

Improving Question Wording in Surveys of Culturally Diverse Populations

RICHARD B. WARNECKE, PhD, TIMOTHY P. JOHNSON, PhD, NOEL CHÁVEZ, PhD, SEYMOUR SUDMAN, PhD, DIANE P. O'ROURKE, MA, LORETTA LACEY, DR PH, AND JOHN HORM, MS

PURPOSE: The purpose of this paper is to briefly describe a theoretical model articulating cognitive theory and sources of potential response bias resulting from racial or ethnic cultural experience to survey questions that deal with health behavior. The theory components are then evaluated using questions obtained from national health surveys conducted by the National Center for Health Statistics and Centers for Disease Control and Prevention. The analysis explores the effects of four cognitive tasks involved in responding to questions as specified by the model: question interpretation, information retrieval from memory, judgment formation, and response editing. Implications for epidemiological research are considered.

METHODS: Data were collected from a purposive sample of 423 adults aged 18 through 50 who were recruited to ensure equal numbers of African American, Puerto Rican, Mexican American, and non-Hispanic white respondents, stratified by age, gender, and education. Individual questions were selected for evaluation to ensure variation by topic and question format. Probes related to each of the cognitive tasks were designed to obtain insight into the underlying cognitive processes used by respondents to answer survey questions. All statistical analyses used logistic regression or ordinary least squares multiple regression as appropriate. RESULTS: Variation by race/ethnicity was found in the way respondents defined physical activity in a series of questions used in the Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System (BRFSS). Gender and race/ethnicity appeared to influence interpretation in the absence of specific cues in the question format about how to respond. Strategies used to retrieve information from memory did not appear to be influenced by respondent culture; however, frequency of the event was associated with the recall strategy in that more frequent or regular events were more likely to result in estimates about frequency, whereas unusual or seldom occurring events were counted. Effects of race/ethnicity on judgment formation seem to be reflected in the propensity of respondents'

of the event was associated with the recall strategy in that more frequent or regular events were more likely to result in estimates about frequency, whereas unusual or seldom occurring events were counted. Effects of race/ethnicity on judgment formation seem to be reflected in the propensity of respondents' willingness to use extreme response categories. Most effects due to race/ethnicity were found in respondent editing of answers. Race/ethnicity was found to be associated with a social desirability trait; with willingness to disclose socially undesirable behavior, particularly to interviews from racial or ethnic groups that differed from the respondent; and with the tendency to overreport socially desirable behavior. CONCLUSIONS: Overall, the results of this research suggest several ways in which the validity of questions about risk behavior can be improved. In designing such questions, the investigator should envision the interview as a structured conversation in which ordinary conversational norms apply. Thus, questions that might request redundant information or that are threatening to the respondent need to be asked in ways that minimize these effects. Using interviewers of the same racial or ethnic group is important. Attending to the order of questions to ensure that redundant information is not requested is important. Writing questions to ensure that where response cues occur they lead the respondent to answer in unbiased ways is also important. Testing questions for potential racial or ethnic bias before using them is also important, even if the questions have been used successfully with population groups other than that or those included in a study.

Ann Epidemiol 1997;7:334–342. © 1997 Elsevier Science Inc.

KEY WORDS: Cognitive Tasks, Health Survey, Information Retrieval, Judgment Formation, Questionnaire Design and Interpretation, Response Editing.

Received March 3, 1997; accepted April 4, 1997.

INTRODUCTION

In 1979, Leon Gordis noted that there was increasing sophistication in the statistical techniques and data processing strategies used by epidemiologists, but that too little attention was being paid to the quality of interview and questionnaire data. He argued that as a result, the quality of the raw data was weakened. He called for more complete publication of data collection instruments, peer review of existing and

From Survey Research Laboratory, University of Illinois, Chicago, IL (R.B.W., T.P.J., S.S., D.P.O'R.); School of Public Health, University of Illinois, Chicago, IL (N.C., L.L.); College of Commerce and Business Administration, University of Illinois, Urbana-Champaign (S.S.); and the National Center for Health Statistics, Bethesda, MD (J.H.).

Address reprint requests to: Richard B. Warnecke, PhD, Survey Research Laboratory, University of Illinois, Chicago, IL, 910 W. Van Burin, 500, MC 336, Chicago, IL 60607.

Selected Abbreviations and Acronyms

CNSTAT = Committee on National Statistics
ZUMA = Zentrum für Umfragen Methoden und Analysen
QDRL = Questionnaire Design Research Laboratory

proposed data collection instruments, and establishment of a bank of tested questionnaires to be maintained under the auspices of the Society for Epidemiologic Research (1). While clearly standardization and peer review of questionnaires can contribute to the comparability of data across studies and improved question wording, these measures will not improve respondent understanding and response to the questions, which are also critical to the quality of the raw data.

At about the same time that Gordis's paper was published (1), an international group of university scholars and government scientists interested in cognitive psychology and survey methods began to discuss the common relevance of their ongoing research to the work in each others' fields (2). This discussion produced two important meetings: one in the United States, convened by the National Academy of Sciences' Committee on National Statistics (CNSTAT) (3) and another in West Germany organized by the Zentrum für Umfragen Methoden und Analysen (ZUMA) (4). The result of these meetings has been the establishment of an international effort to examine the linkages between cognitive psychology and survey question validity. In the United States, the National Center for Health Statistics (NCHS) established the Questionnaire Design Research Laboratory (QDRL), the first permanent government laboratory in which cognitive or intensive interviews were conducted as a routine part of questionnaire development, and an associated research program that formalized the collaboration between cognitive psychologists and survey methodologists at NCHS. Over the next decade, cognitive laboratories emerged at the Bureau of the Census and at many other federal statistical agencies and academic and quasi-academic survey centers. The resulting collaboration between cognitive psychologists and survey statisticians has produced a theoretical model articulating cognitive processes used by respondents to answer survey interview questions. The purpose of this paper is to briefly summarize this model and its impact on how survey questions are answered by persons of varying racial and ethnic backgrounds. We briefly describe four tasks that respondents perform when answering questions. The effects of race and ethnicity on these tasks are then illustrated using questions from various federal surveys that are often used to collect histories of behaviors or exposures that might be associated with risk of disease.

Theory of Survey Response

The four basic tasks respondents perform when they answer a survey question are (a) interpretation, (b) memory retrieval, (c) judgment formation, and (d) response editing (5).

Interpretation. Cross-cultural scholars distinguish between etic constructs, which are thought to be universal and understood in a common manner across all cultural groups, and emic constructs, which are culturally specific and have important meaning within certain cultural groups, but are understood either differently or not at all by other cultural groups. In multicultural societies such as the United States, a so-called "standard question" presumed to be etic may in fact be emic and, hence, answered differently by respondents of varying educational, racial, or ethnic backgrounds. When constructs that are emic are treated as etic, a category fallacy (6) results, the practical significance of which may constitute a problem in generalizing data across respondents and/or a failure of the respondent to answer the question being asked by the investigator. As a result, the respondent and questioner may have different understandings of the question's meaning, even though the question is in a language common to both. If there is variation in how a question is understood among respondents or between the respondent and investigator, the resulting data are vulnerable to misinterpretation. Moreover, if the interpretive context being used by respondents is not clarified, the problem will be further exacerbated across studies, if, as is sometimes the case, "equivalent" wording is substituted for exact replication. At that point, interpretation is further confounded with variation in guestion wording. Finally, interpretation of identically worded questions across studies may be influenced by order effects resulting from variation in placement of questions in different versions of the questionnaire.

Given that individuals use their cultural experience to interpret and respond to information and requests for information, the call for researchers to standardize their questions across studies invites category fallacy. If questions presumed to be etic are in fact emic because of the influence of such respondent characteristics as race, ethnicity, or education, then it is important to evaluate questions used in epidemiologic surveys for their susceptibility to category fallacy each time they are used with a new population of respondents.

Memory Retrieval. Once understood or interpreted, a reply to a question requires that either the answer or the information required to formulate the answer be retrieved from memory. Information may be retrieved either episodically, as discrete events, or semantically, in the form of schemas or generalizations (7). Episodic data are stored as detailed, specific, or unique information; schemas are generic descriptions of classes of events and are stored with little or no detail about specific episodes. Thus, frequent, routine, or regularly recurring events are more likely to be reported as schemas than are unique, unusual, or otherwise remarkable events (8). Frequently, schemas are also culturally conditioned and are reported as representative events of the respondent's community or cultural tradition.

Judgment Formation. When a precise response to a question can be accessed directly during memory retrieval, no

further judgment is required. However, when the task requires a judgment to be formed about information that has been retrieved from memory, the task becomes more complex and is likely to be affected by racial or cultural background. For example, in assessing perceived risk, the respondent has to make several kinds of judgments including whether there has been an exposure to something that creates a risk and then whether the exposure actually creates a risk. Then, the respondent is in a position to assess whether the risk exists. In sum, judgment formation processes may involve rating, estimating frequency of the event, estimating confidence in the responses, agreeing or disagreeing with a particular position, or evaluating the relative importance of conflicting information retrieved from memory in order to form a judgment. Generally speaking, the more the respondent has thought about the topic for which a judgment is requested, the more readily information retrieved is used for making judgments (9). On the other hand, because of the overwhelming amount of information to which a person is exposed, not all of it may be accessible for judgment formation.

Forming judgments is influenced by cues contained in the question. The range of a scale, for example, may indicate to the respondent the investigator's expectations about frequency (10). The choices made by respondents to the metric in a scale may be culturally influenced; for example, African American and Hispanic survey respondents are less likely to qualify their responses than are non-Hispanic whites (11, 12), and use of modifiers by Hispanics may increase with acculturation (13). In contrast, Asians tend to avoid extreme responses (14). Such preferences for various forms of response style have been attributed to cultural variation in conversational norms, such as emphasis on sincerity as opposed to emphasis on modesty in social interaction (15). Other research suggests that cultural variation in probabilistic thinking (e.g., the ability to express thoughts in terms of the degree of uncertainty) may also affect judgment formation (16).

Conversational norms about avoiding redundant information and providing new information also operate in forming judgments (17). For example, in a study of how cancer patients assess quality of life while undergoing chemotherapy, we led respondents to make distinct judgments about their satisfaction with their spouse and about their sex life by how we ordered these two questions. By asking respondents about satisfaction with their sex life before asking about satisfaction with their spouse, we avoided a potential redundancy about the sexual aspect of spousal relationships. Probing indicated that by putting the question about the respondents' sex life first, we obtained better information about satisfaction with other aspects of the marital/partner relationship, since the question about their sex life had already been considered. This was particularly true among Mexican

American respondents for whom this redundancy was particularly salient (18).

Response Editing. When respondents are concerned about the social acceptability of their judgments, response editing is often encountered (19). For example, socially desirable behaviors such as exercise and good nutrition may be frequently overreported, whereas undesirable behaviors such as drinking or smoking may be underreported. Available information suggests that definitions of socially desirable behavior vary culturally (20, 21). Thus, socially desirable response patterns may be compatible with the commonly observed pattern of social interaction in Hispanic cultures referred to as *simpatía*, the expectation that interpersonal relationships will be guided by harmony and the absence of confrontation (22). Similar expectations have also been found to affect how Asians respond to surveys (23).

We have already shown how conversational norms affect judgment; they also affect editing. For example, acquiescence, or the tendency to agree with a statement regardless of content, is related to social desirability, and hence, acquiescence is a form of editing often observed as a strategy employed by minority respondents when they perceive social distance between themselves and the interviewer (15, 24). Acquiescence may also result from the category fallacy: When respondents are unsure about what the question means, they may "play it safe" by acquiescing rather than risk appearing foolish or ignorant by admitting they do not understand the question (25).

Generally speaking, all forms of editing, including acquiescence, are likely to occur in situations where there is social or cultural distance between the interviewer and respondent due to gender, ethnicity, educational level, or other status indicators (26, 27). There is also evidence that bilingual respondents may edit, depending upon the language in which the question is asked and the cultural significance of the question (28).

METHODOLOGY

In summary, this paper will examine how respondent's race and ethnicity influence the performance of four cognitive tasks involved in responding to survey questions: question interpretation, information retrieval, judgment formation, and editing. The data presented here were obtained as part of an inquiry into cultural differences in social cognition (25), funded by Cooperative Agreement #U83/CCU508663 between the University of Illinois Survey Research Laboratory and the National Center for Health Statistics. The study population for this research was recruited from the four largest cultural groups in the Chicago metropolitan area: African Americans, Mexican Americans, Puerto Ricans, and non-Hispanic whites. Mexican Americans and Puerto Ricans were selected separately in recognition of

TABLE 1. Logistic regression analysis of physical activity question meaning probes (Unstandardized Coefficients Reported) *Survey Question:* "What types of physical activity or exercise did you perform during the past month?" Probe: "Which, if any, of the following would you (also) consider to be physical activity?"

			Work related	
Variable	Walking	Housework	activity	Yard work
Behavioral frequency	0.22*	0.02	0.05	0.11
Male $(1 = Yes)$	-0.90°	-0.37^{b}	0.44	-0.41
African American (1 = Yes)	0.08	0.44 ^b	-0.04	0.59 ^h
Mexican American (1 = Yes)	0.45	0.26	0.26	0.42
Puerto Rican $(1 = Yes)$	0.16	0.35	-0.23	0.40
Age	0.07*	0.01	-0.01	0.00
Education	0.28	0.16	0.04	0.02
Income	-0.22	-0.06	-0.28	-0.10
Model chi-square	33.41°	19.80 ^b	12.38	19.06 ^b
Sample size	357	364	363	363

 $^{^{\}circ}p < .05.$

their many linguistic and cultural differences (24), despite the core culture common to all persons of Hispanic origin. Cognitive interviews were conducted with 423 adults aged 18 through 50. Questions evaluated cognitively were selected from a large pool of health questions previously used in national health surveys. Individual items were deliberately chosen for variation in question topic and format. Specific probes appropriate to each of the four cognitive tasks were designed to obtain insights into the underlying cognitive processes used by participants when answering the substantive questions (29).

Interviews were conducted by the research team of investigators and research assistants at the Survey Research Laboratory at the University of Illinois at Chicago. Respondents were recruited primarily through media ads and community organizations. Respondent selection ensured that approximately one-quarter represented each cultural group (Mexican American, Puerto Rican, African American, and non-Hispanic white) and also allowed stratification by gender, age, and education. Screened as ineligible were: (a) persons who were currently employed by a physician or other health care provider; (b) those under the age of 18 or over the age of 50; and (c) persons not self-identifying as African American, Hispanic, or non-Hispanic white. Hispanics or Latinos who self-identified as other than Mexican American (or Mexican or Chicano/a) or Puerto Rican were also ineligible.

Interviews were conducted between July 1993 and April 1994 in English by same-sex interviewers and averaged approximately an hour in length. With respondent consent, each interview was tape-recorded and transcribed. During transcription, responses to unstructured probes were reviewed by several members of the research team and assigned codes representing the content of each respondent's answer.

Multiple codes were used for most probes to capture as much information as possible.

The interview questions were selected from various surveys conducted by the National Center for Health Statistics and included the Health Interview Survey, the Behavioral Risk Factor Surveillance System (BRFSS), the National Household Survey on Drug Abuse (NIDA), the CES (D) Depression Scale, and other various surveys.

Bivariate analyses were done on all of the questions for which the effects were analyzed. Final models were developed using logistic regression or ordinary least squares (OLS) as appropriate. In the following section are described the techniques used for each analysis.

RESULTS

Question Interpretation

The introduction to a series of questions about physical activity that are asked in the Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System (BRFSS) is open ended and reads as follows: "What types of physical activity or exercise did you perform during the past month?" The respondent is then asked how often in the past month each listed activity was performed. Variation by race/ethnicity was found in respondent interpretation of what constitutes physical activity through probes asking whether they considered walking, housework, work-related activity, and vard work to be physical activities. In logistic regression analyses controlling for age, gender, education and income, housework was more often included as a physical activity by African American respondents than by non-Hispanic white respondents and was less often considered physical activity by men than by women (see Table 1). Men included work-related activity as physical activity but

 $^{^{+}}p < .01.$

 $^{^{\}circ}p < .001.$

considered yard work to be physical activity less often than females did. However, African Americans were more likely to consider yard work physical activity than were non-Hispanic whites.

The open-ended question about physical activity gave the respondent no cues about the expected content of responses. Gender and race/ethnicity apparently influenced the interpretation of physical activity in the absence of specific cues from the question format. If the question provided broad response categories of kinds of physical activity, the respondents might have made similar interpretations of what the investigator was expecting by way of response. Thus, an alternative wording might be: "Did you perform any of the following kinds of physical activity during the past month?" Answer categories might refer to broad groupings of activity under which respondents could list specific activities such as sports, housework, yard work, or work-related activities. This format provides distinct cues about what information is desired, does not require the respondent to decide what is a physical activity, and may help circumvent some of the cross-cultural differences in question interpretation likely to be found in the open-ended format.

Recall Strategies

As noted earlier, recall tasks are usually performed using either episodic or semantic recall strategies. We examined the strategies used to answer several behavioral frequency questions, including questions on physical activity, cigarette use, telephone calls, physician visits, and marijuana use. Using logistic regression models, we found no effects of respondent culture upon the recall strategy used. The reported frequency with which an event was performed, however, was consistently associated with recall strategy, with greater frequency always predictive of semantic recall.

Judgment Formation

As discussed earlier, previous research has found culture to influence the use of extreme responses in reporting judgments on rating scales. An 18-item health locus-of-control scale was included to measure use of extreme responses (30). Each item had four response options: "strongly agree," "somewhat agree," "somewhat disagree," and "strongly disagree." Extreme response preference was assessed by summing the number of times the respondent chose the "strongly agree" or "strongly disagree" categories. All three groups of minority respondents were more likely than the non-Hispanic white respondents to choose extreme response categories. Puerto Rican respondents chose an extreme category an average of 8.7 times, Mexican Americans averaged extreme choices 8.1 times, and African Americans averaged 7.8 extreme responses. In contrast, the non-Hispanic white respondents chose an extreme category an average of 6.7 times. Using OLS regression to control for age,

TABLE 2. Multiple regression analysis of predictors of extreme response style scale

Variable	В	(SE)	ВЕТА
Male (1 = Yes)	-0.21	0.43	-0.02
African American (1 = Yes)	0.98	0.60	0.10
Mexican American $(1 = Yes)$	1.43	0.60	0.15°
Puerto Rican (1 = Yes)	1.56	0.65	0.14^{a}
Age	0.03	0.03	0.07
Education	-0.60	0.20	-0.16^{a}
Income	-0.21	0.15	-0.08
R ² (adjusted)		0.06	
F-value	4.40 ^b		
Sample size	384		

 $^{^{}a} p < .01.$

B, unstandardized Beta; (SE), standard error; BETA, standardized Beta.

gender, education, and income, cultural differences in extreme response preference were found between non-Hispanic whites and both Mexican American and Puerto Rican respondents. These findings are presented in Table 2.

Response Editing

A consistent finding in the literature on ethnicity and response to surveys has been a relationship between race/ethnicity and social desirability. All respondents were asked to complete a Marlowe-Crowne social desirability questionnaire (31) where we assessed the likelihood with which they would answer questions in a socially desirable manner (i.e., respond positively to questions about socially desirable behaviors and attitudes and negatively to those considered to be undesirable). Using multiple regression, (Table 3) we examined the predictors of this social desirability trait and found gender and cultural effects: Being male was negatively associated with the social desirability trait, whereas being African American or Mexican American was positively re-

TABLE 3. Multiple regression analysis of predictors of Marlow-Crowne social desirability scale

Variable in equation	В	(SE)	BETA	
Male (1 = Yes)	-0.65	0.23	-0.14^{a}	
African American (1 = Yes)	0.75	0.31	0.14a	
Mexican American (1 = Yes)	1.16	0.32	0.22	
Puerto Rican (1 = Yes)	0.46	0.34	0.08	
Age	0.06	0.01	0.20b	
Education	-0.12	0.11	0.06	
Income	0.051	0.08	0.04	
R ² (adjusted)		0.06		
F-value		4.96 ^b		
Sample size		405		

² p < .01.

b p < .001

b p < .001

B, unstandardized Beta; (SE), standard error; BETA, standardized Beta.

lated. Older respondents were also more likely to express socially desirable behaviors and attitudes.

Using multiple regression, (not shown) we then modeled predictors of frequency of reported physical activity, a socially desirable behavior. The strongest predictor in the model was the social desirability trait. The positive correlation between the scale score for social desirability and frequency of reported physical activity suggests that editing of the reported frequency of these kinds of activities in the direction of appearing to be more active may be taking place. Since race/ethnicity influences social desirability, this editing is likely to occur most often among African American and Mexican American respondents who, relative to non-Hispanic whites, were found to demonstrate the social desirability trait.

We also used projective probes—asking about persons other than the respondent—to assess the extent to which respondents might select socially desirable responses. Projective probes have been found to reduce respondent uneasiness if responding to threatening questions about their own behavior (19). Two examples explore a socially desirable behavior and a socially undesirable behavior. The socially desirable question asked about number of portions of vegetables the respondent consumed daily. Respondents were first asked about how many servings of vegetables they consumed daily. Then they were asked, "In general, do you feel that people might purposely say they eat more vegetables than they do, say they eat fewer vegetables than they do, or would they try to answer accurately?" A majority (57%) felt that people would report accurately the amount of vegetables they had eaten. Smaller proportions felt people would overreport (29%) or underreport (11%). About two-thirds (68%) of the non-Hispanic white respondents said that most would accurately report; 60% of the Mexican Americans shared this opinion, while smaller proportions of Puerto Rican (50%) and African American (51%) respondents felt that others would try to respond accurately. A logistic regression model that controlled for age, gender, education, income, and vegetable consumption was not significant (Table 4), but nonetheless indicated that African American respondents were significantly more likely than Puerto Rican, Mexican American, and non-Hispanic white respondents to believe that editing would occur.

When a socially undesirable behavior, alcohol consumption, was examined, no racial/ethnic differences were apparent. However, a probe dealing with potential interviewer effects when the interviewer was from a different racial/cultural group than the respondent indicted potential cultural effects. Respondents were asked to assess how comfortable most members of their cultural group would be reporting their drinking patterns to an interviewer of their own cultural group or to an interviewer from a different cultural group. A large majority of all respondents (85%) indicated that members of their cultural group would be comfortable

TABLE 4. Logistic regression analysis of projected misreporting of vegetable consumption

Survey Question: "About how many servings of vegetables do you eat per day or per week, not counting salad or potatoes?" Probe: "In general, do you feel that people might purposely say they eat more vegetables than they do, say they eat fewer vegetables, or would they try to answer accurately?"

	Would misreport (1 = yes)	
Variable	В	(SE)
Reported vegetable consumption (per month)	-0.06	0.05
Male $(1 = Yes)$	0.03	0.11
African American (1 = Yes)	0.34	0.15
Mexican American (1 = Yes)	0.14	0.16
Puerto Rican $(1 = Yes)$	0.04	0.17
Age	0.02	0.01
Education	-0.08	0.10
Income	0.10	0.08
Model chi-square		12.55
Sample size		370

 $^{^{\}rm a} p < .05$.

B, unstandardized Beta; (SE), standard error.

talking about their drinking behavior to an interviewer from the same group. However, in a logistic regression model, the contrast between Mexican Americans and non-Hispanic white respondents was significant, holding other demographic traits and alcohol consumption constant. This finding suggested that Mexican Americans were less likely than white respondents to believe that members of their cultural group would be comfortable talking about their drinking behavior with an interviewer from their own cultural group.

Even greater differences were observed when these respondents were asked about discussing drinking behavior with an interviewer from a cultural group different than their own. In response to the question "Do you feel this is a question [drinks per day] that (RESPONDENT'S CUL-TURAL GROUP) would be comfortable or uncomfortable talking about with a survey interviewer who is not (RE-SPONDENT'S CULTURAL GROUP)?" over 90% of non-Hispanic white respondents reported that they felt others of their group would be comfortable. In comparison, only 60% of the African Americans felt that other African Americans would be comfortable and 60% of the Mexican Americans and 70% of the Puerto Ricans felt others from their cultural groups would be comfortable answering when the interviewer differed culturally. These findings were confirmed by logistic regression, controlling for age, gender, education, and self-reported alcohol consumption (Table 5).

The health locus-of-control scale was also used to assess the relationship between race and ethnicity, and the tendency to acquiesce. This assessment was made by summing the number of times each respondent chose the response

TABLE 5. Logistic regression analysis of projected respondent comfort discussing alcohol consumption *Survey Question*: "About how many drinks did you *usually* have in a *day* on the days that you drank during the past 30 days?" *Probe*: "Do you feel this is a question that (RESPONDENT'S CULTURAL GROUP) would be comfortable or uncomfortable talking about with: (a) a (RESPONDENT'S CULTURAL GROUP) survey interviewer; (b) a survey interviewer who is *not* (RESPONDENT'S CULTURAL GROUP)?"

	(a) Interviewer same culture (1 = would be uncomfortable)		(b) Interviewer same culture (1 = would be uncomfortable)	
	В	(SE)	В	(SE)
Reported alcohol consumption (days in past month)	0.37	0.22	0.27	0.16
Male $(1 = Yes)$	-0.57	0.183	-0.21	0.13
African American (1 = Yes)	0.27	0.26	0.77	0.18 ^b
Mexican American $(1 = Yes)$	0.79	0.24 ^b	0.78	0.19 ^b
Puerto Rican $(1 = Yes)$	0.30	0.26	0.55	0.19
Age	0.04	0.02	0.00	0.01
Education	0.08	0.15	-0.10	0.08
Income	-0.10	0.12	-0.10	0.08
Model chi-square	26.76 ^b		32.32°	
Sample size	347		365	

 $^{^{}a} p < .01.$

categories "strongly agree" or "somewhat agree," regardless of the question content, into a scale of respondent acquiescence. Cultural differences were found in the tendency to agree. African American and Mexican American respondents each gave an average of 10.0 positive responses; Puerto Ricans averaged 9.4. In contrast, non-Hispanic white respondents averaged 8.8 positive responses. An OLS regression model confirmed the tendency of African American and Mexican American respondents to give more acquiescent responses than non-Hispanic white respondents. Less educated, lower income, and male respondents were also more likely to be acquiescent (Table 6).

TABLE 6. Multiple regression analysis of predictors of respondent acquiescence scale

Variable in equation	В	(SE)	BETA	
Male (1 = Yes)	1.38	0.28	0.23 ^b	
African American (1 = Yes)	1.00	0.38	0.15°	
Mexican American (1 = Yes)	0.98	0.38	0.15a	
Puerto Rican (1 = Yes)	0.27	0.41	0.04	
Age	-0.02	0.02	-0.05	
Education	-0.55	0.13	-0.21^{b}	
Income	-0.33	0.09	-0.18^{b}	
R ² (adjusted)		0.18		
F-value		12.63 ^b		
Sample size		384		

p < .01.

DISCUSSION AND CONCLUSIONS

We have examined how race and ethnicity are associated with performance of the four tasks that cognitive theory indicates are associated with responding to survey questions. We found evidence in support of differences in question interpretation related to ethnicity. Our findings would suggest that providing cues in the question that help respondents understand what is needed may address these problems. We found no evidence that information retrieval in response to questions about frequency of behavioral acts was related to race or ethnicity. Our analysis of the effects of race or ethnicity on judgment formation supported the finding in the literature that the tendency to choose extreme response categories is related to race and ethnicity.

We also found a number of effects of race and ethnicity on responses to questions associated with health and risk behavior in the form of response editing. Overall, our analysis indicated that African Americans and Mexican Americans were most likely to exhibit the social desirability trait.

Further evidence of the variable effects of social desirability was found when we used projective probes in which respondents were asked to characterize the willingness of others in their cultural group to discuss socially desirable and socially undesirable behaviors. In response to a question about the consumption of vegetables, after controlling for age, gender, education, and vegetable consumption, we found that Mexican American, Puerto Rican, and African American respondents were more likely than non-Hispanic white respondents to believe that in general people would tend to overreport socially desirable behavior such as eating

b p < .001.

B, unstandardized Beta; (SE), standard error.

 $^{^{}b}p < .001.$

B, unstandardized Beta; (SE), standard error; BETA, standardized Beta.

vegetables. This finding was consistent with the finding that individuals expressing the social desirability trait tended to overreport physical activity, and African Americans and Mexican Americans were most likely to exhibit that trait.

We did not find direct effects of racial or cultural bias in response to a question about a socially undesirable behavior, alcohol consumption. However, we did find that Mexican American respondents believed that individuals in their ethnic group were less likely to feel comfortable discussing their alcohol consumption with interviewers from any cultural group. African American and Puerto Rican respondents also felt that others in their cultural group would be less comfortable discussing their drinking habits with interviewers of ethnic backgrounds different from their own, compared with non-Hispanic white respondents.

Overall, these findings suggest ways in which the quality of raw data, which so concerned Gordis, can be improved by conducting cognitive interviews with respondents about how they perform these four cognitive tasks when answering survey questions about risk or behavior. Survey interviews should be treated as structured conversations, subject to the same norms as normal conversations. Thus, before survey questions are finalized, conversational norms that inhibit providing redundant information, that might lead to editing responses to threatening questions, or that might in other ways intrude upon how the respondents interpret, form judgments about, and/or edit their responses to questions need to be explored. Qualitative methods such as those described here provide data that improve the validity of questions by giving insight into both how well the respondent understands the task presented by the question, what information is available to the respondent on which to base an answer, how that information is used to format a response, and how much of the judgment is likely to be reported in the interview.

Even when the questions are so-called standard items, that have a long history of use and a strong psychometric pedigree, think-aloud or cognitive interviews with structured probes can identify items where responses might be affected by racial or cultural experience. The results of these analyses indicate that even with questions that have been used with apparent success, without such analyses, researchers cannot be certain that responses are consistent across persons with varying cultural experiences.

Studying cognitive processes used by respondents will aid in the construction of cognitively equivalent questions. However, simply standardizing instruments without understanding the manner in which respondents provide information may lead to institutionalized invalidity that may be evident in the data only as weak associations between variables or lower odds ratios. The relevance of how the questions are being answered may be overlooked or dismissed. These are subtle effects but can certainly impede our understanding of the relationship between our data and how

respondents perform the tasks implicit in responding to our questions.

It should be noted that the interviews conducted for this project were not based on a random sample of respondents. Respondents were selected from those who answered advertisements designed to recruit individuals with specific characteristics. However, the findings reported here are being replicated in other similar studies. For example, in another study using this methodology we found that respondents who did not speak English were acquiescing to a measurement scale which contained words they did not understand. Bilingual respondents simply read these words in English. however, Mexican respondents who spoke no English simply selected a scale point at random. An analysis of the discriminant validity of the scale revealed such extensive overlap that the individual points had no discriminant validity. This could only have been ascertained by the methodology described here. Moreover, the authors of the scale in question had used it with other multicultural populations and found it to have very strong psychometric properties suggesting that the linguistic and cultural effects were being masked using "standard" psychometric measurement (18, 32).

Clearly, there is a need to evaluate the effects of such testing on the validity of data in a more systematic way than the current project permitted. To some extent we were able to do this with the Quality of Life study cited above. Moreover, the Survey Research Laboratory and many other survey groups are systematically evaluating all questions that are used in current studies and are finding variation in the level of understanding by race and ethnicity. The next phase will be to initiate studies where the principles reported here are evaluated across standard survey formats with more systematic assessment of the "pay-off."

Preparation of this paper was supported in part by Cooperative Agreement #U83/CCU508663 between the University of Illinois Survey Research Laboratory and the National Center for Health Statistics.

REFERENCES

- Gordis L. Assuring the quality of questionnaire data. Am J Epidemiol. 1979;109:21–24.
- 2. Jobe JB, Mingay DJ. Cognition and survey measurement: History and overview. Appl Cognitive Psychol. 1991;5:175–192.
- Jabine TB, Straf ML, Tanur JM, Tourangeau R, eds. Cognitive Aspects
 of Survey Methodology: Building a Bridge Between Disciplines. Washington, DC: National Academy Press; 1984.
- Hippler HJ, Schwarz N, Sudman S, eds. Social Information Processing and Survey Methodology. New York: Springer—Verlag; 1987.
- Strack F, Martin FF. Thinking, judging and communicating: A process account of context effects in attitude surveys. In: Hippler H-J, Schwarz N, Sudman S, eds. Social Information Processing and Survey Methodology. New York: Springer-Verlag; 1987:123-148.
- Kleinman A. Depression, somatization, and the "new cross-cultural psychiatry." Soc Sci Med. 1977;11:3–10.

- Tulving E. Elements of Episodic Memory. New York: Oxford University Press; 1983.
- Menon G. Judgments of behavioral frequencies: Memory search and retrieval strategies. In: Schwarz N, Sudman S, eds. Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer-Verlag; 1994:161–172.
- Dovidio JF, Fazio RH. New technologies for the direct and indirect assessment of attitudes. In: Tanur J, ed. Questions About Questions: Inquiries Into the Cognitive Bases of Surveys. New York: Russell Sage Foundation; 1992:204–237.
- Schwarz N. Assessing frequency of reports of mundane behaviors: Contributions of cognitive psychology to questionnaire construction. Res Methods Pers Soc Psychol (Rev Pers Soc Psychol). 1990;11: 98–119.
- 11. Bachman JG, O'Malley PM. Black-white differences in self esteem: Are they affected by response styles? Am J Sociology. 1984;90:624–639.
- Hui CH, Triandis HC. Effects of culture and response format on extreme response style. J Cross-Cultural Psychol. 1989;20:296–309.
- 13. Marín G, Gamba RJ, Marín BV. Extreme response style and acquiescence among Hispanics: The role of acculturation. J Cross-Cultural Psychol. 1992;23:498–509.
- Zax M, Takahashi S. Cultural influences on response style: Comparisons of Japanese and American high school students. J Soc Psychol. 1967;71:3–10.
- Marín G, Marín BV. Research with Hispanic Populations. Newbury Park, CA: Sage;1989.
- Wright GN, Phillips LD, Whalley PC, Choo GT, Ng KO, Tan I, Wisudha A. Cultural differences in probabilistic thinking. J Cross-Cultural Psychol. 1978;9:285–299.
- 17. Schwarz N. Judgment in a social context: Biases, shortcomings, and the logic of conversation. Adv Exp Soc Psychol. 1994;26:123–162.
- Warnecke RB, Ferrans CE, Johnson TP, et al. Measuring quality of life in culturally diverse populations. Monogr Natl Cancer Inst. 1996;20:29–39.
- 19. Bradburn NM, Sudman S. Improving Interview Method and Question-

- naire Design: Response Effects to Threatening Questions in Survey Research. San Francisco: Jossey-Bass;1979.
- Angel R, Thoits P. The impact of culture on the cognitive structure of illness. Cult Med Psychiatry. 1987;11:465

 –494.
- Dohrenwend B. Social status and psychological disorder: An issue of substance and an issue of method. Am Sociological Rev. 1966;31: 14–34.
- Triandis HC, Marín G, Lisansky J, Betancourt H. Simpatía as a cultural script for Hispanics. J Pers Soc Psycho. 1984;47:1363–1375.
- Deutcher I. Asking questions: Linguistic comparability. In: Warwick D, Osherson S, eds. Comparative Research Methods. Englewood Cliffs, NJ: Prentice Hall;1973:163–186.
- Aday LA, Chiu GY, Andersen R. Methodological issues in health care surveys of the Spanish heritage population. Am J Public Health. 1980;70:367–374.
- 25. Johnson TP, O'Rourke DP, Chávez N, et al. Social cognition and responses to survey questions among culturally diverse populations. In: Lyberg L, Biemer P, Collins M, et al. eds. Survey Measurement and Process Quality. New York: John Wiley & Sons;1997:87–113.
- Cotter PR, Cohen J, Coulter PB. Race-of-interviewer effects in telephone interviews. Public Opinion Q. 1982;46:278–286.
- Reese SD, Danielson WA, Shoemaker PJ, Chang TK, Hsu HL. Ethnicity-of-interviewer effects among Mexican-Americans and Anglos. Public Opinion Q. 1986;50:563–572.
- Bond MH, Yang KS. Ethnic affirmation versus cross-cultural accommodation: The variable impact of question language on Chinese bilinguals in Hong Kong. J Cross-Cultural Psychol. 1982;13:245–249.
- Belson WA. The Design and Understanding of Survey Questions. Aldershot, England:Gower;1981.
- Wallston KA, Wallston BS, DeVillis R. Development of the multidimensional health locus of control (MHLV) scales. Health Educ Monogr. 1978;6:160–170.
- 31. Strahan R, Martin LL. Short, homogenous versions of the Marlowe-Crowne social desirability scale. J Clin Psychol. 1972;28:191–193.
- 32. Ferrans CE, Powers MJ. Quality of Life Index: Development and Psychometric Properties. ANS Adv Nurs Sci. 1985;8:15–25.