sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 353–386, http://dx.doi.org/10.2478/JOS-2019-0017

Measuring Trust in Medical Researchers: Adding Insights from Cognitive Interviews to Examine Agree-Disagree and Construct-Specific Survey Questions

Jennifer Dykema[1], *Dana Garbarski*[2], *Ian F. Wall*[3], *and Dorothy Farrar Edwards*[4]

While scales measuring subjective constructs historically rely on agree-disagree (AD) questions, recent research demonstrates that construct-specific (CS) questions clarify underlying response dimensions that AD questions leave implicit and CS questions often yield higher measures of data quality. Given acknowledged issues with AD questions and certain established advantages of CS items, the evidence for the superiority of CS questions is more mixed than one might expect. We build on previous investigations by using cognitive interviewing to deepen understanding of AD and CS response processing and potential sources of measurement error. We randomized 64 participants to receive an AD or CS version of a scale measuring trust in medical researchers. We examine several indicators of data quality and cognitive response processing including: reliability, concurrent validity, recency, response latencies, and indicators of response processing difficulties (e.g., uncodable answers). Overall, results indicate reliability is higher for the AD scale, neither scale is more valid, and the CS scale is more susceptible to recency effects for certain questions. Results for response latencies and behavioral indicators provide evidence that the CS questions promote deeper processing. Qualitative analysis reveals five sources of difficulties with response processing that shed light on under-examined reasons why AD and CS questions can produce different results, with CS not always yielding higher measures of data quality than AD.

Key words: Agree-disagree questions; questionnaire design; cognitive interviewing; response processes; data quality; construct-specific questions.

[1] University of Wisconsin Survey Center (UWSC), 4308 Sterling Hall, 475 N. Charter St. Madison, WI 53706, U.S.A. Email: dykema@ssc.wisc.edu

[2] Loyola University Chicago, Coffey Hall 440, 1032 W. Sheridan Rd. Chicago, IL 60660, U.S.A. Email: dgarbarski@luc.edu

[3] Steelcase, 901 44th Street SE, Grand Rapids, MI, 49508, U.S.A. Email: ianfwall@gmail.com

[4] University of Wisconsin-Madison, 2176 Medical Science Center, 1300 University Avenue Madison, WI 53706, U.S.A. Email: dfedwards@education.wisc.edu

Acknowledgments: This study was funded by NIMHD (National Institute on Minority Health and Health Disparities) grant P60MD003428 (PD: A. Adams). Project: Increasing Participation of Underrepresented Minorities in Biomarker Research (PI: D. Farrar Edwards). Additional support was provided by the University of Wisconsin Survey Center (UWSC), which receives support from the College of Letters and Science at the University of Wisconsin-Madison, and the facilities of the Social Science Computing Cooperative and the Center for Demography and Ecology (NICHD core grant P2C HD047873). The authors thank: Tara Piché, Kenneth D. Croes, and Nathan Jones for their contributions in the development and analysis of the cognitive interviews; Nadia Assad, Jesus Renteria, and Ray Garza for research assistance; and Steven Blixt, Nora Cate Schaeffer, and John Stevenson for their comments on earlier drafts. Opinions expressed here are those of the authors and do not necessarily reflect those of the sponsors or related organizations.

1. Introduction

Questions that measure subjective constructs or evaluations historically have used an agree-disagree (AD) response format that presents respondents with a statement and asks them to indicate whether they agree or disagree with the statement or to rate their level of agreement. For example, the following question, administered for decades in the General Social Survey (GSS) is part of a scale designed to measure political efficacy: "The average citizen has considerable influence on politics. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?" (Smith et al. 2013).

While researchers have advocated for the positive psychometric properties of AD questions (see Willits et al. 2016), the ubiquity of these items primarily stems from their ease of use. Scales comprised of AD items are "easy to write" and efficient to administer; the same response categories can be used for each statement included in a battery of questions regardless of the content or complexity of the statement (Krosnick and Presser 2010). However, these positive features may be offset by increased burden for respondents and interviewers, which may ultimately lead to reductions in data quality. For example, AD questions may be more subject to response effects like acquiescence (the tendency to agree) or extreme responding (the tendency to select the lowest or highest response categories) (Krosnick and Presser 2010; Liu et al. 2015).

In recent writing, questionnaire designers eschew AD formats and advocate for construct-specific (CS) response formats (Fowler and Cosenza 2009; Krosnick and Presser 2010; Saris et al. 2010). Instead of asking participants to rate their level of agreement, CS questions directly ask about the item's underlying response dimension and provide construct-specific response categories. For example, a CS version of the GSS political efficacy question would be: "How much influence does the average citizen have on politics: none, a little, some, quite a bit, or a great deal?" The direct method of questioning offered by the CS format is argued to yield more reliable and valid data because it is less cognitively burdensome, less likely to be misinterpreted, and less likely to be associated with response effects.

In the current study, we use a mixed methods approach to evaluate the measurement properties of questions about trust in medical researchers using AD or CS questions. Trust is a central concept in the social and medical sciences because of its effect on decision-making and association with behavior. Trust is also a key component in social exchange theory, which posits that individuals are more likely to respond positively to a request to participate in research when they trust the originator of the request and perceive the ratio of rewards to costs to be personally acceptable (Dillman et al. 2014). Many have suggested that challenges recruiting and retaining research participants from underrepresented groups, such as racial and ethnic minorities, is rooted in a general distrust of medical researchers (Corbie-Smith et al. 2002; Scharff et al. 2010). Indeed, individuals with lower levels of trust indicate being less willing to participate in a future research study (Hall et al. 2006; Mainous et al. 2006; Braunstein et al. 2008). To better understand the public's trust in medical researchers, researchers need to measure the construct with sufficient reliability and validity. However, most scales use AD questions (e.g., Hall et al. 2006), which may lower data quality. Thus, we sought to improve on the measurement of trust in medical researchers by using CS questions.

1.1. Cognitive Processing of AD and CS Questions

Tourangeau et al. (2000) discuss four stages that a respondent progresses through in constructing an answer to a survey question, including comprehension of the question, retrieval of relevant information from memory to answer the question, use of retrieved information to make judgments, and selection and reporting of an answer. Researchers have expanded on this model to describe the unique cognitive steps required to answer an AD question (see Carpenter and Just 1975; Fowler and Cosenza 2009; Saris et al. 2010; Höhne and Lenzner 2018; Dykema et al. 2019).

Consider the response process embarked on by a respondent answering the AD question: "Medical researchers work very hard to make sure the participants in their studies are safe. Do you strongly agree, agree, neither agree nor disagree, disagree, strongly disagree." To answer this question, the respondent must first comprehend the literal and pragmatic meaning of the statement "Medical researchers work very hard to make sure the participants in their studies are safe" (Comprehension). Next, the respondent has to identify the question's underlying response dimension (Identification), which is the intensity of "working hard," (i.e., how hard medical researchers work to ensure the safety of research participants). Identification is accomplished by understanding the meaning of the statement as well as attending to any threshold words (e.g., "very") in the statement (Saris et al. 2010). Threshold words are those often included in AD questions that establish a threshold without presenting the full range of scale options. These include intensifiers (e.g., "extremely"), frequency markers (e.g., "rarely"), and quantifiers (e.g., "most"). After they identify the underlying response dimension, respondents must generate their own response or internal value to this response dimension (Generation). Here, our fictional respondent generates an internal value of "pretty hard" to the response dimension "how hard medical researchers work" and places that internal value on the underlying response dimension (Placement). The ensuing steps encompass a set of complicated cognitive processes in which the respondent evaluates the distance between their internal value of "pretty hard" and the threshold value of "very hard" (Threshold evaluation), and then assesses whether the distance between their internal value and the threshold value indicates "agreement," "disagreement," or "neutrality" (Polarity evaluation). Finally, guided by their evaluations of thresholds and polarity, the respondent must map their internal value onto one of the discrete categories offered in the "agreement" or "disagreement" range or select the midpoint if offered (Mapping).

The cognitive processing steps undertaken by a respondent answering the same item formatted in a construct-specific manner – that is, "How hard do medical researchers work to make sure that the participants in their studies are safe: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?" – is greatly simplified and predicted to be less burdensome. As with the AD version, the respondent must first comprehend the question (Comprehension). Next, they determine the underlying response dimension (Identification), which is reinforced by both the wording and ordering of the response categories (e.g., "not at all hard," "a little hard," etc.). Similar to processing with the AD question, the respondent generates an internal value of "pretty hard" (Generation), but placement of the internal value is done directly by mapping the internal value onto one of the discrete categories offered (Mapping), thereby circumventing the steps of Placement, Threshold evaluation, and Polarity evaluation.

Of course, the model for processing AD questions assumes respondents are optimally engaged with the task of responding and attentively progress through the steps. Höhne and colleagues (Höhne and Krebs 2018; Höhne and Lenzner 2018; Höhne et al. 2017) argue this may not be the case. Because AD questions are usually presented in multi-item batteries in which the wording of the statements vary but the response categories remain the same – always some form of agreement to disagreement – they encourage superficial processing. In contrast, when multiple CS questions are grouped together, they will likely use different construct-specific response categories, encouraging deeper processing and motivating effort. In support of this proposition, researchers demonstrated that respondents in an eye-tracking study attended to CS response categories more when they varied from question to question (Höhne and Lenzner 2018), but there were no differences in processing times between AD and CS questions when the questions were presented in grids in which the response categories did not vary for either question format (Höhne et al. 2017).

1.2. Experimental Evidence Comparing AD and CS Questions

Despite strong recommendations among questionnaire designers to use CS questions in lieu of AD questions, only a handful of experimental studies demonstrate CS questions yield higher data quality (Lelkes and Weiss 2015). The most compelling evidence is provided by Saris et al. (2010), who compared AD and CS response formats for items using split-ballot multitrait-multimethod (MTMM) designs. The experiments were conducted in face-to-face interviews that used show cards or self-administered questions across multiple countries in the European Social Survey (ESS). Overall, CS questions yielded higher estimates of reliability and validity. These findings were replicated by Revilla and Ochoa (2015), who also used split-ballot MTMM experiments and found much lower quality for AD than CS questions in data collected in Mexico and Columbia.

Dykema et al. (2012) assessed the measurement properties of items about political efficacy from the General Social Survey. Findings indicated a trend such that CS items were associated with higher internal consistency reliability (see also Hanson (2015) who reported higher test-retest reliability for CS items). This study also examined whether behaviors produced by interviewers and respondents varied by the format of the question, focusing on behaviors that were associated with lower data quality in prior research (e.g., interviewers misreading questions and respondents qualifying responses or saying “don’t know”) (Dijkstra and Ongena 2006; Dykema et al. 1997; Schaeffer and Dykema 2011). The authors found that AD items yielded more instances of interviewers misreading questions and more disfluency tokens (e.g., “um”). Kuru and Pasek (2016) used confirmatory factor analyses (CFA) and structural equation modeling within an experimental design about Facebook use and demonstrated a methodological bias due to AD response formats. Controlling for the method effect reduced reliability and validity estimates for the AD items, but not the CS items, indicating AD items may inflate reliability and regression estimates more than CS items. In validity tests, the criterion relationships yielded stronger relationships with the CS items. In addition, Höhne and

Krebs (2018) reported the AD response format was more susceptible to response scale direction effects than the CS format for internally-focused, self-administered questions about achievement and intrinsic job motivation, but not for externally-focused questions.

Other studies, however, have not reported greater data quality for CS questions. Lelkes and Weiss (2015) and Liu et al. (2015) both analyzed an experiment comparing AD and CS questions about political efficacy embedded in the American National Election Study. Lelkes and Weiss (2015) reported no differences in reliability or concurrent validity for the response formats, and neither format was more valid among those respondents susceptible to acquiescence. In addition, Liu et al. (2015) reported extreme response style was present for both AD and CS formats based on latent class factor analysis. Finally, in a web survey comparing questions presented stand-alone or in grids, Höhne et al. (2017) reported no differences between AD and CS questions for data quality as indicated by non-differentiation and dropping out of the survey before completion.

1.3. *Limitations of Past Comparisons between AD and CS Items*

Although many prior experimental studies provide evidence in support of CS items yielding more reliable and valid responses, most of the analyses are limited to a comparison of how AD and CS items differ with regard to the closed-ended responses they yield, leaving out potentially crucial information about the *response process* respondents undertake when answering questions. This information would be useful when designing and testing new questions. Dykema et al. (2012) began to look at the response process by examining interviewer and respondent behaviors. However, they were limited in their ability to provide clear insights about what characteristics of items may be most difficult or what aspects of the response process may cause cognitive difficulties because they could only examine behaviors produced during the process of answering standardized survey questions.

Comparing responses to AD and CS items and evaluating the response process with cognitive interviews in which respondents are asked to describe what they are thinking about while answering survey questions may prove fruitful for developing targeted approaches to improve data quality. In the current study, we evaluate both close-ended response tendencies as well as aspects of the response process using cognitive interviewing techniques. This approach allows us to incorporate quantitative and qualitative data in order to provide insight about *why* differences in response tendencies occur between AD and CS items.

1.4. *Current Study*

The goal of the current study is to provide an in-depth, mixed-methods analysis of a scale designed to measure the general public's trust in medical researchers. We randomized participants to either an AD or CS version of the scale, and conducted cognitive interviews that included follow-up questions designed to identify problems during the response process. This study is motivated by two main questions: (1) how do closed-ended survey responses differ between the versions of the scale, and (2) what aspects of the response process might explain differences in response tendencies? Based on the empirical research reviewed above, which suggests the CS response categories may be more demanding to process, more likely to encourage deeper processing, and often are associated with higher data quality, we predict:

- H_1: The CS scale will yield closed-ended responses with higher reliability than the AD scale.
- H_2: The CS scale will yield closed-ended responses with higher validity than the AD scale.
- H_3: The CS scale will be associated with greater recency effects than the AD scale.
- H_4: Responses to CS questions will yield longer processing times than AD questions.
- H_5: Responding to CS items will involve more instances of behavioral indicators of response difficulty (e.g., respondents providing uncodable responses).

We further explore the motivations for these hypotheses in Subsection 2.5 where we describe our measures and analytic strategy. Immediately following the analysis of data based on these quantitatively-focused predictions, we explore and leverage the qualitative data generated during cognitive testing to help illuminate why the AD and CS items behaved in predicted or unpredicted ways. The qualitative portion of the study is exploratory and not grounded in previous research, and we do not put forth hypotheses for these analyses.

2. Methods

2.1. *Overview of the Cognitive Interviewing Phase of the Voices Heard Survey Development*

We conducted cognitive interviews to evaluate questions for inclusion in the Voices Heard Survey, a telephone survey targeting members of minority groups underrepresented in biomedical research (Edwards 2015). The primary goal of the survey was to measure perceptions of the barriers and facilitators to participating in medical research studies that collect biomarkers (e.g., saliva and blood), and to document whether there were important differences among groups identified by their race and ethnicity. To develop questions for the survey, we first conducted key informant interviews to identify major themes around which to write questions (e.g., mistrust of medical researchers, logistical constraints, fear of discomfort and pain). Next, we tested the questions in two rounds of cognitive interviewing.

2.2. *Sample for the Cognitive Interviews*

We conducted 64 cognitive interviews, 32 in two rounds, from 2012 to 2013 using a community-based, quota sampling strategy to recruit participants. Members of the project team recruited participants through connections with leaders in specific racial and ethnic communities, by visiting churches and community centers, by attending events sponsored by groups (e.g., pow-wows held by several American Indian tribes), and by posting flyers at targeted locations in communities. We confined recruiting to southern Wisconsin. The quota strategy yielded equal numbers of African American, American Indian, Latino, and white participants, distributed nearly uniformly by gender (male versus female), age (between 30–55 years versus 56 years or more), and education (high school or less versus

Table 1. Descriptive statistics of participant characteristics and participation measures for the agree-disagree (AD) and construct-specific (CS) experimental groups.

	AD		CS			
Panel A: Participant characteristics	Proportion or Mean (S.D.)	n	Proportion or Mean (S.D.)	n	Test	p-value
Age						
30–55 years	0.47	15	0.53	17	$\chi^2 = 0.25$	0.80
56 years or more	0.53	17	0.47	15		
Female	0.50	32	0.50	32	$\chi^2 = 0.00$	1.00
Race/Ethnicity						
African American	0.25	8	0.25	8	$\chi^2 = 0.00$	1.00
American Indian	0.25	8	0.25	8		
White	0.25	8	0.25	8		
Latino	0.25	8	0.25	8		
Education						
High school or less	0.47	15	0.50	16	$\chi^2 = 0.06$	1.00
Some college or more	0.53	17	0.50	16		
Panel B: Participation measures						
Participated in medical research in past	0.28	32	0.35	31	$\chi^2 = 0.39$	0.53
Expressed likelihood to participate in research involving						
Answering questions	3.34 (0.79)	32	3.09 (0.86)	32	t = 1.22	0.23
Providing saliva	3.10 (1.19)	31	2.78 (1.24)	32	t = 1.03	0.31
Providing blood	2.97 (1.11)	31	2.35 (1.43)	31	t = 1.89	0.06
Providing tissue	2.27 (1.28)	30	1.84 (1.46)	31	t = 1.13	0.23
Providing cerebrospinal fluid	1.10 (1.25)	31	0.91 (1.35)	32	t = 0.85	0.56

Note: p-values for Chi-squared tests are from Fisher's exact tests.

some college or more) (see Panel A, Table 1). Interviews were conducted at locations that were convenient for the participants, including homes, libraries, and offices. Participants were remunerated for their time and effort.

2.3. Interviewing and Transcription

Following a format commonly employed in cognitive interviewing (Willis 2005; Fortune-Greeley et al. 2009; Willis and Miller 2011), interviewers asked participants a question being tested for use in the survey interview, and then after participants provided their response to the closed-ended survey question, interviewers administered a series of structured, open-ended probes and follow-up questions. We designed the probes and follow-up questions to uncover how participants formulated their answers to the survey questions, to reveal any problems they had with comprehension of specific terms or retrieval of information from memory, and to document issues participants faced in mapping their responses onto the response categories.

Interviewers received a full day of training on cognitive interviewing tailored for the study, and they were required to complete a practice interview before being certified.

We matched interviewers and participants on race/ethnicity and, for all cases except one, on gender. While interviews were primarily conducted in English, seven participants (distributed nearly evenly between the AD and CS conditions) elected to be interviewed in Spanish by a Spanish-speaking interviewer. On average, interviews took approximately an hour to complete (M = 61.10 minutes, SD = 20.17). Interviews were audio-recorded and digital files were created. In order to facilitate coding and analysis, interviews were transcribed verbatim on a question-by-question basis into Excel.

2.4. Trust/Mistrust Scale Development and Experimental Design

We examine respondents' answers and their response processes during administration of an 11-item scale measuring trust in medical researchers (see Appendix A, Subsection 6.1), for the exact wording of the items, which varied slightly between the two rounds of cognitive interviewing). We randomly assigned participants to the AD or CS scale using a between-subjects design. Items in a given scale appeared in the same sequence (i.e., they were not randomized), roughly 30 minutes into the interview.

To develop the scale, we conducted a literature review that identified approximately 100 questions from 12 studies about trust in medical care providers and researchers (Anderson and Dedrick 1990; Hayman et al. 2001; Corbie-Smith et al. 2002; Hall et al. 2002a; Hall et al. 2002b; Zheng et al. 2002; Thompson et al. 2004; Hall et al. 2006; Mainous et al. 2006; Egede and Ellis 2008; Henderson et al. 2008; Williams et al. 2010). The majority of the questions used AD response formats. From the pool of candidate questions, we modified 11 for the AD scale. We generated items for the CS scale by rewriting the AD version of the question to ask about the underlying response dimension implied by the question. For example, the AD item about informed consent ("hide information") asked, "*Medical researchers never hide information about the possible risks of participating. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?*" To translate to the CS format, we used the threshold word "never" to identify "frequency" as the underlying dimension: "*How often do medical researchers hide information about the possible risks of participating: never, rarely, sometimes, very often, or extremely often?*" Thus, the CS item directly asked participants for their evaluation of the relative frequency with which medical researchers' hide information, rather than having participants rate their level of agreement with a statement about medical researchers "never" hiding information.

AD items used the same five response categories for each item ("strongly agree" to "strongly disagree") but the CS items had response categories that varied depending on the underlying dimension (e.g., "never" to "extremely often" for frequency). The final scale was balanced and included a roughly equal number of positively and negatively valenced items. Positively valenced items are those in which a higher valued response category (e.g., "strongly agree" for AD questions and "a great deal" for CS) indicated most trust; negatively valenced items are those in which a higher valued response category indicated least trust. Thus, when the item is positively valenced, the direction of the response scale is ordered from most to least trust for the AD questions and least to most trust for the CS questions; when the item is negatively valenced, the scale is ordered from least to most trust for AD questions and most to least trust for CS questions.

2.5. Measures and Analytic Strategy

We present three sets of analysis using a mixed-methods approach (Johnson and Onwuegbuzie 2004): (1) measures of the quality of survey responses (e.g., reliability, validity, and recency); (2) response latencies and behavioral indicators of response difficulty; and (3) sources of response difficulties captured during the cognitive interviewing response process.

2.5.1. Trust Scale Summary Statistics and Reliability Measures

We examine trust scale summary statistics, including item nonresponse, mean trust scale scores, and reliability estimates. We assess the effect of item-missing data and other measures described below by estimating aggregate-level regression models that evaluate the items in a scale collectively by treating each question answered by the participant as a separate observation. Models estimate robust standard errors to correct for the fact that individual observations are independent across participants but dependent within a given participant (Rogers 1994). For item-missing values, estimates are from a logistic regression model with response format coded “1” for CS, “0” for AD. We score trust items from 0 to 4, with lower scores indicating less trust in medical researchers (e.g., depending on whether the item is positively or negatively valenced, “strongly disagree” may be coded as 0 or 4; see Appendix A, Subsection 6.1) and compute scale values by summing across the items. We impute cases with missing values with the median value for the non-missing cases on an item-by-item basis, separately for the AD and CS response formats (Hall et al. 2006). We evaluate internal consistency reliability using Cronbach’s alpha (Streiner et al. 2015), and test for significance by treating the alpha coefficients as correlations and applying Fisher’s r-to-z transformation (Tourangeau et al. 2004).

2.5.2. Concurrent Validity

Past research demonstrates a strong association between the public’s trust in medical researchers and their actual or expressed likelihood of participating in a medical research study (Hall et al. 2006; Mainous et al. 2006; Braunstein et al. 2008). We assess concurrent validity by examining whether the relationship between trust and participation is stronger for the AD or CS response format. Questions assess whether participants ever participated in medical research (coded “1” if “yes;” “0” if “no”) and their expressed likelihood of participating in medical research studies involving answering questions or providing samples of saliva, blood, tissue, and cerebrospinal fluid.

Because we were testing these questions for inclusion in a larger survey, their wording and response categories varied between rounds of interviewing, particularly for the expressed likelihood of participating measures. For example, Round 1 used the response categories “not at all likely, a little likely, somewhat likely, pretty likely, and very likely,” while Round 2 used the response categories “very likely, somewhat likely, neither likely nor unlikely, somewhat unlikely, and very unlikely.” We conducted a series of exploratory analyses to determine whether responses could be combined across rounds. First, treating the measures as continuous, we converted the raw scores into z-scores. Next, we tested for measurement invariance of the items using correlations between the standardized scores (tested using the *sem* command in Stata 14). Results (not shown) indicated that the response set from each round performed as a parallel measure, supporting the use of combining them across rounds. In analysis, we present values for the expressed likelihood to participate measures by scoring the items 0 to 4 for least to most likely to participate.

We regress each of the participation measures on trust scores separately for the AD and CS experimental groups and then for a model that includes the trust score, response format (coded "1" if CS; "0" if AD), and the interaction between these. We use logistic regression when the dependent variable is dichotomous (past participation) and ordinary least squares (OLS) regression with the continuous (expressed likelihood to participate) dependent variables.

2.5.3. Recency Effects

A recency effect refers to the tendency for respondents to be more likely to select a response category when it appears later in the list regardless of their true answer (Krosnick and Presser 2010). Recency effects are more likely when questions are presented orally (i.e., by an interviewer). According to satisficing theory, recency effects are also more likely when questions are more cognitively demanding: respondents will be more likely to select the last category if it seems reasonable (Holbrook 2008). Because the response categories vary from question to question for the CS questions, we predicted they would be more demanding to process and recall, and respondents would be more likely to select the final response category. However, we expected this effect would be more pronounced for the positively valenced items for which the final category for the CS questions indicates a higher level of trust and may be perceived as a more "reasonable" or agreeable answer.

We assess recency by examining whether the proportion of responses selecting the final category in the list is higher by response format for each question and aggregating across the positively and negatively valenced items with aggregate-level logistic regression models.

2.5.4. Response Latencies

Response latencies (RLs) capture the length of time participants spend processing while they are formulating answers to survey questions (Bassili and Scott 1996; Draisma and Dijkstra 2004). We predicted longer latencies for the CS questions because they use variable construct-specific response categories, encouraging deeper processing (Höhne and Lenzner 2018; Höhne et al. 2017).

Coders timed RLs using audio recordings of the interviews and the visual waveform functionality in the audio software Audacity (Audacity Developer Team 2008). Audacity provides a visual representation of the sound wave, on which coders were able to highlight sections of audio precisely, timing RLs to the thousandths of seconds. Coders began timing after interviewers read the last word of the question during their initial reading of the question. Timing continued through all utterances, including any subsequent readings of the question, and ended when participants uttered the first sound of a word that unambiguously answered the question (e.g., by providing a response category offered by the question). We code interruptions (where a codable response was offered before the entire question was read) and final dispositions that do not accompany a codable response (i.e., don't knows, refusals, and uncodable responses) as missing RLs.

Because they generally have a skewed distribution and outliers, we followed recommendations and top- and bottom-coded values at the 95th and 5th percentiles within each item and use logged values (Yan and Tourangeau 2008). We top- and bottom-coded at the 95th and 5th percentiles and the not the 99th and 1st percentiles because our sample size is small and there are no observations at the 99th and 1st percentiles. ICC between raters using the transformed data is 0.89, which is considered excellent reliability (Landis and Koch 1977). We examine differences for individual questions using t-tests and aggregated across all of the items using OLS regression and aggregate-level tests.

2.5.5. Behavioral Indicators of Response Difficulty (BIRDs)

We predicted the CS questions would be associated with higher levels of BIRDs. As noted, we expected the varying construct-specific response categories to encourage a more elaborated cognitive processing of the questions and the AD questions to encourage a more superficial processing of the questions. Within the context of the cognitive interview, the behavioral indicators of response difficulty offer evidence of a more elaborated processing.

For each question-response interaction (one per participant for each of the trust questions), we tallied the occurrence or non-occurrence of behavioral indicators of response processing difficulties among participants (described below). These indicators are not necessarily final dispositions: a participant may initially say they do not know how to respond to a question, which would be coded as an occurrence of "don't know/refusal," but the interviewer may repeat the question and obtain a codable response. One team member coded all interactions and another member independently coded two thirds of the interactions. All behaviors yielded kappa values with good to excellent agreement (Fleiss 1981).

- *Codable response with qualification* (kappa = 0.83). Participant provides a codable answer (one of the response categories), but qualifies it by adding "probably," "I guess," "maybe," "depends," etc.
- *Codable response with elaboration* (kappa = .87). Participant provides a codable answer, but also provides additional information during their initial response. If the additional information contradicts the codable answer, the response is coded as an "uncodable response."
- *Uncodable response* (kappa = 0.81). Answer does not answer the question or cannot be coded into the response categories (e.g., unrelated report or ambiguous answer).
- *Seeks clarification* (kappa = 0.96). Participant asks for clarification of all or part of a survey question, asks that all or part of a question or response categories be repeated, or repeats part of the question in a way that sounds like a question. These are coded whether or not a codable response is part of the utterance.
- *Question repeated* (kappa = .70). Coded any time the full question and/or response categories are read more than one time before a final disposition (codable response or otherwise) is achieved. This code supplements "seeks clarification," because interviewers sometimes repeat questions without a formal request by the participant.

- *Don't know/refusal* (kappa $= 0.83$). Instead of or in addition to providing an answer, the participant says "don't know" (or the equivalent) and/or refuses to answer the question.

We examine differences between the AD and CS response format for each behavioral indicator aggregated across questions within a scale using logistic regression and aggregate-level tests.

2.5.6. Cognitive Interviewing Data

Lastly, we incorporate qualitative data from the cognitive interviews to explore differences between response processes resulting from AD and CS items. Answers to questions and probes were analyzed qualitatively. The goal was to identify problems that arose, potentially involving misunderstandings of terms, interpreting the intent of the question in different ways, and issues mapping responses onto the categories provided (Tourangeau et al. 2000; Willis 2005). Preliminary codes were based on potential sources of problems during the response process and more specific codes arose during preliminary assessment of the transcripts (Ryan and Bernard 2003). Once the coding scheme was finalized, each interview was coded.

3. Results

3.1. Baseline Comparison of Experimental Groups

As anticipated, the randomization of participants to experimental groups was effective: there are no significant differences ($p < .05$) between the experimental groups based on participants' characteristics (Table 1, Panel A), or by the participation measures used in the validity analysis (Table 1, Panel B), although participants in the CS group reported slightly higher levels of expressed likelihood of providing blood ($p = .06$).

3.2. Trust Scale Summary Statistics and Reliability Measures

Panel A in Table 2 presents summary statistics and reliability coefficients for the scales. Aggregating across 704 question administrations (64 participants × eleven questions), we find the CS scale is associated with significantly higher levels of missing data than the AD scale. Mean trust scores, however, do not significantly differ between the response formats, regardless of whether we impute for missing values. Contrary to expectations, the alpha coefficient, a measure of internal consistency reliability, is significantly higher for the AD than the CS scale.

3.3. Concurrent Validity

We predicted a positive association between trust and participation. Results are in the expected direction for four of the participation measures for the AD scale and five of the participation measures for the CS scale (Table 3). These bivariate associations, however, are only significant for providing tissue and cerebrospinal fluid for the AD scale. Further, the interaction term is only significant for providing tissue: participants' level of expressed

Table 2. Proportion and mean level of trust scale summary statistics, reliability estimates, recency measures, response latencies, and behavioral indicators of response difficulty by AD and CS response formats.

	AD		CS				
Data quality outcomes	Proportion or mean (S.D.)	n	Proportion or mean (S.D.)	n	Difference	Test	p-value
Panel A: Trust scale summary statistics and reliability							
Item-missing aggregate-level	0.03	352	0.10	352	−0.07	b = 1.23; s.e. = 0.44	0.01
Scale score							
List-wise deletion of missing	24.66 (7.31)	32	24.84 (7.35)	32	−0.19	t = −0.10	0.92
Missing imputed	27.53 (4.93)	32	25.63 (7.27)	32	1.91	t = 1.23	0.22
Cronbach's alpha							
List-wise deletion of missing	0.85	32	0.60	32	−0.26	z = 2.09	0.02
Missing imputed	0.84	32	0.58	32	−0.26	z = 2.13	0.03
Panel B: Recency							
Positive valence							
General trust	0.03	31	0.10	31	−0.07	$\chi^2 = 1.07$	0.30
Participants' interest	0.07	30	0.21	29	−0.14	$\chi^2 = 2.47$	0.15
Participants' safety	0.03	32	0.38	26	−0.35	$\chi^2 = 11.65$	< 0.01
Tell about risks	0.03	30	0.37	30	−0.34	$\chi^2 = 10.42$	< 0.01
Treat fairly	0.03	31	0.36	28	−0.33	$\chi^2 = 10.24$	< 0.01
Protect privacy	0.03	32	0.50	30	−0.47	$\chi^2 = 17.77$	< 0.00
Aggregate-level results	0.04	186	0.32	174	−0.28	b = 2.47; s.e. = 0.92	0.01
Negative valence							
Researchers' interest	0.19	31	0.45	29	−0.26	$\chi^2 = 4.49$	0.05
Select minorities	0.17	29	0.13	32	0.04	$\chi^2 = 0.27$	0.72
Hide information	0.06	31	0.04	26	0.02	$\chi^2 = 0.19$	1.00
Treat like guinea pig	0.16	32	0.11	28	0.05	$\chi^2 = 0.31$	0.71
Know more	0.09	32	0.29	28	−0.20	$\chi^2 = 3.38$	0.09
Aggregate-level results	0.14	155	0.20	143	−0.07	b = 0.48; s.e. = 0.41	0.24

Table 2. Continued.

	AD		CS				
Data quality outcomes	Proportion or mean (S.D.)	n	Proportion or mean (S.D.)	n	Difference	Test	p-value
Panel C: Response latency							
General trust	1.29 (1.43)	31	0.64 (1.16)	31	0.65	t = 1.97	0.05
Participants' interests	1.15 (1.36)	31	1.46 (1.73)	29	−0.31	t = −0.77	0.45
Participants' safety	0.82 (1.60)	31	1.21 (1.28)	27	−0.38	t = −1.00	0.32
Tell about risks	0.60 (1.53)	29	1.14 (1.29)	30	−0.54	t = −1.47	0.15
Treat fairly	0.90 (1.49)	31	1.26 (2.05)	27	−0.35	t = −0.76	0.45
Protect privacy	0.43 (1.71)	31	1.33 (1.80)	30	−0.90	t = −2.01	0.05
Researchers' interest	1.44 (1.47)	32	2.07 (1.42)	29	−0.63	t = −1.70	0.09
Select minorities	1.07 (1.46)	26	2.09 (1.60)	32	−1.03	t = −2.52	0.01
Hide information	1.02 (1.48)	30	1.24 (1.65)	25	−0.22	t = −0.52	0.61
Treat like guinea pig	1.46 (1.60)	31	1.59 (1.71)	28	−0.13	t = −0.31	0.76
Know more	1.20 (1.54)	32	0.99 (1.60)	29	0.21	t = 0.51	0.61
Aggregate-level results	1.04 (1.53)	335	1.37 (1.62)	317	−0.33	b = 0.33; s.e. = 0.19	0.09
Panel D: BIRDs							
Aggregate-level results							
Codable + qualification	0.08	352	0.17	352	−0.09	b = 0.89; s.e. = 0.35	0.01
Codable + elaboration	0.24	352	0.12	352	0.12	b = −0.85; s.e. = 0.32	0.01
Uncodable	0.07	352	0.13	352	−0.06	b = 0.74; s.e. = 0.33	0.02
Seeks clarification	0.20	352	0.20	352	0.00	b = 0.02; s.e. = 0.22	0.94
Question repeated	0.19	352	0.24	352	−0.05	b = 0.30; s.e. = 0.22	0.17
"Don't know" or refusal	0.07	352	0.11	352	−0.04	b = 0.46; s.e. = 0.35	0.19

Notes: Aggregate-level tests assess the effect of the measure (e.g., item-missing responses) evaluated collectively across questions treating each question answered by the participant as a separate observation. p-values for Chi-squared tests are from Fisher's exact tests.

Table 3. Concurrent validity analysis: Regression results using trust scores to predict participation for various types of medical research.

	AD only		CS only		AD and CS	
	b	(S.E.)	b	(S.E.)	b	(S.E.)
Past participation						
Trust score	−0.050	(0.056)	0.014	(0.077)	−0.050	(0.056)
Response format					−1.307	(2.606)
Trust score x response format					0.064	(0.095)
Answering questions						
Trust score	−0.008	(0.024)	0.022	(0.037)	−0.008	(0.025)
Response format					−1.114	(1.214)
Trust score x response format					0.030	(0.044)
Providing saliva						
Trust score	0.013	(0.025)	0.013	(0.037)	0.013	(0.025)
Response format					−0.237	(1.223)
Trust score x response format					−0.000	(0.044)
Providing blood						
Trust score	0.024	(0.022)	0.032	(0.038)	0.024	(0.024)
Response format					−0.673	(1.190)
Trust score x response format					0.008	(0.043)
Providing tissue						
Trust score	0.069**	(0.021)	−0.046	(0.036)	0.069**	(0.023)
Response format					2.806*	(1.128)
Trust score x response format					−0.115**	(0.041)
Providing cerebrospinal fluid						
Trust score	0.065**	(0.021)	0.023	(0.038)	0.065**	(0.024)
Response format					0.894	(1.155)
Trust score x response format					−0.042	(0.042)

Notes: Regression coefficients are from logistic regression for past participation and OLS regression for answering questions and providing saliva, blood, tissue, and cerebrospinal fluid. Trust scale scores are computed by summarizing the z-scores across questions within experimental group.
* $p < .05$; ** $p < .01$

likelihood to participate in research by providing tissue is significantly lower with the CS scale. Overall, the scales appear equally valid in predicting participation.

3.4. Recency

For positively valenced items, we predicted the proportion of responses using the last category for the CS scale (which varied by question but for which the last category indicates more trust) would be higher than the proportion of responses using the last categories for the AD scale (which is always "strongly disagree" such that the last value indicates less trust). Indeed, the CS scale yields more responses using the last category (Panel B, Table 2) for all of the positively valenced items, and the difference is significant for four of the items and for the aggregate-level test. In contrast, for the negatively valenced items, the difference is only significant for one item and the aggregate-level test is not significant.

3.5. Response Latencies (RLs)

We predicted RLs would be longer for the CS scale because the changing construct-specific response categories encourage deeper processing. Overall, nine of the eleven CS items have longer mean RLs than the parallel AD item (Panel C, Table 2); two of these ("protect privacy" and "select minorities") are statistically significant and the aggregate-level test is marginally significant ($p < .09$).

Interestingly, RLs are significantly longer for the AD scale for the first question administered as part of the scale ("general trust"). Here respondents are hearing the question and response categories read for the first time, and the longer response time could be evidence for the more cognitively burdensome response task offered by the AD response format and distinct from the effect of grouping AD questions in a battery.

3.6. Behavioral Indicators of Response Difficulty (BIRDs)

We predicted the CS questions would be associated with higher levels of BIRDs. Panel D in Table 2 presents aggregate-level logistic regression tests for the BIRDs indicators (see Appendix B, Subsection 6.2, for question-by-question results). We find significantly higher levels of codable answers with qualifications and uncodable answers for the CS questions versus higher levels of codable answers with elaborations for the AD questions; there are no differences between response formats for seeking clarification, asking to have a question repeated, or providing a don't know or refusal response. These results help interpret the longer response latencies found with the CS items. In almost all cases, indicators of response difficulty, such as providing a qualification or uncodable response require more interactional time to resolve and result in longer response latencies. In contrast, elaborations tend to follow codable answers and so would not be part of the latency.

3.7. Cognitive Interviewing Data

Quantitative analyses demonstrate the AD and CS scales differ on several of the data quality indicators examined, but these analyses do not provide insight about *why* this is so. A strength of the current study is that we embedded the AD-CS comparison in cognitive interviews in which participants first answered the survey questions that comprised the quantitative analysis, and then answered open-ended follow-up questions about their answers. We incorporate participants' qualitative responses to further examine *why* the response formats produced different results. The qualitative analysis revealed five potential sources of difficulties:

1. Understanding the intent of questions: interpreting questions about opinions as knowledge questions,
2. Understanding the intent of questions: managing comparisons between target objects in a question,

3. Difficulty mapping: dealing with a lack of knowledge or ambivalence,
4. Difficulty mapping: remembering CS categories, and
5. Difficulty mapping: mismatched vocabulary.

We assess how these difficulties with the response task varied depending on whether the participant received the AD or CS scale, and we provide excerpts from the cognitive interviews to illustrate how these potential sources of error manifested.

3.7.1. Understanding the Intent of Questions: Interpreting Questions About Opinions as Knowledge Questions

Trust is a subjective evaluation that may or may not depend on facts about events and behaviors related to the trustee, in this case medical researchers (Hall et al. 2001). Several participants said they did not have enough information to answer the trust questions, indicating they interpreted questions as asking about their knowledge rather than for their evaluation (Excerpt 1, Table 4). The intent of the question about "participants' interests" was to gauge each participants' **attitudes** about the relative frequency with which medical researchers have the best interests of participants from their racial/ethnic group in mind. The intent was not to measure objectively how often medical researchers actually engage in the behavior. This participant's responses, however, indicated she felt we were asking her to "project [her] thoughts into another person," While we observed this issue for both AD and CS formats, the higher proportion of administrations yielding "don't knows" for the CS scale (11%, compared to 7% for the AD scale) indicates this may have been more of a problem for CS questions.

3.7.2. Difficulty Interpreting Intent of Questions: Managing Comparisons between Target Objects in a Question

During follow-up questioning, we documented several participants unintentionally reversing the direction of a comparison, providing a codable answer incongruent with their reasoning. The question about "researchers' interests" asked participants to evaluate how much they believe medical researchers care about their research **compared** to the participants in their studies. In Excerpt 2, the interviewer realizes the participant's reasoning did not match her initial response of "a great deal." While the interviewer catches the incongruence, it is not possible to tally how often this type of mismatch occurred and went unnoticed by the participant or the interviewer, especially if the participant provided vague reasoning for their answer. Not only did respondents flip the direction of the comparison so that their answers were about researchers caring more about their participants than their research, but they needed more time to provide a codable answer. The CS version of the "researchers' interests" question had the fourth longest response latency, suggesting comparisons using CS may be particularly burdensome.

3.7.3. Difficulty Mapping: Dealing With a Lack of Knowledge or Ambivalence

Participants reported feeling uninformed or ambivalent about matters related to medical researchers. They dealt with this ambivalence differently depending on whether they were attempting to map responses onto AD or CS categories. With AD items, when a participant expressed that they needed more knowledge on the topic in order to answer the question,

Table 4. Excerpts from the Cognitive Interviews.

Actor	Text
Excerpt 1:	**"Participants' interests," CS Response Format**
Interviewer	When they are conducting research, how often do medical researchers have the best interests of participants from your racial or ethnic group in mind: never, rarely, sometimes, very often, or always?
Participant	There would be no way for me to know the answer to that question.
Interviewer	Tell me why.
Participant	You're, you're asking me to, um, project my thoughts into another person. I, I, I can't do that. I can only answer questions that directly involve me, I guess.
Excerpt 2:	**"Researchers' interests," CS Response Format**
Interviewer	To what extent do medical researchers care more about their research than they do about the participants in their studies: not at all, a little, somewhat, quite a bit, or a great deal?
Participant	A great deal.
Interviewer	And can you tell me more about why you answered a great deal?
Participant	Because you need the participants to even figure out what's going on in the study that they're performing. So I would think they would care a lot, just as much as they do for the study.
Interviewer	Okay. Um, this question is a little confusing, so I'm going to reread it to you. Um, to what extent do medical researchers care more about their research than they do about the participants in their studies? And you said a great deal. Um, but then you were.
Participant	Oh, you were saying do they care more about the research than they do the peoples that are participating in it.
Interviewer	Yeah. I believe that's what the question is asking.
Participant	No, I don't think.
Interviewer	Okay. So would you say not at all, a little, somewhat, quite a bit?
Participant	Not at all.
Interviewer	Okay. And you said that, that you, I'm sorry. Could
Participant	I said that because, um, they need the participants to figure this stuff out for the study. So I would think they would, it would be equal, even.
Excerpt 3:	**"Participants' interests," AD Response Format**
Participant	Well, see, that, that, I'm going to either, you know, agree and disagree at the same time or whatever, neither disagree or disagree, because, again, I, I'm ignorant. I don't have that much knowledge on, uh, people that do these kinds of things, these studies and everything. So I don't know if they, they're more interested in a certain ethnic group or a certain category or age or whatever. I, I have no knowledge on, uh, why a person does medical interviews, you know, so I, I don't agree or disagree. I have no, no knowledge.

Excerpt 4:	**“Treat like a guinea pig,” CS Response Format**
Interviewer	How often do medical researchers treat participants from your racial or ethnic group like guinea pigs in their studies: never, rarely, sometimes, very often, or extremely often?
Participant	Well, let me put never, because, in reality, I don’t know.
Interviewer	Okay. And the question following up is tell me more about why you answered never for this question.
Participant	It’s because I don’t know. I’m not informed in that aspect.
Excerpt 5:	**“Know more,” CS Response Format**
Interviewer	How often do medical researchers want to know more than they need to know: never, rarely, sometimes, very often, or extremely often?
Participant	Hmm, the categories again are what?
Interviewer	Never, rarely, sometimes, very often, or extremely often?
Participant	And now I forgot the question [L].
Interviewer	How often do medical researchers want to know more than they need to know?
Participant	Okay. Um, rarely.
Excerpt 6:	**“Researchers’ interests,” CS Response Format**
Participant	Oh, I’m sorry, can, can you repeat the question one more time?
Interviewer	The question or the responses?
Participant	The question.
Interviewer	Okay. To what extent do medical researchers care more about their research than they do about the participants in their studies: not at all, a little, somewhat, quite a bit, or a great deal?
Participant	Well, for me, the, the answers you’re giving me are, are difficult to use to w-, to make this response.
Interviewer	Okay. Can you tell me a little bit more about why it’s difficult?
Participant	Sure. It, it’s not the type of vocabulary that I use. I don’t use quite a bit or a great deal a lot for any, for anything, so I don’t have any, any, any f-, any meaning in anything that I say.
Interviewer	Okay. And if you had to choose one, what answer would you give me?
Participant	Well, the, the questions, the answers you gave me before, those were easier to use.

they were often able to provide a codable answer by responding with “neither agree nor disagree.” Because they are bipolar, AD items include a middle category that appeared to be a reasonable option for participants who did not feel they had enough knowledge or who were not inclined to answer one way or the other (Excerpt 3). In contrast, the CS items are unipolar and lack a clear “neutral” (middle) category. In order for an uncertain or ambivalent participant to conclude the interaction, she could either pick a category that was not an exact match to her “true” state or provide an uncodable response (e.g., “don’t know”). There is some evidence that at least two of these behaviors occurred during responses to CS items; the participant in Excerpt 4 does both.

3.7.4. Difficulty Mapping: Remembering CS Categories

One clear difference between the AD and CS items is the format of the response categories: AD questions use the same response categories for each item while the CS categories vary by question. A potential source of difficulty is that the CS categories were harder for participants to remember, an issue exacerbated by the aural presentation of the items and the number of items in the scale (the scale included 11 items). In Excerpt 5, the participant requested to hear the response categories a second time, but by the time the participant had a handle on the categories, they had forgotten the content of the question.

3.7.5. Difficulty Mapping: Mismatched Vocabulary

In a few cases participants reported the CS categories did not use their common language. In Excerpt 6, the participant indicated difficulty using the categories and elaborated that particular phrases like "quite a bit" or "a great deal" are not in his usual vocabulary, so he does not have "meaning in anything" he says. This participant was Latino and the interview was conducted in Spanish, which could have further complicated category interpretability. If participants are uncomfortable or unfamiliar with vocabulary and interpret and use the categories differently, reliability may be lower. This problem may be particularly relevant in cross-cultural research and research with diverse samples, such as this study.

4. Discussion

Based on past research, we formulated several hypotheses about how the closed-ended survey responses would differ between the AD and CS versions of the scale. We expected the CS scale would yield higher reliability and validity than the AD scale. Results, however, indicated higher reliability for the AD scale and neither scale appeared more valid in predicting participation. While these results seem to favor the AD scale, AD responses may be more internally consistent, as indicated by the significantly higher value of coefficient alpha, because factors like acquiescence and the use of the same response categories increases common method variance and not because of the scale's ability to reliably measure the underlying construct. Unfortunately, the small sample sizes in our study precluded more sophisticated analyses, such as using structural equation models to account for potential method effects that may have biased estimates of reliability and validity, particularly for the AD scale (e.g., Kuru and Pasek 2016). In addition, we were limited in the availability of criterion measures for the validity analysis. We selected the past participation and expressed likelihood to participate measures because of their demonstrated relationship with trust in past research, but they are not ideal; their wording varied somewhat between rounds of interviewing, their response format was more similar to the CS than AD format, and they were not strongly associated with trust scores in this study.

In developing scales, experts recommend that they be balanced with half of the items measuring one direction of the construct (in our case high trust with the positively valenced items) and half measuring the other direction (in our case low trust with the negatively valenced items) (Streiner et al. 2015). Because the CS response categories vary from question to question, they are likely to be more demanding to process and recall (Höhne and Lenzner 2018), and we hypothesized that the CS scale would be associated with greater recency effects than the AD scale, particularly for the positively valenced CS questions for which the final category indicated a higher level of trust and possibily a more "reasonable" or agreeable answer. Consistent with our expectations, we found a higher proportion of responses in the last category for the CS questions overall.

That participants often struggled to remember the wording of the variable CS categories was also observed in the qualitative data. Much of the past research comparing AD and CS questions has been conducted using self-administered questions or with visual aids. We did not provide showcards because the cognitive interviews were the first step in a study with the purpose of developing a telephone survey. In this unfamiliar context, participants may have been unable or unwilling to dedicate cognitive resources to remembering the variable CS categories, often necessitating a reread of CS items. Further, the trust scale was quite long – it included 11 items. Most previously tested comparisons of AD and CS questions involve many fewer items. When CS scales are long and contain many questions with variable response categories, they may be more problematic when presented aurally. It is possible that the variable nature of the categories coupled with the length of the scale contributed to the lower reliability we documented with the CS scale.

We also predicted that responses to CS questions would yield longer processing times than AD questions and that responding to CS items would involve more instances of behavioral indicators of response difficulty. Overall, results match our expectations. Aggregating across questions, the CS scale is associated with a marginally significant higher mean response latency and higher levels of qualifications and uncodable answers, behaviors that tend to increase response latencies. We often interpret longer response latencies and the presence of the behavioral indicators we measured as signs of cognitive response difficulty (e.g., Bassili and Scott 1996; Schaeffer and Dykema 2011). However, recent theorizing and evidence suggests these outcomes may be desirable when comparing AD and CS questions (Höhne and Lenzner 2018; Höhne et al. 2017). Because CS questions use construct-specific response categories, which will likely vary on a question-by-question basis in a battery, they encourage deeper processing, which will increase processing time and likely result in respondents producing other behaviors when attempting to respond in an optimal manner.

Being able to turn to qualitative data from our cognitive interviews added substantially to our understanding of these mixed quantitative results. They revealed several sources of difficulties for respondents that varied by the AD or CS questioning format. For example, the CS items may have confused the intent of the question. Even if AD items generally are more cognitively burdensome than CS items, the response dimension "agreement" is a reminder to respondents that the question seeks their evaluation about the topic. For CS items, depending on the response dimension (e.g., intensity, frequency, or quantity), the intent of the question can be less clear. For example, several items seemed to ask about knowledge or facts related to the target object (e.g., "how hard do medical researchers work to ensure participants in their studies are safe") rather than respondents' evaluations about the target object (e.g., "how confident are you that medical researchers work hard to ensure participants in their studies are safe"). They also focused on evaluations of external objects (e.g., "medical researchers") and to a lesser extent on internal or self-focused objects (e.g., "you"). Although others have reported higher validity for CS questions that focus on evaluations of the characteristics or qualities of external objects such as Facebook or doctors (e.g., Kuru and Pasek 2015; Saris et al. 2010), we recommend future research explore whether AD or CS questions yield more desirable data quality outcomes for questions of the type examined here.

Experiments evaluating data quality for the inclusion of middle categories for bipolar questions has been mixed (see Krosnick and Presser 2010), and another important finding from the qualitative analysis was a description of participants' use of the middle "neither agree nor disagree" category to deal with a lack of knowledge and express ambivalence with the bipolar AD questions. In contrast, the unipolar CS response categories do not include a clear middle or "neutral" category. In this unfamiliar context, participants sought ways to express ambivalence, but struggled to do so with the CS response categories, often resulting in significantly higher level of item-nonresponse and uncodable answers. In contrast, participants who were asked AD items often selected "neither agree nor disagree" to express uncertainty or ambivalence. Our findings are consistent with those of Sturgis et al. (2014) who probed respondents selecting the "neither/nor" middle category during the administration of three attitudinal questions to determine why respondents selected that category. Overwhelmingly respondents reported selecting the middle category because they did not have an opinion on the issue. Further, this strategy was employed more often among respondents who indicated more interest in the topic under consideration, possibly as a way to "save face" and avoid having to say "don't know" outright. From a measurement perspective, respondents use of the "neither/nor" middle category is highly problematic: while respondents may reliably select this middle option, their response is not a valid measure of the construct being assessed. Researchers have noted problems with the interpretation of the middle category with AD questions and often suggest that responses using this category should be analyzed separately and not as the middle value between agree and disagree (Willits et al. 2016).

5. Conclusions

For survey methodologists, one important consideration is that AD and CS items seem to demand different levels of cognitive effort, which may vary depending on characteristics of the questions and the mode of administration including: (1) valence (whether the question is positively or negative valenced); (2) offered response dimension (whether the offered response dimension measures intensity, frequency, or quantity; the offered response dimension for an AD question is by definition intensity – the intensity of agreement – but the offered response dimension – the dimension that is explicit with a CS question – will likely vary); (3) number of response categories (most comparisons use five categories, but some experiments use seven or eleven categories); (4) labeling of categories (whether categories are fully labeled versus end-point-only labeled); (5) direction of response categories (whether the categories increase in value – "not at all" to "extremely" – or decrease in value – "extremely" to "not at all"); and (6) polarity (whether the question is bipolar or unipolar; AD questions are always bipolar, CS questions can vary). In addition, if questions are bipolar, an important feature is the inclusion (or exclusion) of a middle category (e.g., "neither agree nor disagree") and how it is labeled (Dykema et al. 2019).

In our experiment, the AD and CS questions varied based on their offered response dimension (the response dimensions for the CS questions were construct-specific by design and tapped into the dimensions of intensity, frequency, and quantity), the direction of the response categories (the AD response categories were ordered from high to low – "strongly agree" to "strongly disagree" – while the CS categories were ordered from low to high – "not at all" to "a great deal," "never" to "always"), and their polarity (the AD questions were bipolar; the CS were unipolar). While our design does not allow us to estimate the unique effects of these characteristics, we encourage future work using multifactorial designs that will provide researchers with the ability to estimate the effects of particular characteristics.

With regard to the mode of administration, one critical difference between interviewer-administered and self-administered modes is that respondents need to encode and recall the response categories in order to map their response. Providing showcards for all CS items during in-person interviews may reduce cognitive burden on respondents. However, this solution is not easily applicable to telephone interviews and CS scales that include many items with variable response categories may be problematic. Another possibility is to select response options that vary less dramatically from question to question and that use the everyday language of respondents, which may introduce an additional challenge if the item is to be used in cross-cultural research. These issues are likely to receive increased scrutiny as surveys that mix interviewer- and self-administration grow and researchers continue to explore methods to measure and reduce mode effects (De Leeuw and Berzelak 2016).

We note several other limitations of this study. First, while prior research indicates AD questions are more problematic for respondents with lower education (e.g., Schuman and Presser 1996), our analytic sample was small, precluding subgroup analyses based on education or other socio-demographic variables. Second, if particular aspects of trust are more salient for certain groups (such as those defined by race/ethnicity) or groups use response categories differently, measurement nonequivalence may help explain a portion of the current results (Davidov et al. 2014). Future work investigating whether AD or CS items yield higher levels of measurement equivalence is needed, especially for cross-cultural research and research involving diverse samples, such as the current study.

Third, our data were collected in a unique situation, that of a cognitive interview. While it is possible that the semi-structured nature of this interviewing situation affected outcomes in ways that would not generalize to a standardized survey interview, the format allowed us to collect data on how participants process AD and CS questions, a major contribution of this study. Given that this study yielded some unexpected findings, pretesting of new CS items remains crucial, particularly if the sample is diverse, the target object being evaluated is unfamiliar to participants, and if the evaluation being solicited is complex (e.g., trust, evaluations of others).

6. Appendix

6.1. *Appendix A*

Appendix A. Exact wording of the agree-disagree (AD) and construct-specific (CS) questions from the trust in medical researchers scale, by round of cognitive interviewing.

	Round 1		Round 2	
Question label	AD version	CS version	AD version	CS version
Positive valence (most positively valenced category – e.g., "strongly agree" for AD and "a great deal" for CS – indicates most trust)				
General trust	All things considered, you trust medical researchers a great deal. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	All things considered, how much do you trust medical researchers: none, a little, some, quite a bit, or a great deal?	All things considered, you trust medical researchers a great deal. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	All things considered, how much do you trust medical researchers: none, a little, some, quite a bit, or a great deal?
Participants' interests	Medical researchers always have the best interests of participants from your racial or ethnic group in mind. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How much of the time do medical researchers have the best interests of participants from your racial or ethnic group in mind: none of the time, a little of the time, some of the time, most of the time, or all of the time?	**When they are conducting research,** medical researchers always have the best interests of participants from your racial or ethnic group in mind. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	**When they are conducting research**, how **often** do medical researchers have the best interests of participants from your racial or ethnic group in mind: **never, rarely, sometimes, very often, or always**?
Participants' safety	Medical researchers work hard to make sure that the participants in their studies are safe. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How hard do medical researchers work to make sure that the participants in their studies are safe: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?	Medical researchers work hard to make sure that the participants in their studies are safe. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How hard do medical researchers work to make sure that the participants in their studies are safe: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?

Appendix A. Continued.

	Round 1		Round 2	
Question label	AD version	CS version	AD version	CS version
Tell about risks	Medical researchers always tell participants everything they need to know about the risks of participating in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers tell participants everything they need to know about the risks of participating in their studies: never, rarely, sometimes, very often, or extremely often?	Medical researchers always tell participants everything they need to know about the risks of participating in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers tell participants everything they need to know about the risks of participating in their studies: never, rarely, sometimes, very often, or **always**?
Treat fairly	Medical researchers treat participants from your racial or ethnic group the same as participants from other racial or ethnic groups. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers treat participants from your racial or ethnic group the same as participants from other racial or ethnic groups: never, rarely, sometimes, very often, or extremely often?	Medical researchers **always** treat participants from your racial or ethnic group the same as participants from other racial or ethnic groups. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers treat participants from your racial or ethnic group the same as participants from other racial or ethnic groups: never, rarely, sometimes, very often, or **always**?
Protect privacy	Medical researchers work hard to make sure they keep information from participants private and secure. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How hard do medical researchers work to make sure they keep information from participants private and secure: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?	Medical researchers work **extremely** hard to make sure they keep information from participants private and secure. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How hard do medical researchers work to make sure they keep information from participants private and secure: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?

Appendix A. Continued.

	Round 1		Round 2	
Question label	AD version	CS version	AD version	CS version
Negative valence (most positively valenced category – e.g., “strongly agree” for AD and “a great deal” for CS – indicates least trust)				
Researchers’ interests	Medical researchers care more about their research than they do about the participants in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	Compared to their research, do medical researchers care a lot less, somewhat less, about the same, somewhat more, or a lot more about the participants in their studies?	Medical researchers care more about their research than they do about the participants in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	**To what extent do medical researchers care more about their research than they do about the participants in their studies: not at all, a little, somewhat, quite a bit, a great deal?**
Select minorities	Medical researchers are more likely to select minorities for their most risky studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How likely are medical researchers to select minorities for their most risky studies: not at all likely, a little likely, somewhat likely, very likely, or extremely likely?	**When selecting participants for their most risky studies**, medical researchers are more likely to select minorities. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	**When selecting participants for their most risky studies**, how likely are medical researchers to select minorities: not at all likely, a little likely, somewhat likely, very likely, or extremely likely?
Hide information	Medical researchers never hide information about the possible risks of participating. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers hide information about the possible risks of participating: never, rarely, sometimes, very often, or extremely often?	Medical researchers **often** hide information about the possible risks of participating **in medical research studies**. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers hide information about the possible risks of participating **in medical research studies**: never, rarely, sometimes, very often, or extremely often?

Appendix A. Continued.

	Round 1		Round 2	
Question label	AD version	CS version	AD version	CS version
Treat like guinea pig	Medical researchers treat participants from your racial or ethnic group like guinea pigs in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers treat participants from your racial or ethnic group like guinea pigs in their studies: never, rarely, sometimes, very often, or extremely often?	Medical researchers **often** treat participants from your racial or ethnic group like guinea pigs in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers treat participants from your racial or ethnic group like guinea pigs in their studies: never, rarely, sometimes, very often, or extremely often?
Know more	Medical researchers often want to know more than they need to know. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers want to know more than they need to know: never, rarely sometimes, very often, or extremely often?	Medical researchers often want to know more than they need to know. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers want to know more than they need to know: never, rarely, sometimes, very often, or extremely often?

Notes: Differences in question wording between rounds are shown in bold.

6.2. Appendix B

Appendix B. Proportion of participants exhibiting a given behavioral indicator of response difficulty (BIRD), by question and experimental group.

	Codable response + qualification				Codable response + elaboration				Uncodable Response				Seeks Clarification			
	AD	CS	Difference	p-value	AD	CS	Difference	p-value	AD	CS	Difference	p-value	AD	CS	Difference	p-value
Positive valence																
General trust	0.03	0.03	0.00	1.00	0.09	0.00	0.09	0.24	0.06	0.09	−0.03	1.00	0.19	0.03	0.16	0.10
Participants' interests	0.00	0.13	−0.13	0.11	0.19	0.09	0.10	0.47	0.03	0.19	−0.16	0.10	0.31	0.19	0.12	0.39
Participants' safety	0.09	0.22	−0.13	0.30	0.28	0.16	0.12	0.37	0.03	0.06	−0.03	1.00	0.13	0.19	−0.06	0.73
Tell about risks	0.00	0.25	−0.25	0.01	0.25	0.13	0.12	0.34	0.03	0.06	−0.03	1.00	0.22	0.06	0.14	0.15
Treat fairly	0.09	0.09	0.00	1.00	0.38	0.13	0.25	0.04	0.03	0.22	−0.19	0.05	0.22	0.31	−0.09	0.57
Protect privacy	0.13	0.13	0.00	1.00	0.06	0.13	−0.07	0.67	0.03	0.03	0.00	1.00	0.19	0.25	−0.06	0.76
Negative valence																
Researchers' interest	0.13	0.31	−0.18	0.13	0.22	0.25	−0.03	1.00	0.13	0.13	0.00	1.00	0.25	0.13	0.12	0.29
Select minorities	0.16	0.28	−0.12	0.37	0.31	0.09	0.22	0.06	0.09	0.19	−0.10	0.47	0.16	0.28	−0.12	0.37
Hide information	0.03	0.22	−0.19	0.05	0.28	0.09	0.19	0.11	0.16	0.19	−0.03	1.00	0.16	0.09	0.07	0.71
Treat like guinea pig	0.13	0.13	0.00	1.00	0.41	0.13	0.28	0.02	0.13	0.16	−0.03	1.00	0.19	0.25	−0.06	0.76
Know more	0.09	0.13	−0.04	1.00	0.19	0.13	0.06	0.73	0.03	0.16	−0.13	0.20	0.19	0.16	0.03	1.00

Appendix B. Continued.

	Question Repeated				Don't Know or Refusal			
	AD	CS	Difference	p-value	AD	CS	Difference	p-value
Positive valence								
General trust	0.19	0.03	0.16	0.10	0.06	0.00	0.06	0.49
Participants' interests	0.25	0.38	−0.13	0.42	0.19	0.09	0.10	0.47
Participants' safety	0.16	0.19	−0.03	1.00	0.00	0.16	−0.16	0.05
Tell about risks	0.25	0.16	0.09	0.54	0.09	0.06	0.03	1.00
Treat fairly	0.25	0.31	−0.06	0.78	0.09	0.19	−0.10	0.47
Protect privacy	0.22	0.28	−0.06	0.77	0.09	0.09	0.00	1.00
Negative valence								
Researchers' interest	0.22	0.44	−0.22	0.11	0.03	0.06	−0.03	1.00
Select minorities	0.16	0.38	−0.22	0.09	0.16	0.06	0.10	0.43
Hide information	0.19	0.03	0.16	0.10	0.00	0.22	−0.22	0.01
Treat like guinea pig	0.19	0.31	−0.12	0.39	0.00	0.13	−0.13	0.11
Know more	0.06	0.19	−0.13	0.26	0.06	0.13	−0.07	0.67

Notes: For each question, n = 32 participants for the AD scale and n = 32 participants for the CS scale. p-values are from Chi-squared tests and Fisher's exact tests.

7. References

Anderson, L.A. and R.F. Dedrick. 1990. “Development of the Trust in Physician Scale: A Measure to Assess Interpersonal Trust in Patient-physician Relationships.” *Psychological Reports* 67: 1091–1100. Doi: https://doi.org/10.2466/pr0.1990.67.3f.1091.

Audacity Developer Team. 2008. Audacity (Version 1.2.6) [Computer Software]: Available at: http://www.audacityteam.org/download/ (accessed April 2019).

Bassili, J.N. and B.S. Scott. 1996. “Response Latency as a Signal to Question Problems in Survey Research.” *Public Opinion Quarterly* 60: 390–399. Doi: https://doi.org/10.1086/297760.

Braunstein, J.B., N.S. Sherber, S.P. Schulman, E.L. Ding, and N.R. Powe. 2008. “Race, Medical Researcher Distrust, Perceived Harm, and Willingness to Participate in Cardiovascular Prevention Trials.” *Medicine* 87: 1–9. Doi: https://doi.org/10.1097/MD.0b013e3181625d78.

Carpenter, P.A. and M.A. Just. 1975. “Sentence Comprehension: A Psycholinguistic Processing Model of Verification.” *Psychological Review* 82: 45–73. Available at: http://psycnet.apa.org/doi/10.1037/h0076248 (accessed April 2019).

Corbie-Smith, G., S.B. Thomas, and D.M.M. St. George. 2002. “Distrust, Race, and Research.” *Archives of Internal Medicine* 162: 2458–2463. Doi: https://doi.org/10.1001/archinte.162.21.2458.

Davidov, E., B. Meuleman, J. Cieciuch, P. Schmidt, and J. Billiet. 2014. “Measurement Equivalence in Cross-National Research.” *Annual Review of Sociology* 40: 55–75. Doi: https://doi.org/10.1146/annurev-soc-071913-043137.

De Leeuw, E. and N. Berzelak. 2016. “Survey Mode or Survey Modes?” In *The SAGE Handbook of Survey Methodology*, edited by C. Wolf, J. Dominique, T.W. Smith, and F. Yang-chih, 142–156. Los Angeles: SAGE Publications Ltd. Available at: https://books.google.com/books?hl=en&lr=&id=g8OMDAAAQBAJ&oi=fnd&pg=PA142&dq=survey+mode+or+modes+berzelak&ots=DyqMiBT1oS&sig=hGg7pa80-bI535N5GgSUwvLmLfY#v=onepage&q=survey%20mode%20or%20modes%20berzelak&f=false (accessed April 2019).

Dijkstra, W. and Y. Ongena. 2006. “Question-Answer Sequences in Survey-Interviews.” *Quality & Quantity* 40: 983–1011. Doi: https://doi.org/10.1007/s11135-005-5076-4.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th edition). Hoboken, NJ: John Wiley.

Draisma, S. and W. Dijkstra. 2004. “Response Latency and (Para)linguistic Expression as Indicators of Response Error.” 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, 131–148. New York: Springer-Verlag. Doi: https://doi.org/10.1002/0471654728.ch7.

Dykema, J., J.M. Lepkowski, and S. Blixt. 1997. “The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study.” In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 287–310. N.Y: Wiley-Interscience. Available at: https://onlinelibrary.wiley.com/doi/10.1002/9781118490013.ch12 (accessed April 2019).

Dykema, J., N.C. Schaeffer, and D. Garbarski. 2012. “Effects of Agree-Disagree Versus Construct-Specific Items on Reliability, Validity, and Interviewer-Respondent Interaction.” Presented at the American Association for Public Opinion Research, May 17–20. 2012. Orlando, Florida, U.S.A.

Dykema, J., N.C. Schaeffer, and D. Garbarski. 2019. “Towards a Reconsideration of the Use of Agree-Disagree Questions in Measuring Subjective Evaluations.” Unpublished manuscript, University of Wisconsin-Madison, Madison-WI.

Edwards, D.F. 2015. “Voices Heard.” Presented at the Health Equity Leadership Institute, Madison, WI.

Egede, L.E. and C. Ellis. 2008. “Development and Testing of the Multidimensional Trust in Health Care Systems Scale.” *Journal of General Internal Medicine* 23: 808–815. Doi: https://doi.org/10.1007/s11606-008-0613-1.

Fleiss, J.L. 1981. *Statistical Methods for Rates and Proportions*, 2nd edition. New York: Wiley.

Fortune-Greeley, A.K., K.E. Flynn, D.D. Jeffery, M.S. Williams, F.J. Keefe, R.B. Reeve, G.B. Willis, and K.P. Weinfurt. 2009. “Using Cognitive Interviews to Evaluate Items for Measuring Sexual Functioning Across Cancer Populations: Improvements and Remaining Challenges.” *Quality of Life Research* 18: 1085–1093. Doi: https://doi.org/10.1007/s11136-009-9523-x.

Fowler, F.J. and C. Cosenza. 2009. “Design and Evaluation of Survey Questions.” In *The Sage Handbook of Applied Social Research Methods*, edited by L. Bickman and D.J. Rog, 375–412. Thousand Oaks, CA: Sage.

Hall, M.A., F. Camacho, E. Dugan, and R. Balkrishnan. 2002a. “Trust in the Medical Profession: Conceptual and Measurement Issues.” *Health Services Research* 37: 1419–1439. Doi: https://doi.org/10.1111/1475-6773.01070.

Hall, M.A., F. Camacho, J.S. Lawlor, V. DePuy, J. Sugarman, and K. Weinfurt. 2006. “Measuring Trust in Medical Researchers.” *Medical Care* 44: 1048–1053. Available at: http://www.jstor.org/stable/41219560 (accessed April 2019).

Hall, M.A., E. Dugan, B. Zheng, and A.K. Mishra. 2001. “Trust in Physicians and Medical Institutions: What is It, Can It be Measured, and Does It Matter?” *Milbank Quarterly* 79: 613–639. Doi: https://doi.org/10.1111/1468-0009.00223.

Hall, M.A., B. Zheng, E. Dugan, F. Camacho, K.E. Kidd, A. Mishra, and R. Balkrishnan. 2002b. “Measuring Patients’ Trust in their Primary Care Providers.” *Medical Care Research and Review* 59: 293–318. Doi: https://doi.org/10.1177/1077558702059003004.

Hanson, T. 2015. “Comparing Agreement and Item-Specific Response Scales: Results from an Experiment.” *Social Research Practice* 1: 17–25. Available at: http://the-sra.org.uk/wp-content/uploads/social-research-practice-journal-issue-01-winter-2015.pdf (accessed April 2019).

Hayman, R.M., B.J. Taylor, N.S. Peart, B.C. Galland, and R.M. Sayers. 2001. “Participation in Research: Informed Consent, Motivation and Influence.” *Journal of Paediatrics and Child Health* 37: 51–54. Available at: https://doi.org/10.1046/j.1440-1754.2001.00612.x (accessed April 2019).

Henderson, G., J. Garrett, J. Bussey-Jones, M.E. Moloney, C. Blumenthal, and G. Corbie-Smith. 2008. “Great Expectations: Views of Genetic Research Participants Regarding

Current and Future Genetic Studies." *Genetics in Medicine* 10: 193–200. Doi: https://doi.org/10.1097/GIM.0b013e318164e4f5.

Höhne, J.K. and D. Krebs. 2018. "Scale Direction Effects in Agree/Disagree and Item-Specific Questions: A Comparison of Question Formats." *International Journal of Social Research Methodology* 21: 91–103. Doi: https://doi.org/10.1080/13645579.2017.1325566.

Höhne, J.K. and T. Lenzner. 2018. "New Insights on the Cognitive Processing of Agree/Disagree and Item-Specific Questions." *Journal of Survey Statistics and Methodology* 6: 401–417. Doi: https://doi.org/10.1093/jssam/smx028.

Höhne, J.K., S. Schlosser, and D. Krebs. 2017. "Investigating Cognitive Effort and Response Quality of Question Formats in Web Surveys Using Paradata." *Field Methods* 29: 365–382. Doi: https://doi.org/10.1177/1525822x17710640.

Holbrook, A.L. 2008. "Recency Effect." In *Encyclopedia of Survey Research Methodology*, edited by P.J. Lavrakas, 695–696. Newbury Park, CA: Sage.

Johnson, R.B. and A.J. Onwuegbuzie. 2004. "Mixed Methods Research: A Research Paradigm Whose Time Has Come." *Educational Researcher* 33: 14–26. Doi: https://doi.org/10.3102/0013189X033007014.

Krosnick, J.A. and S. Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research*, Second Edition, edited by P.V. Marsden and J.D. Wright, 263–313. Bingley, UK: Emerald Group Publishing Limited.

Kuru, O. and J. Pasek. 2016. "Improving Social Media Measurement in Surveys: Avoiding Acquiescence Bias in Facebook Research." *Computers in Human Behavior* 57: 82–92. Available at: https://doi.org/10.1016/j.chb.2015.12.008 (accessed April 2019).

Landis, J.R. and G.G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33: 159–174. Doi: https://doi.org/10.2307/2529310.

Lelkes, Y. and R. Weiss. 2015. "Much Ado about Acquiescence: The Relative Validity and Reliability of Construct-Specific and Agree-Disagree Questions." *Research and Politics* 2: 1–8. Doi: https://doi.org/10.1177/2053168015604173.

Liu, M., S. Lee, and F.G. Conrad. 2015. "Comparing Extreme Response Styles between Agree-Disagree and Item-Specific Scales." *Public Opinion Quarterly* 79: 952–975. Doi: https://doi.org/10.1093/poq/nfv034.

Mainous, A.G., D.W. Smith, M.E. Geesey, and B.C. Tilley. 2006. "Development of a Measure to Assess Patient Trust in Medical Researchers." *Annals of Family Medicine* 4: 247–252. Doi: https://doi.org/10.1370/afm.541.

Revilla, M. and C. Ochoa. 2015. "Quality of Different Scales in an Online Survey in Mexico and Columbia." *Journal of Politics in Latin America* 7: 157–177. Available at: https://journals.sub.uni-hamburg.de/giga/jpla/article/view/903/910 (accessed April 2019).

Rogers, W. 1994. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 13. Available at: https://ideas.repec.org/a/tsj/stbull/y1994v3i13sg17.html (accessed April 2019).

Ryan, G.W. and H.R. Bernard. 2003. "Techniques to Identify Themes." *Field Methods* 15: 85–109. Doi: https://doi.org/10.1177/1525822x02239569.

Saris, W.E., M. Revilla, J.A. Krosnick, and E.M. Shaeffer. 2010. "Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response

Options." *Survey Research Methods* 4: 61–79. Doi: https://ojs.ub.uni-konstanz.de/srm/article/view/2682/3971.

Schaeffer, N.C. and J. Dykema. 2011. "Response 1 to Fowler's Chapter: 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, A. Maitland, and G. Willis, 23–39. Hoboken, NJ: John Wiley & Sons, Inc. Available at: https://doi.org/10.1002/9781118037003.ch3.

Scharff, D.P., K.J. Mathews, P. Jackson, J. Hoffsuemmer, E. Martin, and D. Edwards. 2010. "More than Tuskegee: Understanding Mistrust about Research Participation." *Journal of Health Care for the Poor and Underserved* 21: 879–897. Doi: https://doi.org/10.1353/hpu.0.0323.

Schuman, H. and S. Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage Publications, Inc.

Smith, T.W., P.V. Marsden, and M. Hout. 2013. General Social Survey, 1972–2010 [Cumulative File]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2013-02-07. Doi: https://doi.org/10.3886/ICPSR31521.v1.

Streiner, D.L., G.R. Norman, and J. Cairney. 2015. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford, UK: Oxford University Press.

Sturgis, P., C. Roberts, and P. Smith. 2014. "Middle Alternatives Revisited: How the neither/nor Response Acts as a Way of Saying "I Don't Know"?" *Sociological Methods & Research* 43: 15–38. Doi: https://doi.org/10.1177/0049124112452527.

Thompson, H.S., H.B. Valdimarsdottir, G. Winkel, L. Jandorf, and W.W. Redd. 2004. "The Group-Based Medical Mistrust Scale: Psychometric Properties and Association with Breast Cancer Screening." *Preventive Medicine* 38: 209–218. Doi: https://doi.org/10.1016/j.ypmed.2003.09.041.

Tourangeau, R., M.C. Couper, and F. Conrad. 2004. "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68: 368–393. Doi: https://doi.org/10.1093/poq/nfh035.

Tourangeau, R., L.J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.

Williams, M.M., D.P. Scharff, K.J. Mathews, J.S. Hoffsuemmer, P. Jackson, J.C. Morris, and D.F. Edwards. 2010. "Barriers and Facilitators of African American Participation in Alzheimer Disease Biomarker Research." *Alzheimer Disease & Associated Disorders* 24: S24–S29. Available at: https://journals.lww.com/alzheimerjournal/Fulltext/2010/07001/Barriers_and_Facilitators_of_African_American.6.aspx (accessed April 2019).

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

Willis, G.B. and K. Miller. 2011. "Cross-Cultural Cognitive Interviewing: Seeking Comparability and Enhancing Understanding." *Field Methods* 23: 331–341. Doi: https://doi.org/10.1177/1525822x11416092.

Willits, F.K., G.L. Theodori, and A.E. Luloff. 2016. "Another Look at Likert Scales." *Journal of Rural Social Sciences* 31: 126–139. Available at: http://journalofruralsocialsciences.org/pages/Articles/JRSS%202016%2031/3/JRSS%202016%2031%203%20126-139.pdf (accessed April 2019).

Yan, T. and R. Tourangeau. 2008. “Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times.” *Applied Cognitive Psychology* 22: 51–68. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.1331 (accessed April 2019).

Zheng, B., M.A. Hall, E. Dugan, K.E. Kidd, and D. Levine. 2002. “Development of a Scale to Measure Patients’ Trust in Health Insurers.” *Health Services Research* 37: 185–200. Doi: https://doi.org/10.1111/1475-6773.00145.

Received October 2017
Revised December 2018
Accepted January 2019