International Journal of Public Opinion Research Advance Access published September 2, 2014

International Journal of Public Opinion Research

doi:10.1093/ijpor/edu024

Sources of Interactional Problems in a Survey of Racial/Ethnic Discrimination

Timothy P. Johnson[1], Salma Shariff-Marco[2,3], Gordon Willis[4], Young Ik Cho[5], Nancy Breen[4], Gilbert C. Gee[6], Nancy Krieger[7], David Grant[8], Margarita Alegria[9], Vickie M. Mays[10], David R. Williams[11], Hope Landrine[12], Benmei Liu[3], Bryce B. Reeve[13], David Takeuchi[14] and Ninez A. Ponce[6]

[1]Survey Research Laboratory, University of Illinois at Chicago, Chicago, IL, USA; [2]Cancer Prevention Institute of California, Fremont, CA, USA; [3]Stanford University, School of Medicine; [4]National Cancer Institute, Rockville, MD, USA; [5]University of Wisconsin-Milwaukee, Milwaukee, WI, USA; [6]University of California, Los Angeles, CA, USA; [7]Harvard University, School of Public Health, Boston MA, USA; [8]UCLA Center for Health Policy Research, Los Angeles, CA, USA; [9]Center for Multicultural Mental Health Research, Somerville, MA, USA; [10]UCLA School of Public Health, Los Angeles, CA, USA; [11]Department of Society, Human Development and Health, School of Public Health, Harvard University, Boston, MA, USA; [12]Center for Health Disparities Research, East Carolina University, Greenville, NC, USA; [13]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [14]Boston College, MA, USA

Abstract

Cross-cultural variability in respondent processing of survey questions may bias results from multiethnic samples. We analyzed behavior codes, which identify difficulties in the interactions of respondents and interviewers, from a discrimination module contained within a field test of the 2007 California Health Interview Survey. In all, 553 (English) telephone interviews yielded 13,999 interactions involving 22 items. Multilevel logistic regression modeling revealed that respondent age and several item characteristics (response format, customized questions, length, and first item

All correspondence concerning this article should be addressed to Timothy P. Johnson, Survey Research Laboratory, University of Illinois at Chicago (MC 336), 629 CUPPA Hall, 412 S. Peoria, Chicago, IL 60607-7064, USA. E-mail: timj@uic.edu; tjohnson@srl.uic.edu

with new response format), but not race/ethnicity, were associated with interactional problems. These findings suggest that item function within a multi-cultural, albeit English language, survey may be largely influenced by question features, as opposed to respondent characteristics such as race/ethnicity.

Introduction

A persistent challenge in survey methodology, for both attitudinal and behavioral measures, is understanding the relative influence of respondent and question characteristics because both influence data quality through the survey response process. A paucity of clear criterion measures makes it difficult to assess response error directly, but a common proxy measure of measurement error used is the quality of the interaction between interviewer and respondent, as assessed by behavior coding (Cannell, Fowler, & Marquis, 1968; Fowler, 2011; Schaeffer & Dykema, 2011). Behavior coding entails the systematic assignment of codes to the behaviors of respondents and interviewers during survey interviews (Dijkstra, 2008). Originally intended as a means for assessing interviewer performance, behavior coding quickly became focused on respondent behavior as well to identify individual problematic survey questions. More recently, researchers have further modified the design of investigations, not to identify individual items that are problematic, but rather the codeable features of items (e.g., length, concreteness–abstractness) that are associated with the frequency of assignment of behavior codes (see Schaeffer & Dykema, 2011 for a review of the issue of definition of question characteristics). Further, some of these investigations also extend beyond the realm of question features and additionally focus on respondent characteristics that influence the interaction. For example, Cannell et al. (1968) and van der Zoewen and Smit (2004) found that older respondents tend to produce higher frequencies of potentially problematic interactions than do younger ones.

More recently, researchers have attempted to account for both question and respondent characteristics within the same investigation. From the perspective of statistical analysis, this undertaking requires accounting for the hierarchical design, in which items are nested within respondent, through the use of hierarchical linear modeling (HLM) or equivalent statistical approaches (Holbrook, Cho, & Johnson, 2006). Such studies have become increasingly prominent, as investigators assess the influence of racial, ethnic, and cultural group membership on question validity, and as surveys become increasingly heterogeneous in their language of administration and respondent composition. For example, Holbrook et al. (2006) examined the behavior of 423 adult African–Americans, Mexican Americans, Puerto Ricans, and non-Hispanic whites when administered a set of common health-related items used in U.S. federal population surveys such as the US Centers for Disease Control and Prevention (CDC) National Health Interview Survey, or the National Health and Nutrition Examination Survey. Questions were coded according to five characteristics: Abstraction level, length, reading level, response format, and whether the item was complicated by the requirement to perform a qualified judgment (e.g., involving a specific time interval, or that specifically excluded particular categories). Measures of both comprehension difficulty and difficulty in selecting a response option ("mapping") were used as indicators of interactional problems for the respondent. Overall, Holbrook et al. found that comprehension difficulties were associated (positively) with question length, abstraction level, requirement of a complex qualified judgment, use of an open-ended numeric response format, and reading level. With respect to respondent characteristics, race/ethnicity was associated with comprehension difficulty. Response mapping difficulties were associated with several of the same item characteristics, but not with race/ethnicity; rather, older respondents exhibited more difficulties in selecting a response. It is notable, however, that no studies have investigated how both respondent and interviewer behavior, as well question characteristics, influence the survey response process, as assessed via behavior coding.

The possibility that racial/ethnic group differences may exist in the survey response process is intriguing and a cause for concern, as they may undermine the validity of survey research as a methodological tool. Holbrook et al. (2006) reported greater question comprehension difficulties, as measured by behavior codes, among the several minority populations examined, relative to non-Latino whites, a finding that suggests the existence of group disparities in ability to answer questions as phrased during a survey interview. Scant other evidence is currently available, however, to confirm this finding or to address this concern. To extend this finding, we conducted the current investigation as part of efforts designed to evaluate a discrimination module (DM) developed for use with the California Health Interview Survey (CHIS) across a broad range of racial/ethnic groups: Latinos, African–Americans, American Indians/Alaska Natives (AI/ANs), Asian–Americans/Native Hawaiians and Other Pacific Islanders (AANHOPIs), non-Latino whites, and those indicating multiracial background. The CHIS DM study presented a valuable opportunity to investigate measurement disparities across a wide range of racial/ethnic groups. Moreover, because this questionnaire, which we are using as a vehicle for assessing interactional difficulties, addressed the topic of self-reported racial/ethnic discrimination, it would presumably bring out group-specific issues that might also be reflected in response processes. Groups with varying discrimination experiences, for example, might react differently in terms of seeking clarifications regarding the meaning of questions included in the DM.

As in the Holbrook et al. study, we sought to assess the relative effects of question characteristics and respondent characteristics on observable behavior likely to be indicative of survey interview quality. Finally, we also investigated the effects of the quality of the interviewer's presentation of each survey question on the quality of respondent's answers.

Methods

Data

The data for this analysis are based on 553 respondents selected from the CHIS DM behavior coding study. The behavior coding study was nested within a larger CHIS DM field test, which was conducted in English, ($N = 7401$; response rate = 21.1%) between June 20, 2007 and March 3, 2008 as part of the CHIS 2007 (California Health Interview Survey, 2009). The objective of the field test was to evaluate the validity and reliability of the DM and to compare two of the most commonly used approaches for asking about racial/ethnic discrimination in a population-based, diverse racial/ethnic sample (Shariff-Marco et al., 2009, 2011). One uses a two-stage approach (Version A), first asking about unfair treatment generally and then asking about attribution for those experiences (e.g., gender, age, race/ethnicity). The other uses a one-stage approach asking, within a single question, about unfair treatment due to the respondent's race/ethnicity (Version B). For the field test, a split-sample design was used to randomly assign respondents to instruments using either the one-stage or two-stage approach. The version of the instrument using the one-stage approach included 22 questions, and the version using the two-stage approach included 36 questions (sample items from both versions are illustrated in Table 1). The study was approved by the Institutional Review Board (IRB) at the University of California at Los Angeles (UCLA) and exempted from IRB review at the National Institutes of Health.

Within the overall CHIS DM field test, 553 respondents were assigned to participate in the behavior coding study. The behavior coding sample was stratified to yield a final sample that included African–Americans, AI/ANs, AANHOPIs, Latinos, non-Latino whites, and persons of multiracial backgrounds. The telephone interviews were recorded and subsequently behavior coded, applying a modified set of the Cannell, Lawson, and Hausser (1975) codes.

Behavior coders were instructed to code only the first exchange between interviewers and respondents. A total of three respondent codes and one interviewer code were available to be assigned to the initial exchange associated with each question. The specific definitions of each behavior code are reported by Kudela, Stark, Hantmann, Seak, and Newsome (2009) and reproduced in Table 2.

Table 1
Sample items from the CHIS Discrimination Module by approach

One-stage approach—ask specifically about discrimination based on race/ethnicity.	Two-stage approach—ask about unfair treatment, then ask about the reasons for this unfair treatment.
Recent everyday discrimination and appraisal of discrimination as stressful.	
B1. In the past 12 months, how often have you been treated with less respect than other people because you are {FILL WITH PREFERRED SELF-REPORTED RACE/ETHNICITY}? Would you say . . . Never, Rarely, Sometimes, OR Often	B1. In the past 12 months, how often have you been treated with less respect than other people? Would you say . . . Never, Rarely, Sometimes, OR Often
B2. In the past 12 months, how often have you been treated unfairly or been discriminated against at restaurants or stores because you are [FILL]? Would you say . . . Never, Rarely, Sometimes, OR Often	B2. In the past 12 months, how often have you been treated unfairly at restaurants or stores? Would you say . . . Never, Rarely, Sometimes, OR Often
B5. [In the past 12 months,] how often have people acted as if they are afraid of you because you are [FILL]? [Would you say . . . Never, Rarely, Sometimes, OR Often]	B5. [In the past 12 months,] how often have people acted as if they are afraid of you? [Would you say . . . Never, Rarely, Sometimes, OR Often]
If yes to one or more items, B1–B8:	If yes to one or more items, B1–B8:
	B9A: For each of the following, please tell me if it was a reason why you were treated unfairly, in the past 12 months. First, because of your ancestry or national origin? Because of your gender or sex? Because of your race or skin color? Because of your age? Because of the way you speak English? Because of some other reason? IF yes to more than one: B9B. Which of these do you think is the main reason why you were treated unfairly, in the past 12 months?
B10. In the past 12 months, how stressful have these experiences of unfair treatment usually been for you? Would you say . . . Not at all stressful, A little stressful, Somewhat stressful, OR Extremely stressful	B10. In the past 12 months, how stressful have these experiences of unfair treatment usually been for you? Would you say . . . Not at all stressful, A little stressful, Somewhat stressful, OR Extremely stressful

(continued)

Table 1
Continued

One-stage approach—ask specifically about discrimination based on race/ethnicity.	Two-stage approach—ask about unfair treatment, then ask about the reasons for this unfair treatment.
Lifetime discrimination and appraisal of discrimination as stressful	
C1. Next we would like for you to think about unfair experiences over your entire lifetime. Over your entire lifetime, how often have you have been treated unfairly or been discriminated against at school because you are [FILL]? Would you say . . . Never, Rarely, Sometimes, OR Often	C1. Next we would like for you to think about unfair experiences over your entire lifetime. Over your entire lifetime, how often have you have been treated unfairly at school? Would you say . . . Never, Rarely, Sometimes, OR Often
C3. [Over your entire lifetime,] how often have you been treated unfairly or been discriminated against when getting medical care because you are [FILL]? [Would you say . . . Never, Rarely, Sometimes, OR Often]	C3. [Over your entire lifetime,] how often have you been treated unfairly when getting medical care? [Would you say . . . Never, Rarely, Sometimes, OR Often]
If yes to one or more items, C1–C5:	If yes to one or more items, C1–C5: C6A: For each of the following, please tell me if it was a reason why you were treated unfairly, over your entire lifetime? First, because of your ancestry or national origin? Because of your gender or sex? Because of your race or skin color? Because of your age? Because of the way you speak English? Because of some other reason? IF yes to more than one: C6B: Which of these do you think is the main reason why you were treated unfairly, over your entire lifetime?
C7. Over your entire lifetime, how stressful have these experiences of unfair treatment usually been for you? Would you say . . . Not at all stressful, A little stressful, Somewhat stressful, OR Extremely stressful	C7. Over your entire lifetime, how stressful have these experiences of unfair treatment usually been for you? Would you say . . . Not at all stressful, A little stressful, Somewhat stressful, OR Extremely stressful
Responses	
D2. Did you get angry or get into an argument or physical fight?	D2. Did you get angry or get into an argument or physical fight?
D4. Did you pray or meditate about the situation?	D4. Did you pray or meditate about the situation?

Table 2
Definitions of Respondent and Interviewer Behavior Codes

Interviewer codes[a]	Definition
Question read correctly	The interviewer asked the question either exactly as written or close enough to its written form that the meaning of the question was not changed.
Question read incorrectly	The interviewer failed to read the question as worded, either by leaving out important words and phrases or rewording the question in some other way that changed the question's meaning. Or, the interviewer failed to administer a question that should have been asked.
Respondent codes[b]	
Respondent interrupts	The respondent interrupted the interviewer's reading of the question.
Respondent requests clarification	The respondent said something to indicate he or she did not hear or understand the question (e.g., asking for a repeat of the question; asking the interviewer what the question or answer categories mean).
Adequate answer	The respondent answered the question using the answer categories on the questionnaire or in a way that can reasonably be classified into one of those categories.
Problem with answer	The respondent answered the question, but the answer did not fit the answer categories (e.g., gives a range instead of a precise number); the respondent was obviously unsure of an answer; the respondent did not know the answer or refused to answer.

Note. [a]Interviewer codes were mutually exclusive.
[b]Respondent codes were not mutually exclusive.
Source. Kudela et al. (2009).

Measures

In the behavior coding study reported here, four key coding categories were assessed. These included the presence or absence of each of the following behaviors: (1) "interviewer reads question incorrectly;" (2) "respondent interrupts;" (3) "respondent requests clarification;" and (4) "respondent has a problem answering." The first of these coding categories emphasizes interviewer behavior, including whether the question was read out loud (or asked) as written, or in a way that altered its intended meaning.

The other three codes focused on respondent behavior. "Respondent interrupts" represents whether the respondent interrupted the interviewer while the question was being read, potentially signaling that a question was perceived as being too long, that the respondent believed that the question did not need to be fully read, or that the interviewer did a poor job reading the question. “Respondent requests clarification” was coded whenever the respondent requested further information—e.g., “What do you mean by unfair treatment”—possibly signaling question vagueness or comprehension difficulty. A third category of “Respondent has a problem answering” was applied if the respondent exhibited difficulty, or was completely unable to provide a response using the intended answer categories. A summary measure of any respondent problem (i.e., respondent interrupts, respondent requests clarification, and/or respondent has a problem answering) was also constructed using these three indicators to reflect any respondent difficulty answering each question. Five coders were responsible for the behavior coding. Kappa coefficients for behavior code intercoder agreement were 0.47 for “respondent has a problem answering”, 0.50 for “respondent interrupts,” 0.61 for “interviewer reads question incorrectly,” and 0.83 for “respondent requests clarification.”

In addition to race/ethnicity, other respondent-level demographic variables available for these analyses included gender, age, and education. Age and education were assessed using 6 and 10 ordinal categories, respectively (see Table 3).

Four sets of question-level characteristics were examined: (1) question response format, in relation to three types of format: (a) yes-no, (b) ordinal responses with verbal labels, and (c) other response formats, referring chiefly to selecting the most important response from a set of previous answers; (2) question length (measured as a simple word count); (3) whether an item was the first asked in a series that used a new response format from those asked in previous questions; and (4) whether an item used customized question text. Customized question text is defined as whether “fills” were used to add customized information into the question (e.g., specific types of events, respondent race/ethnicity). The distributions of these question types are presented in Table 3.

Analysis

Multilevel cross-classified logistic regression models with HLM7 (Raudenbush, Bryk, Cheong, & Congodon, 2011) were used to simultaneously assess the effects of both respondent-level and question-level characteristics on response-level interactional difficulties in answering individual survey questions. The models account for the joint clustering of the response level behavior coded outcomes within each respondent and question (Rodriguez & Elo, 2003). Codes indicating when interviewers misread questions were also included as a covariate at the response level in these models. An example of a model that estimates a behavioral interaction problem by question

Table 3
Respondent and Question Characteristics

Respondent characteristics ($n = 553$)	Mean or %	(SE)
Gender		
Male	49.7%	
Female	50.3%	
Race/ethnicity		
Latino	21.3%	
African–American	19.3%	
American Indian/Alaska Native	4.2%	
Asian American/Native Hawaiian/Other Pacific Islander	19.2%	
Non-Latino white	27.1%	
Multiracial	8.9%	
Age (six categories, 1 = 18–29 years, 6 = 65 + years)	3.88	(1.78)
1. 18–29	13.6%	
2. 30–39	17.4%	
3. 40–44	8.3%	
4. 45–49	11.2%	
5. 50–64	27.5%	
6. 65+	22.1%	
Education (10 categories, 1 = ≤8th grade, 10 = PhD or equivalent)	5.32	(2.44)
1. Grade 1–8	2.2%	
2. Grade 9–11	5.6%	
3. Grade 12/high school diploma	21.7%	
4. Some college	22.4%	
5. Vocational school	2.4%	
6. AA or AS degree	8.0%	
7. BA or BS degree	19.5%	
8. Some graduate School	1.4%	
9. MA or MS degree	11.9%	
10. PhD or equivalent	4.9%	
Questionnaire version[a]		
Version A (one-stage)	50.3%	
Version B (two-stage)	49.7%	
Question-level characteristics ($n = 58$)		
Response format		
Yes-no	44.8%	
Ordinal	51.7%	
Other[b]	3.4%	
Customized question text using fills		
Yes	25.9%	
No	74.1%	
Item length (i.e., word count)	16.55	(1.29)
First item asked using new response format		
Yes	20.7%	
No	79.3%	

Notes: [a]Although questionnaire version is logically a question-related rather than respondent-related attribute, it was randomly assigned to each respondent and is therefore treated as a respondent-level variable. [b]Other response format involved respondents selecting a most important response from a set of previous answers.

characteristics, controlling for respondent attributes, is presented in two equations as follows:

Level-1 Model (response behavior outcome level)

$$\mathrm{Prob}\big(Problem_{ijk} = 1|\pi_{jk}\big) = \phi_{\mathrm{ijk}}$$

$$\log[\phi_{\mathrm{ijk}}/(1-\phi_{\mathrm{ijk}})] = \eta_{ijk}$$

$$\eta_{ijk} = \pi_{0jk} + \pi_{1jk}(Incorrectly\ Read_{ijk})$$

where

π_{0jk} is an intercept, and π_{1jk} is the coefficient of the incorrectly read for the response (coded response behavior) *i*, respondent *j*, and question *k*, predicting the probability of having problem with question *i* by respondent *j*. The logit link function is used for the binary outcome.

Level-2 Model (respondent and question level)

$$\begin{aligned}\pi_{0jk} = \ & \theta_0 + \gamma_{01}(Male_j) + \gamma_{02}(Aage_j) + \gamma_{03}(Latino_j) + \gamma_{04}(African - American_j)\\ & + \gamma_{05}(AI/AN_j) + \gamma_{06}(AANHOPI_j) + \gamma_{07}(Multiracial_j)\\ & + \gamma_{08}(Education_j) + \gamma_{09}(Version_j) + \beta_{01}(Yes - \mathrm{No}_k)\\ & + \beta_{02}(Other\ Type_k) + \beta_{03}(Customized_k) + \beta_{04}(Number\ of\ Words_k)\\ & + \beta_{05}(First\ Item_k) + b_{00j} + c_{00k}\end{aligned}$$

where $\gamma_{on}(n = 1, 2, \ldots, N)$ is Level-2 coefficients of the respondent characteristics, and $\beta_{on}(n = 1, 2, \ldots, N)$ is Level-2 coefficients of the question characteristics.

b_{ooj} and c_{ook} are residual respondent- and question-specific random effects, respectively, representing the deviation of respondent *j*'s and question *k*'s Level-1 intercept, π_{ojk}, an overall probability of reporting any problem. They are assumed to be normal [$b_{ooj} \sim N(0, \tau\ b_{ooj})$, $c_{ook} \sim N(0, \tau\ c_{ooj})$]. All analyses were conducted with unweighted data.

Results

Characteristics of respondents are presented in Table 3. The sample included approximately equal numbers of males and females. Respondents representing each of the six racial/ethnic groups of interest were also included, although the proportions in each group were not equal. Almost equal numbers of respondents were assigned to answer each version of the questionnaire ($n = 275$ for two-stage vs. 278 for one-stage). Also included in Table 3 is a summary of the characteristics of the 58 questions. The mean length of the questions answered was 16.6 words ($SD = 1.29$). Almost 21% of questions (20.7%)

Table 4
Behavior code variables (N = 13,999)

Behavior coded problems	*n*	%
One or more respondent problem	1,848	13.2
Respondent interrupts	492	3.5
Requests clarification	572	4.1
Problem with answer	847	6.1
Interviewer fails to read question correctly	980	7.0

introduced new response formats. Most questions used ordinal sets of verbal response labels (51.7%). Other questions used yes-no response options, except for a small proportion (3.4%) in which respondents were asked to select a previous answer that was deemed to be the "main reason" why they felt they were treated unfairly. Just over a quarter of all questions (25.9%) used "fills" to insert customized question text into questions.

Table 4 presents the frequencies with which each type of behavior was observed. Respondents interrupted interviewers while reading the questions in 3.5% of all instances. They requested clarification in 4.1% of all instances, and respondents had a problem answering 6.1% of all questions. At least one of these behaviors was observed in responses to 13.2% of all questions. In addition, interviewers failed to read the question correctly 7.0% of the time.

The first four columns in Table 5 present the results of a multilevel logistic regression model that evaluates the independent effects of respondent and question-level characteristics on the combined measure of any respondent behaviors that potentially indicated problems. Respondent age was the only demographic measure associated with these potential problems. Several question characteristics, in contrast, were significantly and independently associated with behavior coding problems. Questions using the yes-no response format produced, on average, fewer problems than did ordinal response formats. Ordinal response formats, in turn, generated fewer problems than did other, less common, response formats. Questions using customized text were also found to elicit fewer behavioral problems. In addition, when interviewers were coded as having failed to read a question correctly, there was a much greater likelihood that respondents would also be coded as having a behavioral problem when answering the question. There was also an association between version (one-stage vs. two-stage) of the DM questions that each respondent answered, and the presence or absence of interactional problems. The two-stage version was less likely to elicit an interactional problem. Figure 1 presents the adjusted probabilities that behavior problems were observed by each respondent and question characteristic that was significantly associated with the summary behavior coding measure.

Table 5
Multilevel Logistic Regression Models Predicting Problematic Respondent Behaviors (N = 13,999 question answers; N = 553 respondents)

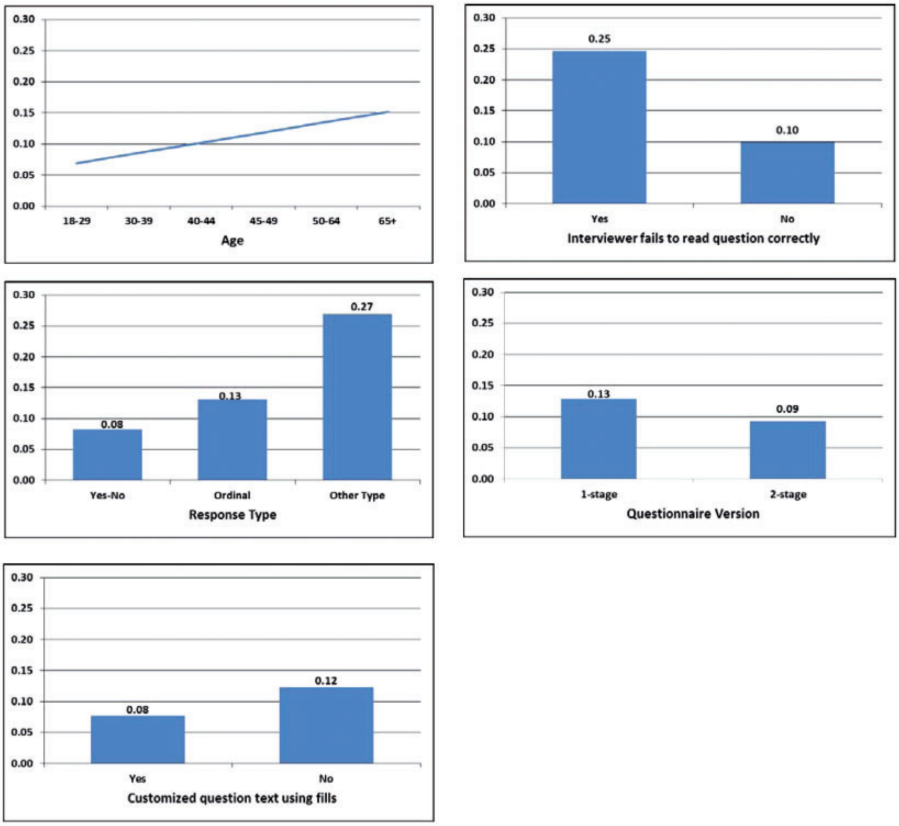
Fixed effect	Any respondent problem			Respondent interrupts			Respondent clarification			Problem with answer		
	OR (odds ratio)	CI (confidence interval)	*p*	OR	CI	*p*	OR	CI	*p*	OR	CI	*p*
Respondent level												
Intercept, θ_o	0.12	(0.11,0.13)	<.001	0.01	(0.01,0.02)	<.001	0.04	(0.03,0.04)	<.001	0.04	(0.03,0.05)	<.001
Gender: men (referent group = women), γ_{o1}	0.97	(0.80,1.15)	.710	1.01	(0.76,1.33)	.950	1.00	(0.80,1.24)	.990	1.02	(0.79,1.31)	.860
Age-group (range 1–6), γ_{o2}	1.19	(1.13,1.25)	<.001	1.18	(1.08,1.29)	<.001	1.08	(1.01,1.16)	.017	1.25	(1.15,1.35)	<.001
Race/ethnicity (referent = non-Latino white)												
Latino, γ_{o3}	0.80	(0.61,1.05)	.114	0.87	(0.56,1.33)	.514	0.98	(0.69,1.37)	.885	0.66	(0.44,0.98)	.039
African–American, γ_{o4}	1.03	(0.79,1.32)	.840	1.00	(0.66,1.49)	.999	1.27	(0.92,1.72)	.138	0.95	(0.66,1.35)	.771
American Indian/Alaska Native, γ_{o5}	1.02	(0.64,1.60)	.943	1.16	(0.58,2.30)	.676	1.19	(0.67,2.09)	.553	0.87	(0.45,1.66)	.665
Asian–American/Native Hawaiian/Other Pacific Islander, γ_{o6}	0.99	(0.75,1.30)	.944	0.98	(0.63,1.51)	.944	0.94	(0.66,1.32)	.736	1.13	(0.77,1.65)	.524
Multiracial, γ_{o7}	0.98	(0.69,1.37)	.906	1.18	(0.69,1.98)	.542	0.79	(0.50,1.23)	.304	1.18	(0.74,1.87)	.472
Education (range 1–10), γ_{o8}	0.98	(0.94,1.02)	.342	0.94	(0.88,0.99)	.034	1.03	(0.98,1.07)	.243	0.97	(0.92,1.02)	.257
Questionnaire version: two-stage (referent = one-stage), γ_{o9}	0.69	(0.52,0.91)	.009	0.56	(0.26,1.19)	.132	0.60	(0.39,0.90)	.015	0.92	(0.60,1.39)	.699
Question level												
Response format (referent = ordinal)												

(continued)

Table 5
Continued

Fixed effect	Any respondent problem			Respondent interrupts			Respondent clarification			Problem with answer		
	OR (odds ratio)	CI (confidence interval)	*p*	OR	CI	*p*	OR	CI	*p*	OR	CI	*p*
Yes-no, β_{01}	0.60	(0.48,0.74)	<.001	0.23	(0.11,0.45)	<.001	0.41	(0.27,0.60)	<.001	1.00	(0.70,1.41)	.977
Other response format (referent = ordinal), β_{02}	2.46	(1.41,4.27)	.002	0.42	(0.07,2.28)	.311	2.58	(0.99,6.67)	.051	4.81	(2.14,10.80)	<.001
Customized question text using fills (referent = non-customized text), β_{03}	0.59	(0.43,0.81)	.001	0.53	(0.21,1.32)	.172	0.40	(0.23,0.67)	<.001	0.83	(0.50,1.37)	.472
Number of words, β_{04}	1.01	(0.99,1.02)	.107	1.07	(1.03,1.11)	.001	0.99	(0.96,1.01)	.467	1.00	(0.97,1.01)	.784
First item asked using new response format (referent: not first item asked using new response format), β_{05}	1.29	(0.98,1.69)	.068	0.65	(0.28,1.47)	.292	1.50	(0.93,2.40)	.09	1.54	(0.99,2.38)	.053
Question response level												
Interviewer fails to read question correctly (referent: interviewer reads question correctly), π_1	2.87	(2.41,3.40)	<.001	8.60	(6.63,11.16)	<.001	1.20	(0.86,1.67)	.264	1.34	(1.03,1.74)	.024

Figure 1
Adjusted probability of having any respondent behavior problem by selected respondent and question characteristics



Additional models presented in Table 5 examine the effects of the same set of respondent, question, and interviewer effects on each of the three individual respondent behavior codes of respondent interrupts (Columns 5–7), respondent requests clarification (Columns 8–10), and problematic answer (Columns 11–13). In general, the directions of the odds ratios for each independent variable are in a consistent direction, although not always reaching significance. Most notable is the consistently positive and significant relationship between respondent age and each observed behavior. Also of note is the very large odds ratio (OR = 8.60) between interviewer misreading a question and respondent interruptions, an effect that seems to overwhelm most others in that model. A fairly large effect of question response formats other than ordinal and yes/no formats (OR = 4.81) was also observed in the model predicting problematic respondent answers. The only model in which an effect of respondent race/ethnicity was found to be significant was the final one in Table 5, in which Latino respondents were found to be less likely to provide problematic answers in comparison with non-Latino white respondents. This was the only significant effect observed among 20 race/ethnicity coefficients examined across the four models in Table 5.

We also re-estimated each model to determine whether there were statistical interactions between respondent race/ethnicity and each question-level characteristic. Of the 100 statistical interactions examined, 9 were found to be significant, with no clear direction in findings. Other models examined the effects of respondent and question characteristics separately for those responses for which interviewers did ($n = 13{,}015$) and did not ($n = 984$) correctly read questions to respondents. For responses for which questions were correctly read, findings replicated very closely those observed for the full sample (in Columns 2–4 of Table 5). For responses in which questions were not read correctly, the direction of observed relationships was generally consistent with those for responses in which questions were read correctly. These latter findings, however, were not all significant, due to the much smaller available sample of misread questions. These additional models are available from the authors.

Discussion

Although questionnaire designers have increasingly been concerned with measurement disparities across racial and ethnic groups and their potential threats to cross-cultural validity, we found only weak evidence that the DM questions we evaluated performed in disparate ways across a broad range of racial and ethnic respondent subgroups—at least in English-language interviews. This is significant given that the questionnaire used was primarily concerned with racial/ethnic discrimination. This finding stands in contrast to those of Holbrook et al. (2006), which examined a broader range of health survey questions and used a larger set of behavior codes specifically linked to question comprehension problems, although with a less diverse range of racial/ethnic groups (African–Americans, Mexican Americans, Puerto Ricans, and non-Latino whites). The lack of differences in respondent expressions of difficulty answering the DM questionnaire may be a consequence of the research invested in addressing this issue, via cognitive interviewing, before fielding the CHIS DM survey (see Reeve et al., 2011 for more details).

Among demographic variables assessed, only respondent age was associated with interactional difficulties. This finding confirms other research that has also indicated increasing likelihood of observing problematic behavior codes among older respondents (Cannell et al., 1968; Johnson et al., 2006; van der Zoewen & Smit, 2004) and is consistent with other findings that have documented declines in the quality of survey reporting with increasing age (Alwin, 2007; Andrews & Herzog, 1986).

On the other hand, several basic question-design features were related to observed difficulties for the DM. The positive contribution of using yes-no response option formats to minimizing interactional problems have been reported in other investigations (Johnson et al., 2006). The positive effects of customizing survey question text via "fills," and the use of a two-stage procedure to collect information regarding discrimination experiences, however, have not been previously reported. Each of these findings is intuitive. The sequential, two-stage, question format likely produces fewer interactional problems as a consequence of separating questions about experience of unfair treatment and attribution into separate items. Likewise, the contribution of survey question customization toward minimizing interactional problems would seem to be further validation of the positive effects of computer technology on interview quality.

Several researchers have recommended that survey items that elicit behavior code problems in >15% of all cases should be of particular concern (Fowler & Cannell, 1996). Using this benchmark, Figure 1 suggests that questions that use "other types" of response options (specifically, selecting a most important response from a set of previous answers) should be used with considerable caution. Likewise, and perhaps not surprisingly, those items that are misread by interviewers were also found to be highly problematic.

The findings outlined above suggest that cognitive processing and providing answers to English versions of survey questions may be more similar than dissimilar when administered across cultural groups, and that the major factors that affect difficulties in response are associated with question presentation formats. Further, most respondents are affected similarly by question structure, with the general exception of persons with increasing age.

This conclusion must be tempered by several additional points. First, as noted above, all interviews in the current study were conducted in English. This likely selects for respondents who are acculturated to U.S. society and therefore more likely to process survey questions somewhat similarly. Further, the questionnaire had been subjected to cognitive testing before being fielded (Reeve et al., 2011), and may have harbored fewer sources of cross-cultural variation than instruments receiving less attention to potential cultural variability when under development. Finally, we note that our basic measure of question function—behavior coding—by design captures only overt problems in the interaction, and is ill-suited to investigate underlying, 'silent misinterpretations' that may vary in frequency or impact across cultural groups. In addition, only a small number of behavior codes were examined in this study. We also cannot rule out the possibility that the meaning of the various behaviors captured by our behavior codes may vary across cultural groups

(Johnson et al., 2006). Available research on this topic, however, reveals similar levels of reliability, validity, and conceptual equivalence for behavior codes across multiple racial/ethnic groups (Johnson, Cho, & Holbrook, 2010; Johnson et al., 2011). We were also unable to examine interviewer characteristics (beyond whether each question was read correctly), which might also be associated with respondent behaviors.

We believe that the current study buttresses a growing literature that focuses both on respondent and item characteristics as potential sources of survey response error. Our results suggest a reconsideration of the relative influence of these factors. Questions featuring characteristics that are problematic present similar difficulties to all respondents. Rather than focusing on cross-cultural comparability only, we may be better served by also paying attention in the first place to basic practices of good questionnaire design, such that we are able to benefit all of our respondents.

References

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken, NJ: John Wiley. Doi:10.1002/9780470146316.

Andrews, F. M., & Herzog, A. R. (1986). The quality of survey data as related to age of respondent. *Journal of the American Statistical Association, 81*, 403–410.

California Health Interview Survey. (2009). *CHIS 2007 methodology series: Report 1—Sample design*. Los Angeles, CA: UCLA Center for Health Policy Research.

Cannell, C. F., Lawson, S. A., & Hausser, D. A. (1975). *A technique for evaluating interviewer performance*. Ann Arbor: Survey Research Center, University of Michigan.

Cannell, C. F., Fowler, F. D., & Marquis, K. H. (1968). The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interviews. *Vital and health statistics, Series 2: Data evaluation and methods research, no. 26*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Dijkstra, W. (2008). Behavior coding. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 53–55). Los Angeles: Sage. Doi.org/10.4135/9781412963947.

Fowler, F. J. (2011). Coding the behavior of interviewers and respondents to evaluate survey questions. In J. Madans, K. Miller, A. Maitland & G. Willis (Eds.), *Question evaluation methods: Contributing to the science of data quality*. New York: Wiley. Doi:10.1002/9781118037003.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 15–36). San Francisco: Jossey-Bass.

Holbrook, A. L., Cho, Y. I., & Johnson, T. P. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties. *Public Opinion Quarterly, 70*, 565–595. doi:10.1093/poq/nfl027.

Johnson, T. P., Cho, Y. I., & Holbrook, A. (2010). *A preliminary assessment of cultural variability in respondent actions captured by behavior codes*. Paper presented at the Joint Statistical Meetings, Vancouver.

Johnson, T. P., Cho, Y. I., Holbrook, A., O'Rourke, D., Warnecke, R., & Chávez, N. (2006). Cultural variability in the effects of question design features on respondent comprehension of health surveys. *Annals of Epidemiology, 16*, 661–68. doi:10.1016/j.annepidem.2005.11.011.

Johnson, T. P., Holbrook, A. L., Shavitt, S., Cho, Y. I., Chávez, N., & Weiner, S. (2011). *Cross-cultural validity of behavior codes*. Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix.

Kudela, M. S., Stark, D., Hantmann, J., Seak, M. A., & Newsome, J. (2009). *Behavior coding of the 2007 CHIS discrimination module: Final report*. Rockville, MD: Westat.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congodon, R. T. (2011). *HLM7: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc..

Reeve, B. B., Willis, G., Shariff-Marco, S., Breen, N., Williams, D. R., Gee, G. C., Alegria, M., Takeuchi, D. T., Kudela, M. S., & Levin, K. Y. (2011). Comparing cognitive interviewing and psychometric methods to evaluate a racial/ethnic discrimination scale. *Field Methods, 23*, 397–419. doi:10.1177/1525822X11416564.

Rodriguez, G., & Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *Stata Journal, 3*, 32–46.

Schaeffer, N. C., & Dykema, J. (2011). Response 1 to Fowler's chapter: Coding the behavior of interviewers and respondents to evaluate survey questions. In J. Madans, K. Miller, A. Maitland & G. Willis (Eds.), *Question evaluation methods: Contributing to the science of data quality* (pp. 23–39). New York: Wiley. Doi:10.1002/9781118037003.

Shariff-Marco, S., Gee, G. C., Breen, N., Willis, G., Reeve, B. B., Grant, D., Ponce, N. A., Krieger, N., Williams, D. R., Landrine, H., Alegria, M., Mays, V., Johnson, T. P., & Brown, E. R. (2009). A mixed-methods approach to developing a self-reported racial/ethnic discrimination measure for use in multiethnic health surveys. *Ethnicity and Disease, 19*, 447–53.

Shariff-Marco, S., Breen, N., Landrine, H., Reeve, B. B., Krieger, N., Gee, G. C., Williams, D. R., Mays, V. M., Ponce, N. A., Alegria, M., Liu, B., Willis, G., & Johnson, T. P. (2011). Measuring everyday racial/ethnic discrimination in health surveys: How best to ask the questions, in one or two stages, across multiple racial/ethnic groups? *Du Bois Review: Social Science Research on Race, 8*, 159–177. doi:10.10170S1742058X11000129.

Van der Zoewen, J., & Smit, J. H. (2004). Evaluating survey questions by analyzing patterns of behavior codes and question-answer sequences. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 109–130). New York: Wiley. Doi:10.1002/0471654728.

Biographical Notes

Timothy P. Johnson is a Professor of Public Administration, Director of the University of Illinois at Chicago (UIC) Survey Research Laboratory and a Fellow at the UIC Institute for Health Research and Policy.

Salma Shariff-Marco is a Research Scientist at the Cancer Prevention Institute of California, Consulting Assistant Professor in the Dept. of Health Research & Policy at Stanford University School of Medicine, and an Associate Member of the Cancer Prevention and Control Program at the Stanford Cancer Institute.

Gordon Willis is a Cognitive Psychologist in the Applied Research Program of the Division of Cancer Control and Population Sciences at the National Cancer Institute.

Young Ik Cho is an Associate Professor at the University of Wisconsin at Milwaukee.

Nancy Breen is an Economist in the Health Services and Economics Branch of the Applied Research Program of the Division of Cancer Control and Population Sciences at the National Cancer Institute.

Gilbert C. Gee is an Associate Professor in the Department of Community Health Sciences in the School of Public Health at the University of California, Los Angeles.

Nancy Krieger is a Professor in the Department of Society, Human Development and Health at the Harvard School of Public Health.

David Grant is the Director of the California Health Interview Survey at the UCLA Center for Health Policy and Research.

Margarita Alegria is the Director of the Center for Multicultural Mental Health Research and a Professor of Psychology in the Department of Psychiatry at Harvard Medical School.

Vickie M. Mays is Professor in the Department of Psychology in the College of Letters and Sciences, as well as a Professor in the Department of Health Services in the School of Public Health. She is also the Director of the UCLA Center on Research, Education, Training and Strategic Communication on Minority Health Disparities.

David R. Williams is the Florence Sprague Norman & Laura Smart Norman Professor of Public Health and a Professor of African and African American Studies and of Sociology at Harvard University.

Hope Landrine is Professor in the Department of Psychology and the Director of the Center for Health Disparities Research at East Carolina University.

Benmei Liu is a Mathematical Statistician in the Surveillance Research Program of the Division of Cancer Control and Population Sciences at the National Cancer Institute.

Bryce B. Reeve is an Associate Professor in the Department of Health Policy and Management at the Gillings School of Public Health at the University of North Caroline at Chapel Hill and member of the Cancer Prevention and Control program at the Lineberger Comprehensive Cancer Center.

David Takeuchi is Professor and the inaugural Dorothy Book Scholar at Boston College Graduate School of Social Work, Boston College. He is also Associate Dean for Research.

Ninez A. Ponce is the Principal Investigator of the California Health Interview Survey at the UCLA Center for Health Policy and Research and an Associate Professor in the Department of Health Services at the UCLA School of Public Health.