity and Crops Stored on This Operation* section (see Figure 1). Interviewers used retrospective probing to understand respondents' response processes (Willis 2005). Data were analyzed using the constant comparative method of analysis (Strauss and Corbin 1990).



Section 3 - Storage Capacity and Crops Stored On This Operation 19

	None		Capacity
1. On March 1, what was the Storage Capacity of all structures normally used to store Whole Grains or Oilseeds on the total acres operated?	☐	Bushels	808

Please account for whole grains and oilseeds stored March 1 on the total acres operated, whether for feed, seed or sale. They may have belonged to you or someone else (sold or unsold), or been stored under a government program (loan, farmer owned reserve, or CCC).

2. On March 1, were any of the following crops on hand or stored on this operation from 2020 and earlier crop years:	No	Yes		Amount on Hand March 1
a. Whole Grain Corn?	☐	☐	How many bushels?	121
b. Soybeans?	☐	☐	How many bushels?	125
c. Oats?	☐	☐	How many bushels?	123

Figure 1. Agricultural Survey Section 3

BEHAVIOR CODING

Behavior coding was performed using recordings from the December 2019 and March 2020 Agricultural Surveys. Due to the closing of centralized call centers during the COVID-19 pandemic, these were the most current recordings available. Twenty interviews from the December 2019 and 40 interviews from the March 2020 Agricultural Surveys were selected. Questions in the *Storage Capacity and Crops Stored on This Operation and Unharvested Crops* section were coded, resulting in the coding of 201 question administrations for December and 333 for March.

Six possible codes were assigned to the interviewers' behavior (see Table 1). The codes *failure to verify*, *major change*, and *question omitted* were considered problematic interviewer behavior. Ten possible codes were assigned to respondent behavior (see Table 2). Problematic respondent behavior included the codes *qualified answer, clarification, verification no response, incorrect format, interrupted, response to intro text,* and *refused*. Problematic behavior that occurs more than 15% of the time is indicative of a problem (Fowler 2011). Two researchers trained in behavior coding coded the interviews. Cohen's kappa was calculated to ensure consistency across coding. The overall kappa was 0.9394, indicating there was substantial agreement between the two coders (Landis and Koch 1977).

GSR

COMBINED USABILITY AND COGNITIVE TESTING

In July 2021, NASS conducted five remote combined usability and cognitive testing interviews with commercial grain facilities. During the interviews, interviewers observed respondents completing the web survey and used

Table 1. Interviewer Codes

Code	Description
Exact Wording	Question asked as worded or with minimal changes.
Verified	Response is verified based on information provided earlier in the interview.
Major Change	Question asked with major changes that can alter interpretation of the question.
Failure to Verify	Information is provided earlier in interview but interviewer records response without verifying.
Question Omitted	Interviewer records response without asking a question or verifying information.
Other	Behavior that does not fit under other interviewer codes.

Table 2. Respondent Codes

Code	Description
Codable Answer	Response appears plausible and matched the response options format.
Qualified Answer	Respondent expressed uncertainty or qualifies his/her response.
Clarification	Respondent requests clarification.
Verification Corrected	Respondent corrects verification.
Verification No Response	Respondent does not confirm or correct verification by interviewer.
Incorrect Format	Response does not match response options format.
Interrupted	Respondent interrupted interviewer while she/he was reading the question.
Response to Intro Text	Respondent provides a response to introductory text.
Refused	Respondent refused to answer question.
Other	Behavior that does not fit under other respondent codes.

retrospective probing to understand respondents' response processes (Willis 2005). Cognitive interview data were analyzed using the constant comparative method (Strauss and Corbin 1990).

REVIEW OF ELECTRONIC SUBMISSION OF FILES

Methodologists met with statisticians in the regional field offices (RFOs) to learn more about the electronic submission process. During these meetings, methodologists gained insight into the need for this mode and reviewed files submitted electronically to NASS and procedures used to process these data.

RESULTS

Agricultural Survey

COGNITIVE TESTING RESULTS

Given that grain and oilseeds are moved off-farm throughout the year for sale or storage in off-farm facilities, respondents must comprehend and adhere to the instructions and key clauses in the questions regarding the reference date and the type and location of the storage facilities to provide accurate answers. In cognitive testing, no major issues were found with the comprehension of survey questions. Respondents reported on-farm grain storage capacity and quantities stored, and properly excluded grain stored off-farm in commercial facilities. Although all respondents understood the questions in this survey as asking about grain stored on-farm, some respondents indicated that the phrase "on hand or stored" in the grain storage question could be interpreted as stored on- and off-farm and having the phrase "on this operation" was essential to

understanding the question. Respondents adhered to the reference date in the questions, but some noted that their answer would be different had they not. Finally, a few respondents did not read the introductory text and incorrectly omitted grain that was stored for others on their operation.

Cognitive testing demonstrated that respondents were interpreting the survey questions as intended. It also demonstrated the importance of reading survey questions in full. If response error was present in the survey data, it was not likely due to question wording. However, a limitation of cognitive testing is that it does not reflect how respondents complete questionnaires during production. Establishment respondents are often responding to surveys during working hours, when they have competing demands and limited time. Additionally, the cognitive testing was only performed on the paper questionnaire and given that a large portion of the data are collected via CATI, an examination of the CATI questionnaire was warranted. Cognitive testing of CATI questionnaires is not always ideal as cognitive interviewers are trained to read survey questions exactly as worded, which may not occur in production. Behavior coding provided an opportunity to evaluate the CATI instrument and substantiate or identify different problems with the questions during production (Fowler 2011).

BEHAVIOR CODING RESULTS

In December 2019 and March 2020, interviewers made major changes to survey questions in half of the question administrations and omitted entire survey questions at high rates (21% and 25%, respectively). NASS interviewers are trained to use conversational interviewing. Therefore, it is not surprising that interviewers did not precisely adhere to the survey script. However, the behavior coding revealed that key phrases in the survey questions, such as the reference date and "on the acres operated" were often omitted. Given the interviewers were using conversational interviewing, it is possible they established that the survey would be asking about grain stored on-farm on the reference date before asking the questions. However, a closer examination revealed that in some interviews, interviewers never read these keys phrases in any of the question administrations, nor were these criteria conveyed earlier in the survey (20% of interviews in 2019 and 33% of interviews in 2020). Despite the high levels of problematic interviewer behavior, there was little indication that respondents were having difficulty answering the survey questions. The only indication of a problem was in December 2019, where respondents provided a response in an incorrect format at a high rate (30%). This was due to interviewers changing crop storage questions from open-ended numeric questions to yes/no questions.

Behavior coding provided an opportunity to evaluate the CATI questionnaire, performance of the questionnaire in a production setting, and interviewer effects. The behavior coding revealed no respondent issues; however, a limitation of behavior coding is that it can be difficult to identify respondent

issues when responses appear plausible, as was the case in this study. The large number of changes interviewers made to the survey questions was concerning; however, interviewer changes to survey questions do not necessarily lead to response error (Dykema, Lepkowski, and Blixt 1997). Additional research would be needed to determine whether interviewer changes led to response error in the survey data. Additionally, without talking to the interviewers, it is difficult to know why they made certain changes to the survey questions.

GSR

COMBINED USABILITY AND COGNITIVE TESTING

No usability issues were found with the GSR web instrument. Respondents navigated the instrument, read the survey instructions and questions, and entered and submitted their responses with no issues. Respondents felt the instructions were clear and took the time to read them before proceeding through the survey questions, and some returned to the instruction screen when questions arose. No major comprehension issues were found with the survey questions, and most respondents indicated they refer to their records when responding to the survey.

Conducting combined usability and cognitive testing allowed NASS to efficiently evaluate both the user experience with the web questionnaire and respondents' comprehension of the survey questions. Recruitment for this population was challenging and only five interviews were completed. Five interviews are sufficient for usability testing (Geisen and Romano Bergstrom 2017), and cognitive interviewing best practices state that a minimum of five interviews should be completed (Willis 2005) or interviews should be conducted until saturation is reached (U.S. Office of Management and Budget 2016). Given that no new issues were uncovered after five interviews, we felt these interviews provided sufficient information on the performance of the GSR. However, combined testing limits the number of cognitive probes that can be asked, and additional cognitive issues may have been uncovered had more interviews been conducted.

REVIEW OF ELECTRONIC SUBMISSION OF FILES

The review of the GSR electronic submission procedures revealed that some large businesses prefer to submit data exports from their records rather than completing survey questionnaires. Often these businesses are reporting for multiple grain storage facilities across multiple states and would need to complete a separate questionnaire for each facility. Files submitted vary in format (e.g., pdfs, Excel spreadsheets) and in content (e.g., different column labels, omission of data) or specificity of data (e.g., unit of measurement). RFO staff hand-edit the data and perform manual data entry. Data processing procedures varied slightly across the RFOs.

The expert review of the GSR electronic submission procedures provided insight into challenges large entities face when responding to the GSR and potential measurement error associated with processing electronic submission of files; however, no direct feedback from respondents was gathered. Interviews with respondents would have been beneficial to better understand the challenges large businesses face when responding to the GSR and to identify ways to ease burden while collecting more standardized data.

DISCUSSION

In 2020, stakeholders raised concern that stored grain was being doubled-counted in the Agricultural Survey and GSR. If double-counting was occurring, it was likely due to misreporting in the Agricultural Survey. Therefore, it would have been easy to limit this evaluation to cognitive testing of the Agricultural Survey. However, given the numerous documented advantages of multimethod testing, NASS decided to evaluate these two surveys using multiple methods with the limited resources available, allowing issues that would not have been discovered with cognitive testing alone to be uncovered and ultimately providing a better understanding of potential sources of measurement error in these two surveys.

Cognitive testing of the two surveys revealed that both surveys were performing well, and respondents understood that grain stored on-farm was to be reported on the Agricultural Survey and grain stored off-farm was to be reported on the GSR. Usability testing revealed no issues with the GSR web form. However, additional testing revealed that error may be present in the other modes. Behavior coding revealed interviewers made major changes to the Agricultural Survey in CATI, which may result in respondents reporting grain that has been moved off-farm either for storage or sale, possibly inflating estimates of grain stored on farm. The expert review of the electronic data submissions revealed that businesses that report for multiple facilities prefer to submit electronic records rather than use the web questionnaire. Submission of non-standardized data files opens the possibility for the introduction of measurement and processing errors.

Based on these findings, several recommendations were made to improve data collection, including improvements to the Agricultural Survey CATI instrument to ease interviewer administration, retraining and monitoring of interviewers and developing standardized ways for businesses to report for multiple facilities at one time in the GSR, whether it be modifying the web instrument or providing standardized templates for data uploads.

There were a few limitations to this research. Although behavior coding revealed significant issues with the administration of the CATI survey, without talking to interviewers, there is no way to know for sure why major changes were made to the survey questions. Interviewer debriefings could provide much needed insight to improving the instrument and interviewer training. It is also not possible to know if interviewer deviation in the Agricultural Survey

led to response error. Additional research, such as a reinterview study could help determine this. It may have been beneficial to conduct more cognitive interviews with GSR respondents to ensure no major cognitive problems existed and to better understand how the web form could be improved for large operations.

NASS had the benefit of having several resources in place that helped limit the cost of this research. Prior experience with remote testing and the standard practice of recording CATI interviews allowed NASS to easily employ cognitive testing, usability testing, and behavior coding at no additional cost. Others who do not have these practices in place may still find multimethod testing cost prohibitive. In these cases, methodologists may want to consider using multiple lower costs methodologies such as expert review and ex-ante methodologies (Maitland and Presser 2018).

In summary, this research provides further evidence of the benefit of using multiple methods and demonstrates that it can be done on a small scale with lower costs. Multimethod pretesting allowed NASS to produce different but complementary findings and ultimately a better understanding of possible sources of measurement error in these two surveys.

Lead author

Heather Ridolfo, Office of Statistical Methods and Research, U.S. Energy Information Administration, (202) 586-6240, Heather.Ridolfo@eia.gov

Disclaimers

The lead author conducted this research while employed for the National Agricultural Statistics Service.

The analysis and conclusions contained in this paper are those of the authors and do not represent the official position of the U.S. Energy Information Administration (EIA) or the U.S. Department of Energy (DOE).

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

Submitted: July 21, 2023 EST, Accepted: September 11, 2023 EST

REFERENCES

Blair, Johnny, Allison Ackermann, Linda Piccinino, and Rachel Levenstein. 2007. "Using Behavior Coding to Validate Cognitive Interview Findings." In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 3896–3900. Anaheim, CA.

Creswell, John W., and Cheryl N. Poth. 2018. *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Thousand Oaks, CA: Sage Publications.

Dykema, Jennifer, James M. Lepkowski, and Steven Blixt. 1997. "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study." *Survey Measurement and Process Quality*, February, 287–310. https://doi.org/10.1002/9781118490013.ch12.

Forsyth, Barbara, Jennifer M. Rothgeb, and Gordon B. Willis. 2004. "Does Pretesting Make a Difference? An Experimental Test." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, and M. P. Couper, 525–46. Hoboken, NJ: John Wiley and Sons. https://doi.org/10.1002/0471654728.ch25.

Fowler, Floyd J., Jr. 2011. "Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions." In *Question Evaluation Methods: Contributing to the Science of Data Quality*, edited by J. Madans, K. Miller, and A. Willis, 5–21. https://doi.org/10.1002/9781118037003.ch2.

Geisen, Emily, and Jennifer Romano Bergstrom. 2017. *Usability Testing for Survey Research*. Cambridge, MA: Morgan Kaufmann Publishers.

Hansen, Sue Ellen, and Mick P. Couper. 2004. "Usability Testing to Evaluate Computer-Assisted Instruments." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer, 337–60. https://doi.org/10.1002/0471654728.ch17.

Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159. https://doi.org/10.2307/2529310.

Maitland, Aaron, and Stanley Presser. 2018. "How Do Question Evaluation Methods Compare in Predicting Problems Observed in Typical Survey Conditions?" *Journal of Survey Statistics and Methodology* 6 (4): 465–90. https://doi.org/10.1093/jssam/smx036.

McCarthy, Jaki S., Kathleen Ott, Heather Ridolfo, Pam McGovern, Robyn Sirkis, and Danna Moore. 2018. "Combining Multiple Methods in Establishment Questionnaire Testing: The 2017 Census of Agriculture Testing Bento Box." *Journal of Official Statistics* 34 (2): 341–64. https://doi.org/10.2478/jos-2018-0016.

National Agricultural Statistics Service. 2023. "Grain Stocks Methodology and Quality Measures." https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Grain_Stocks/01_2022/grstqm22.pdf.

Presser, Stanley, and Johnny Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology* 24: 73. https://doi.org/10.2307/270979.

Romano Bergstrom, Jennifer C. , Jennifer Hunter Childs, Erica Olmsted-Hawala, and Nathan Jurgenson. 2013. "The Efficiency of Conducting Concurrent Cognitive Interviewing and Usability Testing on an Interviewer-Administered Survey." *Survey Practice* 6 (4): 1–9. https://doi.org/10.29115/sp-2013-0022.

Strauss, Anselm C., and Juliet Corbin. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Newbury Park: Sage Publications.

Tourangeau, Roger, Aaron Maitland, Darby Steiger, and Ting Yan. 2019. "A Framework for Making Decisions About Question Evaluation Methods." In *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by P.C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G.B. Willis, and A. Wilmot, 47–73. https://doi.org/10.1002/9781119263685.ch3.

Tuttle, Alfred D., Rebecca L. Morrison, and David H. Galler. 2007. "From Respondent Debriefings to Pilot Test and Beyond: A Comprehensive Redesign of a Questionnaire Measuring Foreign Direct Investment." In *Proceedings of the Third International Conference on Establishment Surveys*. Montreal, Quebec.

U.S. Office of Management and Budget. 2016. "Addendum to the Statistical Policy Directive No. 2: Standards and Guidelines for Cognitive Interviews." Washington, DC: Office of Management and Budget. https://www.whitehouse.gov/wp-content/uploads/2021/04/final_addendum_to_stat_policy_dir_2.pdf.

Willimack, Diane K., Heather Ridolfo, Amy Anderson Riemer, Melissa Cidade, and Kathy Ott. 2023. "Advances in Question(Naire) Development, Pretesting, and Evaluation *." In *Advances in Business Statistics, Methods and Data Collection*, edited by G. Snijkers, M. Bavdaz, S. Bender, J. Jones, S. MacFeely, J. W. Sakshaug, K. J. Thompson, and A. van Delden, 387–411. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781119672333.ch17.

Willis, Gordon. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications, Inc. https://doi.org/10.4135/9781412983655.