Questionnaire versus Culture versus Demographics: What Accounts for Difficulties in Answering Survey Questions?

Gordon Willis, National Cancer Institute, NIH
Elaine Zahnd, Public Health Institute
Kerry Levin, Martha Kudela, Barbara Forsyth, Westat

To be presented at 3MC, Berlin
Session #1: Questionnaire Design
June 26, 2008

Most survey researchers will likely agree that cross-cultural *comparability* of survey measures is a worthy – and even vital – pursuit (Agans, Deeb-Sossa, & Kalsbeek, 2006; Behling, & Law, 2000; European Social Survey, 2002; Johnson, 1988; 2006; Schmidt, & Bullinger, 2003). Our objective is not necessarily to achieve cross-cultural 'equivalence' – as it may be too ambitious to believe that we literally will create exactly equivalent measures across disparate cultural groups. Rather, we strive to produce measures that are comparable in the sense that *the discrepancies across group-specific estimates represent 'true-score' differences as opposed to measurement artifacts that are only due to between-group measurement error.*

We further believe that it is important to distinguish the various potential sources of non-comparability across groups defined according to race, ethnicity, language, or other socio-cultural concepts. First, Harkness and Schoua-Glusberg (1998) in particular have promulgated the view that language translation – the means by which concepts are coded in the form of natural language – is often a potent source of between-group variation. Although in the larger perspective of questionnaire design and survey administration translation is often viewed as a relatively straightforward and even simple step, it is becoming increasingly apparent that appropriate translation is a *process*, rather than a single algorithmic step (Census Bureau, 2004; Forsyth, Kudela,, Levin, Lawrence, & Willis, 2007; Harkness, Van de Vijver, & Mohler, 2003;

McKay, Breslow, Sangster, Gabbard, Reynolds, Nakamoto, & Tarnai, 1996; Martinez, Marín & Schoua-Glusberg, 2006; Pan & de la Puente, 2005). Further, the translation process itself requires several intensive steps which require the thoughtful application of principles specific to the translation process. Currently, a consensus appears to be developing that the procedure previously believed to be the "state of the art" in translation – back translation (Brislin, 1970) – is insufficient in itself, and may even represent a flawed view of the ultimate objective of the translation process.[1] Others have explicated these ideas further, but we will note that back-translation tends to focus attention toward a literal, word-for-word equivalence of terms, when our true intent is to convert the overall meaning of each item, in context of the whole questionnaire.

As such, a preferred approach is one that explicitly focuses on the means by which the intended objective of the question can be represented in another language. Given that this in fact requires both an understanding of the measurement intent of the source question, and the means by which this can be conveyed in a different language in a manner that is consistent with the dictates of the standardized survey questionnaire, researchers have increasingly tackled this complex challenge through the use of a team approach that involves several individuals who each provide many (if not all) of the requisite skills. The team works closely with the designers to interpret the intent of the source question, often identifying sources of vagueness that preclude a single clear representation in a different language. As such, the translation process itself becomes a means by which to determine whether the source questions are sufficiently clear that they represent a single well-defined underlying concept that can in turn be represented by a different coding system (e.g., Is the English version clear enough that its intent can be uniquely represented in Spanish, or Chinese, or Korean?).

[1] The fact that this point was made by Brislin (1970) in his influential manuscript has often been overlooked.

Besides language translation, a second important threat to cross-cultural comparability may exist: As Warnecke, Johnson, Chavez, Sudman, O'Rourke, Lacey, & Horm (1997) have pointed out, it is possible that differences between groups may not be only a function of language, but as well of socio-cultural factors that supersede language. Hence, even when questionnaires are administered in English, Hispanics and Non-Hispanics may simply differ in their cognitive processing of some questions. Further, when administered in translated versions, some between-group variation may not be attributable only to failures of translation, but due to more general failure of the question to be "asking the same question" across group. To address this possibility, researchers now engage in a process of *cultural adaptation* – in brief, establishment of how the questions will function within the context of the culture(s) represented by disparate groups (Miller, 2004). For example, translations of questions on sexual practice and identification might be equally well understood across groups; but these may be received very differently depending on cultural standards of acceptability concerning discussions with a stranger about the underlying content matter.

We have proposed that careful translation procedures involving a group approach, and more general cultural adaptation, are vital procedures. However, all questionnaire development processes are imperfect – so researchers are also interested in testing and verification. The generally accepted approach to quality control of survey questionnaires is to incorporate empirical pretesting processes – in particular behavior coding and cognitive interviewing (Tourangeau, Rips, & Rasinski, 2000). Addressing the latter first, cognitive interviews are generally used early in the developmental process, largely with the intent of elucidating hidden, or covert sources of error associated with the cognitive processing of the evaluated questions (Forsyth & Lessler, 1991; Nápoles-Springer, Santoyo-Olsson, O'Brien, & Stewart, 2006;

Nápoles-Springer,, & Stewart, 2006; Willis, 2005). This approach can be represented as “the part of the iceberg below the surface” -- especially in terms of *silent misinterpretation* (DeMaio and Rothgeb, 1996) that goes undetected unless one uses cognitive probes to venture beyond that which is stated through the usual question asking-answering process. For example, an individual may silently misinterpret the term “dental sealant,” and this may be unapparent to either the interviewer or the respondent until the cognitive interviewer asks “What, to you, is a ‘dental sealant?”).

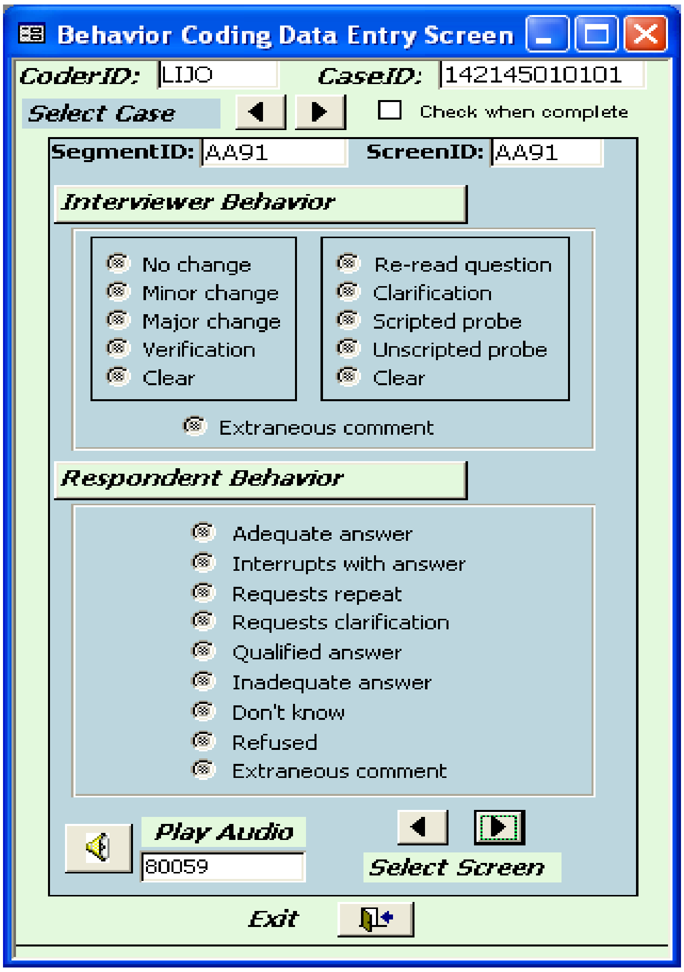
The focus of this paper will be on an alternative pretesting method – Behavior Coding (Cannell, Fowler, & Marquis, 1968) – intended mainly to assess that component of error that is observable, or *overt* -- and that has sometimes been termed “problems in the social interaction” between interviewer and respondent. As such, these problems can be represented, metaphorically, by the part of an iceberg that exists *above* the surface. The relationship between such problems and the existence of measurement (response) error is not necessarily self-evident, however. Developers of behavior coding techniques (Fowler and Cannell, 1996 in particular) have made the case that (a) questions that frequently produce observable problems such as inadequate answers, or request for clarification, are likely to be those that are not well-comprehended by respondents; and (b) there is a direct relationship between comprehension and data quality, in that questions that are poorly understood can be expected to produce response error. On this basis, practitioners of behavior coding argue that at least some overt problems in the interaction, if counted and quantified, serve as an indicator of error (somewhat similar to the way in which non-response rate is often used as an index of potential non-response bias).

More specifically, behavior coding is designed to systematically document both sides of the interaction: (a) problems experienced by interviewers in administering questions (in particular, reading them in a way that departs from the manner intended); and (b) problems experienced by respondents who are delivered the questions. The focus of the current study was the potential of behavior coding to elucidate potential sources of measurement error, and threats to cross-cultural comparability, in the multi-lingual survey environment.

Method

For the current investigation we applied behavior coding to the 2003 California Health Interview Survey, or CHIS – a random-digit-dial (RDD) telephone survey of approximately 40,000 California households which is designed by UCLA and administered by Westat (see Zahnd, Tam, Lordi, Willis, Edwards, Fry, & Grant, (2005). CHIS is an ideal test-bed for methodological work on survey pretesting techniques due to its wide inclusion of racial and ethnic groups, and the fact that it has been ambitious in terms of inclusion of respondents for whom English presents a language barrier. As such, the current study was piggybacked onto the 2003 fielded CHIS survey, such that interviews could be evaluated across five groups, representing to the degree possible a combination of race/ethnicity (Hispanics, Koreans, Other) and language (Spanish, Korean, and English): (1) Hispanics in Spanish; (2) Hispanics in English; (3) Koreans in Korean; (4) Koreans in English; and (5) Non-Hispanic, Non-Koreans in English. The use of these five groups of interest that allowed us to focus both on (a) language of administration and (b) racial group membership/ethnicity – which we will refer to simply as *Culture.*

Behavior coding consisted of the recoding of almost 100 interviews within each of the five groups. We selected a subset of 35 of these items to behavior-code that were of interest to the sponsor (NCI), and that were asked of all respondents (to maximize effective sample size for analysis). Behavior codes selected were those recommended by Fowler, covering both interviewer and respondent behaviors. A major positive feature of the investigation was the development of a digitized, semi-automated system for recording interviews and case-processing. Figure 1 depicts the behavior codes that applied to all coded items. For example, respondents might interrupt, request that the item be repeated, or answer the question with a response category that could not be coded into the categories provided by the respondent.



Behavior Coding Data Entry Screen

CoderID: LIJO   CaseID: 142145010101

Select Case ◀ ▶   □ Check when complete

SegmentID: AA91   ScreenID: AA91

Interviewer Behavior

○ No change
○ Minor change
○ Major change
○ Verification
○ Clear

○ Re-read question
○ Clarification
○ Scripted probe
○ Unscripted probe
○ Clear

○ Extraneous comment

Respondent Behavior

○ Adequate answer
○ Interrupts with answer
○ Requests repeat
○ Requests clarification
○ Qualified answer
○ Inadequate answer
○ Don't know
○ Refused
○ Extraneous comment

Play Audio 80059   ◀ ▶ Select Screen

Exit

Figure 1. Behavor Coding Data Entry Screen

Results

One notable finding was that permission to record the interview varied markedly across group (88-91% in English; 80% in Spanish) – and was found to be much lower for Korean-language interviews (59%). Follow-up investigation suggested that this was not so much because Koreans refused to be recorded, but rather because interviewers sometimes neglected to request consent to record. Hence, we have no particular reason to believe that differences between Koreans and others in behavior coding data was confounded by self-selection factors – although we admittedly have no data that allow us to make an absolute determination concerning this issue.

The first conclusion of note was that the CHIS interviews largely functioned as intended; looking across the universe of 15,433 recorded interaction, 92% of the survey questions were read as worded; that is, without a major misread that altered the question meaning. Also, 76% of the time, the question was answered with a response deemed to be adequate by coders. This of course does not indicate that the answer was correct, only that it produced no overt evidence of problems. Multiple coders were trained and used within each of the three languages: In general, inter-rater reliability was acceptable, although this was actually only at a ‘fair’ level for English language interviews, and was best across the two Korean-language coders. We did attempt to avoid confounding of language and coder by having all coders (including Spanish and Korean-language coders) conduct some coding of the English-language interviews (which controls somewhat for variation that is due to coder variance).

Despite the fact that behavior coding suggested that interviews for the most part functioned fairly well, Korean language interviews tended to be problematic. Although frequency of major misreads was negligent (2-3%) for non-Korean interviews, these occurred fully 33% of the time for Korean-language interviews. Follow-up analysis determined that one Korean interviewer in particular largely adopted her own versions of the questions rather than reading them as scripted. However, eliminating her interviews from the analysis did not come close to eliminating the observed discrepancy between Korean-Koreans and the other groups, as the value of major misreads remained over 20% for Koreans.

Subsequent qualitative review of the Korean translation revealed that the Korean interviewers were not misreading questions in order to spontaneously repair grossly erroneous translations. First, within this group we found a high correlation between interviewer misreads and respondent Problem codes, indicating that the interviewers' "version" of the question tended to create, as opposed to avoid, problems in the interaction. Further, the standardized versions were deemed appropriate by Korean-language reviewers, and the behavior coders reported that those interviewers tended to change questions by reading only segments of them.

Analysis. The usual approach to analyzing behavior coding data is to focus on each coded question, by tabulating problem codes, and flagging those that produce relatively high levels of misreads or problems for respondents. Given that the intent of the current study was not to pretest particular items, but rather to understand patterns related to problems in the interactions, we adopted a different analytic strategy. Our focus was on the relationship between particular independent variables – especially demographics and language -- and as dependent variables, the problems indicated by the behavior codes. Further, we are also interested in how measurable characteristics of survey questions influence respondent behavior in particular. In order to simplify the analysis, we endeavored to produce one overall measure of whether each interaction produced a problem for the respondent. Hence, although there were multiple codes representing non-optimal respondent behavior (e.g., interruption, request for clarification, inadequate answer, etc.), we pooled all of these into one indicator variable, which captured whether any type of behavior other than Adequate Answer had occurred. Hence, for each interaction between respondent and question (1-35), we recoded responses into either 0, if no evidence of a problem, or 1 if a problem was in evidence.

Following this, to produce an overall score for each respondent, we counted the number of questions receiving a '1' score. Because for various reasons not all respondents were administered all 35 questions, we divided this count by the number of questions administered to that respondent, to produce an overall Problem Score' that could in principle range from 0 to 100. We found this overall score to be illuminating when we computed its mean magnitude across each of our five defined groups (Table 1).

Table 1: Frequency of overall behavior codes indicating problems for respondents, by evaluated group.

Group	Percent of interviews with behavior Code indicating problem for respondent
Non-Hispanic/Korean in English:	20.3%
Hispanics in English:	17.9%
Hispanics in Spanish:	23.2%
Koreans in English:	18.9%
Koreans in Korean:	42.7%

We acknowledge that these values are simple univariate measures that are potentially driven by confounding variables, as the groups may have differed markedly by age, gender, or some other influential factor. Still, it is noteworthy that interviews of Koreans that were conducted in Koreans produced many more problems, relative to the other groups, which tended to cluster at around a 20% overall level. Recalling that Korean-language interviews were often read incorrectly, this again suggests that misreads of questions, rather than resolving problems, induced problems for respondents. In order to determine the independent influence of both reading error and demographic variables on our outcome measure, we conducted a series of linear regression analyses that utilized, as independent variables: (a) the frequency with which a Reading Error occurred (Major Misreads), and (b) a series of demographic variables. As the dependent variable we used the computed Problem Score for each respondent (as described above).

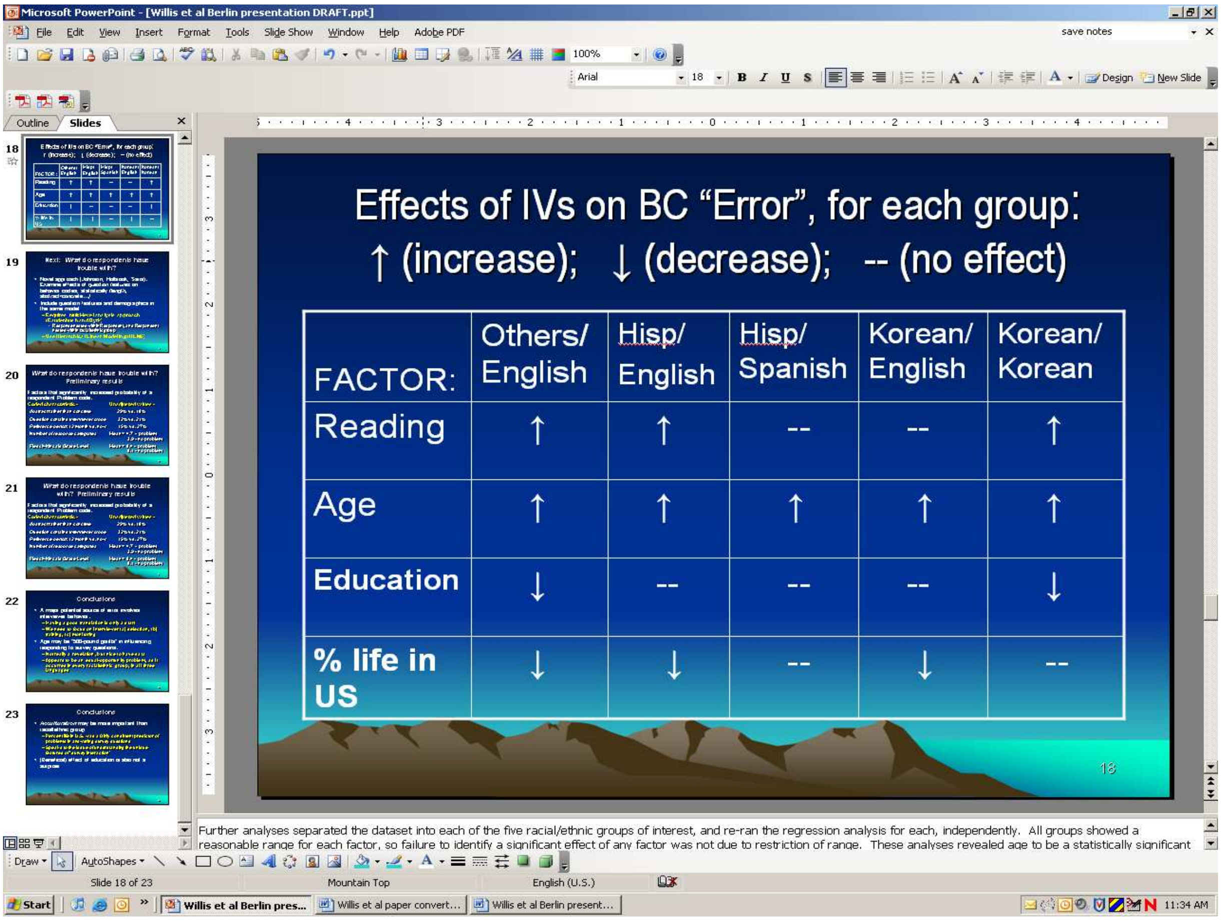
We also note that it was necessary for these analyses to transform our dependent variable, in order to produce a distribution exhibiting acceptable measurement characteristics: Due to skewness of the original Problem Score variable, we found that a square root transformation produced a variable that exhibited reasonable score distribution, as well as homogeneity of variance (and was superior to a natural log transform in this regard). Because the CHIS dataset contains a large number of demographic variables, we conducted a set of analyses which first assessed potential for confounding among independent variables (collinearity), then assessed first-order correlations between independent variables and our dependent measure, and then utilized a multiple regression strategy that involved a series of stepwise regression analyses. Further, analyses were done that both included, and excluded, the Korean-Korean respondent group, as that group was discrepant from the other racial/ethnic groups.

The result of a range of analyses, and combinations of different factors, repeatedly pointed toward a consistent set of predictors of problems for respondents: (a) the interviewer's failure to read the question correctly, (b) respondent age, (c) percent of his/her life the respondent had spent in the US, (d) Educational level, and to a lesser extent, (e) being a member of the Korean ethnic group (as self-reported within CHIS). (see Table 2). Together, these factors explained a fairly large amount of variance in the Problem Score measure, giving an (adjusted) multiple R-squared of .44, indicating that 44% of the variance in respondent behavior could be accounted for by these factors. Further analyses separated the dataset into each of the five racial/ethnic groups of interest, and re-ran the regression analysis for each, independently. All groups showed a reasonable range for each factor, so failure to identify a significant effect of any factor was not due to restriction of range. These analyses revealed age to be a statistically significant factor for every group. Other variables influenced each group to varying degrees: Reading errors were predictive for three of five groups; Educational level was predictive for two of five; and percent life in the US for three of five (Figure 2 illustrates these overall findings in a qualitative sense). Further analyses utilizing the entire dataset that searched explicitly for interaction effects between group membership and these demographic characteristics revealed little of further interest.

Table 2. Linear regression analysis: Significant predictors of respondent-based problems.

1) Reading errors *(risk factor)*	*b = .23*
2) Age *(risk factor)*	*b = .42*
3) *Percent life* spent in US *(protective factor)*	*b = -.18*
4) Educational level *(protective factor)*	*b = -.20*
5) Korean group membership *(risk factor)*	*b = -.21*

Question characteristics as predictors of problems in the exchange. The above results address the issue of *who*, in demographic terms, has problems with survey questions. A natural extension of the analysis is to ask additionally *what* they have trouble with, in terms of question characteristics. For example, are questions that are relatively long, or those that ask about abstract rather than concrete concepts, more problematic? This approach to understanding reactions to survey questions has been spearheaded by Tim Johnson, Alison Holbrook, and colleagues the University of Illinois; and by Wils Saris at the University of Amsterdam.


Microsoft PowerPoint - [Willis et al Berlin presentation DRAFT.ppt]
File Edit View Insert Format Tools Slide Show Window Help Adobe PDF
save notes
100%
Arial
18
Design
New Slide
Outline
Slides
Draw
AutoShapes
Slide 18 of 23
Mountain Top
English (U.S.)
Start
Willis et al Berlin pres...
Willis et al paper convert...
Willis et al Berlin present...
11:34 AM

Effects of IVs on BC "Error", for each group:
↑ (increase); ↓ (decrease); -- (no effect)

FACTOR:	Others/ English	Hisp/ English	Hisp/ Spanish	Korean/ English	Korean/ Korean
Reading	↑	↑	--	--	↑
Age	↑	↑	↑	↑	↑
Education	↓	--	--	--	↓
% life in US	↓	↓	--	↓	--

18

Further analyses separated the dataset into each of the five racial/ethnic groups of interest, and re-ran the regression analysis for each, independently. All groups showed a reasonable range for each factor, so failure to identify a significant effect of any factor was not due to restriction of range. These analyses revealed age to be a statistically significant

Figure 2. Depiction of effects of independent variables on frequency of problems identified through behavior coding.

A "holy grail" associated with this research effort is the simultaneous consideration of demographic factors and question characteristics, in the same model. Traditionally such analyses have proved challenging, because of the inherent multi-level, hierarchical nature of the data (given that question characteristics associated with multiple questions are tested within-subject, but demographic variables such as ethnic group membership exist at a higher, between-subject level). Recently, however, Raudenbush and colleagues in particular have strongly advocated the use of hierarchical linear modeling in such cases – and we have begun expanding our analysis through use of the software program HLM6.

We therefore conducted HLM analyses that enable the determination of effects of a range of coded question characteristics, within the same model that assesses the influence of demographic factors. These analyses revealed that several factors did predict the frequency with which behavior codes are assigned, as determined by our use of a summary Problem measure that treats Problem/No Problem as a dichotomous variable (and that requires logistic modeling of the relationship between predictor and outcome variables; specifically via a logit link function).

Table 2. Factors that significantly increased probability of a respondent Problem code.

Coded characteristic -	**Unadjusted values -**
Abstract rather than concrete	29% vs. 18%
Question contains *interviewer probe*	32% vs. 21%
Reference period: 12 month vs. now	15% vs. 27%
Number of *response categories*	Mean = 4.7 - problem 3.9 - no problem
Flesch-Kincaid Grade Level	Mean = 8.4 - problem 8.1 - no problem

----

Using a model that simultaneously controls for the effects of multiple factors, we found that:

a) Questions that are (subjectively) coded as containing abstract as opposed to concrete concepts produced significantly more respondent behavior codes (the corresponding univariate, or unadjusted values, are 29% versus 18%, respectively).

a) Questions containing an additional interviewer probe – that is, an additional statement to be read as needed -- produced more respondent-based behavior codes. Note that we do not know whether these probes were actually read – only that questions for which the designers believed an additional, optional probe was necessary produced more disruptions in the interaction than those without probes.

b) Interestingly, questions with longer reference periods (such as 12 months) produced significantly fewer codes than did those with shorter reference periods (e.g., now). However, it would be inappropriate at this point to conclude that longer reference periods in some manner ease the interaction between interviewer and respondent. Even though we did control for some potentially confounding factors, questions were not selected in a way that completely balanced the effects of different characteristics – it could simply be that the questions asking about "now" happened to cover a topic that was especially difficult for respondents. In fact, questions about the present were dominated by a series asking the respondent to make judgments about the social characteristics of the neighborhood in which they live, which may present unique difficulties.

c) Number of response categories read to the respondent was positively related to the probability of observing a problem in the interviewer-respondent interaction.

d) Finally, the Flesch-Kincaid Grade Level (a measure of print readability) was also positively related to the frequency of problem codes (note that at this point we have relied on Grade Level in English, even for Spanish and Korean cases, under the unverified assumption that Grade level is highly correlated across languages).

Conclusions

1) We strongly advise survey researchers not to overlook the impact of the interviewer, in analysis of interviewer-administered questionnaires that involve multiple language groups. Given that interviewers must be specifically selected to administer questionnaires in languages such as Korean, it behooves the questionnaire developer to pay careful attention to issues beyond translation, and to focus on interviewer selection, training, and monitoring during the field survey. Even a good quality translation can be rendered ineffective if interviewers do not utilize it in a consistent, standardized manner. Further, it may be inappropriate to assume that interviewers who depart from the translated instrument are doing so in order to produce a "conversational" alternative to an otherwise defective instrument in order to decrease error; they may in fact be making the situation worse.

2) Further, and not surprisingly, given previous methodological work in the area, age appears to be a major factor, within any group, that influences the way in which telephone survey items function. Special efforts can be made to pretest questionnaire on the elderly, as an attempt to alleviate resultant problems.

3) We also propose that acculturation may be a potentially potent factor in explaining cross-cultural differences. We found that percent live in the US was a fairly consistent predictor of problems, which may not be surprising, given that one byproduct of

acculturation may in effect consist of a learning process which renders one more familiar with the peculiar demands associated with responding to survey questionnaires. Parenthetically, a measure often used to represent acculturation – number of years in the US – was not useful, as it is highly correlated with the even more potent factor of age.

4) Finally, the beneficial effect of educational level is not surprising, but it is reassuring to see that this effect occurred, as this lends face validity to the investigation as a whole.

As a caveat – we return to a point made initially: Our performance measure of choice – behavior coding – has not been demonstrated to serve as a direct measure of response error in interviewer-administered surveys, at least to our satisfaction. We have determined clearly that elements of the interaction between interviewer and respondent do differ across demographic groups. What remains is to assess the ramifications of departures from "ideal" interaction, in terms of data quality within the evaluated survey. It may be that problems in the interaction lead directly to problems with data. On the other hand, some degree of disruption of the question asking-answering process may reflect a negotiation of question meaning of the type often discussed by socio-linguists (e.g., Schober & Conrad, 1997), and that have the ultimate effect of reducing error.

References

Agans, R. P., Deeb-Sossa, N., & Kalsbeek, W. D. (2006). Mexican immigrants and the use of cognitive techniques in questionnaire development. *Hispanic Journal of Behavioral Sciences, 28,* 209-230.

Behling, O., & Law, K.S. (2000). *Translating questionnaires and other research instruments: Problems and solutions.* London: Sage.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185-216.

Cannell, C. F., Fowler, F. J., & Marquis, K. (1968). The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interviews. *Vital and Health Statistics*, Series 2, No. 26. Washington, DC: U.S. Government Printing Office.

Census Bureau (U.S.) (2004). *Census Bureau Guideline: Language translation of data collection instruments and supporting materials*. Retrieved August 12, 2007 from http://www.census.gov/cac/www/007585.html.

DeMaio, T. J. and Rothgeb, J. M. (1996). Cognitive Interviewing Techniques: in the Lab and in the Field. In N. Schwarz and S. Sudman (Eds.), *Answering questions* (pp. 177-198). California: Jossey-Bass.

European Social Survey (2002). An outline of ESS translation strategies and procedures. Retrieved April 13, 2006 from http://naticent02.uuhost.uk.uu.net/ess_docs/R3/Methodology/r3_translation_guidelines.pdf.

Forsyth, B. H., Kudela, M. S., Levin, K., Lawrence, D., & Willis, G. B. (2007). Methods for translating an English-language survey questionnaire on tobacco use into Mandarin, Cantonese, Korean, and Vietnamese. *Field Methods, 19,* 264-283.

Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 393–418). New York: Wiley.

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

Harkness, J. A., & Schoua-Glusberg, A. (1998). Questionnaires in translation. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence* (pp. 87-126). Mannheim, Germany: ZUMA .

Harkness, J. A., Van de Vijver, F. J. R., & Mohler, P. (2003). *Cross-cultural survey methods*. Hoboken, N. J.: Wiley.

Johnson T. P. (2006). Methods and frameworks for crosscultural measurement. *Medical Care, 44*, S17–S20.

Johnson, T. P. (1988). Approaches to equivalence in cross-cultural and cross-national survey research. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence* (pp. 1-40). Mannheim, Germany: ZUMA.

Martinez, G., Marín, B. V., & Schoua-Glusberg, A. (2006). Translating from English to Spanish: The 2002 National Survey of Family Growth. *Hispanic Journal of Behavioral Sciences, 28*, 531-545 .

McKay, R. B., Breslow, M. J., Sangster, R. L., Gabbard, S. M., Reynolds, R. W., Nakamoto, J. M., & Tarnai, J. (1996). Translating survey questionnaires: Lessons learned. *New Directions for Evaluation, 70*, 93-105.

Miller, K. (2004). Implications of socio-cultural factors in the question response process. In P. Prufer, M. Rexroth, & F. J. Fowler (Eds.), *Questionnaire evaluation standards* (pp. 172-188). Mannheim, Germany: ZUMA.

Nápoles-Springer, A. M., Santoyo-Olsson, J., O'Brien, H, & Stewart, A. L. (2006). Using cognitive interviews to develop surveys in diverse populations. *Medical Care, 44* (Suppl 3), S21–S30.

Nápoles-Springer, A. M., & Stewart, A. L. (2006). Overview of qualitative methods in research with diverse populations: Making research reflect the population. *Medical Care, 44 (Suppl 3*), S5–S9.

Pan, Y., & de la Puente, M. (2005). *Census Bureau guidelines for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed.* (Research Report Series #2005-06, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.).

Schmidt, S., & Bullinger, M. (2003). Current issues in cross-cultural quality of life instrument development. *Archives of Phys Med Rehabil, 84,* S29-34.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61,* 576–602.

Stewart, A. L., & Nápoles-Springer, A. (2000). Health-related quality-of-life assessments in diverse population groups in the United States. *Medical Care, 38 (Supplement II),* 102-124.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge: Cambridge University Press.

Warnecke, R. B., Johnson, T. P., Chavez, N., Sudman, S., O'Rourke, D. P., Lacey, L., & Horm, J. (1997). Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology, 7,* 334-342.

Willis, G. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Zahnd, E., Tam, T., Lordi, N., Willis, G., Edwards, S., Fry, S., & Grant, D. (2005, July). *Cross-cultural behavior coding: Using the 2003 California Health Interview Survey (CHIS) to assess cultural/language data quality*. Paper presented at the meeting of the European Association for Survey Research, Barcelona.