Behavior Coding across Multiple Languages: The 2003 California Health Interview Survey as a Case Study

W. Sherman Edwards and Stephanie Fry, Westat; Elaine Zahnd and Nicole Lordi, Public Health Institute; Gordon Willis, National Cancer Institute; David Grant, UCLA Center for Health Policy Research

Coding of interviewer and respondent behavior within the fielded interview has become an important tool for evaluating survey questions. Unlike most other evaluation tools, it provides quantitative information on the performance of individual survey questions. The behavior coding system that has evolved into the one most commonly used for this purpose was originally developed at the University of Michigan to assess the effects of interviewer and respondent behavior on the quality of survey responses (Cannell et al, 1966). Some 25 years later, Oksenberg et al (1991) reported on "New Strategies for Pretesting Survey Questions," describing essentially the same coding system applied to a somewhat different purpose. This article and other research (e.g., Zukerberg et al, 1995; Edwards et al, 2002) have focused on issues such as the system's reliability, whether it works equally well with live monitoring and recorded interviews, the sample size needed for pretesting, and the number of codes needed to identify questions that present problems for interviewers and respondents.

This paper will describe an extension of behavior coding to assess questionnaire functioning across language, as part of an evaluation of the degree to which translated versions exhibit cross-cultural equivalence. For this study, behavior coding was performed for several portions of the questionnaire administered to adults (via telephone) in the 2003 California Health Interview Survey (CHIS 2003). Funded by the National Cancer Institute, the study had three principal objectives:

- To evaluate how well particular questions work across cultures and across languages;
- To provide general insights into effects of culture and language on how survey questionnaires function; and
- To assess behavioral coding as a tool for cross-cultural methods research.

This paper will present some overall findings related to the second and third objectives.

Background: The 2003 California Health Interview Survey

Westat conducted CHIS 2003 under contract to the UCLA Center for Health Policy Research; CHIS is a joint undertaking of UCLA, the California Department of Human Services, and the Public Health Institute. CHIS 2003 was a telephone survey using primarily a random-digit-dial sample stratified so as to provide estimates for California's larger counties and groups of smaller counties. Interviews were conducted with about 42,000 sampled adults, with more than 4,000 adolescents in sampled households, and with knowledgeable adults about more than 8,500 children in sampled households. Because of California's diverse population, the RDD sample supports statewide estimates for Latinos, African-Americans, and some Asian groups. The sample was enhanced through geographic stratification and the addition of telephone numbers from surname lists to allow estimates for persons of Korean and Vietnamese ethnicity. The CHIS adult questionnaire includes questions on general health, health conditions and behaviors, women's health, health insurance coverage, and use of health care services, among other topics. The CHIS 2003 screener response rate was 56 percent, using the survival method to allocate noncontacted numbers to household/nonhousehold status, and the adult interview response rate was 60 percent.

CHIS 2003 was conducted in six languages: English, Spanish, Mandarin, Cantonese, Korean, and Vietnamese. The questionnaires were first developed in English, a process that included reviews to ensure that the language was appropriate for persons with a sixth grade reading level, and to identify and revise words, phrases, and concepts that might be unfamiliar or sensitive in different cultures. The English questionnaires were then translated into the other languages and the translations reviewed by independent translators. The translators discussed and resolved differences of opinion with the assistance of CHIS research staff. The adjudicated translations were then reviewed by bilingual staff at Westat who prepared specifications for the CATI program, and by bilingual members of the interviewing team. Again, differences of opinion were adjudicated with the research staff and translators. Bilingual interviewing staff were trained first in English, then in their other language. Some bilingual interviewers worked first in English, then moved to their other language. All interviewers were monitored and their performance reviewed during initial phases of the field work. The research staff held interviewer debriefings early in the period during which each language was fielded; some final questionnaire changes were made during the field period.

Methods

The behavior coding research team decided to sample interviews conducted with adults reporting Hispanic or Latino ethnicity, as well as those reporting themselves to be of Korean background. To examine the effects of both culture and language, the sample was divided into five strata: Latinos interviewed in English and Spanish, Koreans interviewed in English and Korean, and all others interviewed in English. The goal was to record and code portions of about 100 adult interviews in each stratum. The research team selected sections of the adult interview that were of substantive interest (e.g., cancer screening items), that were felt likely to function differently across language or cultures (e.g., sexual orientation), or that had been identified as problematic in a prior behavioral coding exercise for CHIS 2001. Twelve

sections were selected, for a total of 125 question items, 42 of which were asked of all respondents.

Sampling occurred after the initial section of the CHIS 2003 adult interview, which included questions on race and ethnicity. Random samples of Latino/English, Latino/Spanish, and Other/English interviews, and all Korean/English and Korean/Korean interviews, were selected for recording, until the quotas of 100 per stratum were achieved. The CATI program implemented the sample and displayed a consent script for those selected. If the respondent agreed to be recorded, the designated questionnaire sections were digitally recorded to .wav files. The recording process did not involve any further interviewer action, unless the respondent asked for the recording to be stopped, which the interviewer could effect through a toggle key.

The behavior coding scheme is presented in Table 1. The codes are fairly consistent with those used in other studies, with two notable differences. First, because the CHIS 2003 questionnaire included a number of scripted "if needed" probes, and answer categories that were not read as part of the question but could be offered as probes, we distinguished between scripted and unscripted probes among interviewer behaviors. Second, because we were interested in the *gestalt* of the interviews across cultures and languages, we added codes for interviewer and respondent "extraneous comments," or conversation not related to the question-answer interview pattern[1]. Overall, the codes were intended to apply to two levels of exchange – typically a question (interviewer), some response (respondent), a prompt of some kind (interviewer), and another response (respondent). Further behaviors were not coded, except that extraneous comments were coded regardless of what other codes were used.

We trained 6 coders, 3 bilingual in English and Spanish, 2 bilingual in English and Korean, and one English-only. Training in English was conducted over two days, and all coders received initial assignments in English. Monolingual supervisors reviewed coders' work and provided feedback. When most of the English-language cases had been coded, the bilingual coders moved on to Korean and Spanish interviews. Samples of cases in all strata were independently re-coded.

All coding activities were done on the coders' PC desktop. A database program managed the sample, linked to a .wav player, and provided entry screens for the codes. A separate application in CATI presented the screens as seen by the interviewers, and noted the response entered by the interviewer.

We debriefed the coders once during the time they were actively coding, and again when they were done. The first debriefing focused on issues with the coding scheme and general impressions of comparability across cultures and languages, and on coders' perceptions of problem questions. The second coding was driven by the overall and question-specific results of the coding, and focused on differences in the results by language and culture, and on questions that appeared to be particularly problematic generally or with one or more strata.

Results

More than 100 interviews were recorded in each stratum. Cooperation rates ranged between 88 and 91 percent for the English-language strata, 80 percent for Latino/Spanish, and 59 percent for Korean/Korean. The final sample available for analysis included 97 Latino/English interviews, 106 Latino/Spanish, 103 Korean/English, 85 Korean/Korean, and 103 Other/English. Some interviews were not used for analysis because of technical problems or because the respondent changed his/her mind about being recorded after initially giving consent.

Table 2 presents aggregate totals for each of the behavior codes across all questions, by stratum. For the interviewer behavior "asking questions," there is little difference across strata except for the Korean Korean. While 86 to 89 percent of questions were read as written in English and Spanish, and only 2-3 percent were coded as "major change," about a third of the Korean Korean questions were in each of the change categories, and there were three times as many "verification" codes as for the other strata.

The respondent behavior codes combine first and second level responses. Not surprisingly given their high rates of misreading, interviews in Korean had the lowest rate of "adequate answer" and highest rate of "inadequate answer." Korean-speaking respondents provided adequate answers by the second exchange for less than two-thirds of questions coded, compared with 86-90 percent for other languages, with Spanish-speaking respondents having the highest rate. Twenty-eight percent of items coded for Korean-speaking respondents included an inadequate answer; Spanish-speaking respondents had the next highest rate, at 15 percent. Korean-language interviews also had the highest rate of requesting clarification (for 9 percent of questions coded), Korean respondents interviewed in English also had a higher rate of requesting clarification than other groups, at 6 percent. Spanish-speaking respondents had the lowest rate of interrupting the reading of the question with an answer (2 percent), while Korean-language interviews and "other" English interviews had the highest rates (5 percent). Qualified answers were relatively rare for all groups, but particularly for Spanish-speaking and Korean-speaking respondents. Overall, the Korean-language interviews had about twice as many questions with one or more respondent problem behaviors (any code other than "adequate answer") as any other group.

Korean-language interviews had the highest rate of interviewer follow-up behavior, with more than one-quarter of question exchanges resulting in follow-up. Unscripted probes

[1] Similar codes were part of the original Michigan coding scheme (Cannell et al, 1966), but were not retained over time.

accounted for about half of these behaviors, and offering clarification another quarter. Scripted probes were used most often in Spanish-language interviews, about three times as often as for other interviews.[2]

Relatively few of the question item exchanges included extraneous comments, with respondents (at 4 percent of question items coded) about twice as likely as interviewers (2 percent) to make extraneous remarks. "Other" English-language respondents made the most extraneous comments (6 percent), while Koreans (2 percent) and Latinos (3 percent) interviewed in English made the fewest.

Table 3 describes the level of inter-coder reliability for interviews coded in each stratum. All of the coders, as well as two supervisors, are included in the English-language strata, while only the 2 Korean and 3 Spanish bilingual coders are included in those languages. Reliability is measured using Kappa, a statistic that uses expected marginal totals to mitigate the effects of apparent high agreement for low-frequency behaviors. Kappa is a proportion ranging from 1.0 (perfect agreement) to -1.0 (perfect disagreement). The two interviewer codes each have multiple values, while the respondent and extraneous comment codes are discrete "0/1" variables. Kappas of about 0.8 are desirable for behavior coding applications.

Most Kappas for the interviewer behaviors are between 0.4 and 0.7, indicating fair to good agreement. The Korean bilingual coders had the highest rates of agreement, while the Spanish bilingual coders had the lowest, particularly for asking questions. The respondent codes show higher agreement in general than the interviewer codes, with the notable exception of qualified answer. This behavior was particularly rare in the non-English interviews, and the coders did not agree on the few instances when either noted it. Again, the Korean bilingual coders had generally the highest rates of agreement and the Spanish bilingual coders the lowest, although this pattern did not hold for all respondent codes. Kappas for the extraneous comment codes were lower than for other behaviors, and lower for interviewer comments than for respondent comments (except for the Korean-language cases).

Discussion

The coding results and the inter-coder reliability analysis reveal some differences across language of interview and respondent ethnicity. Since we did not go into this study with a priori hypotheses about differences, our interpretations are post hoc. Some are informed by comments made by the coders during debriefing, and others by listening to English-language interviews.

The most striking difference by stratum is the high rates of interviewer mis-reading in the Korean-language interviews. There are several apparent causes for this result. The Korean translation was described by the coders as being inconsistent, particularly with regard to the level of politeness indicated by particular word forms. This discrepancy was due in part to carrying over translations from CHIS 2001 for some items, and having other items newly translated. Even if the translation had been consistent, the appropriate level of politeness differs by the age and gender of interviewer and respondent, so "mis-reading" questions is culturally appropriate in many cases. There were relatively few Korean-language interviewers, and coders noted that the interviewers differed considerably in their adherence to the script. The level of respondent problems in Korean-language interviews is certainly attributable, at least in part, to the mis-readings of the interviewers. Coders also noted that some specific questions included terms most respondents were unlikely to know, adding to the likelihood of respondent difficulties.

Other differences are more subtle. The Spanish-language interviews had the highest rate of scripted probes, the lowest rates of respondents interrupting with answers and providing qualified answers, and the highest rate of providing adequate answers despite having relatively high rates of inadequate answers and "don't knows." These differences perhaps point to a more passive role by Spanish-language respondents, and also may indicate that the scripted probes were particularly helpful in the Spanish-language interviews. The Spanish bilingual coders, all of whom had prior survey experience, said that the translation was particularly good.

One reasonable post hoc hypothesis is that respondents behave somewhat differently when they are being interviewed in their first language, i.e., they may feel more comfortable and hence be more active. One finding in support of this hypothesis is that respondent extraneous comments were lowest in the English-language interviews of Korean and Latino respondents. Respondent interruptions were more likely in the Korean/Korean and "Other"/English interviews than in the other groups (although lowest overall, as noted, in the Spanish-language interviews).

Question-specific results (not shown) included situations where particular questions appeared problematic for all groups, for example:

> "Which statement best describes the rules about smoking INSIDE your home? Smoking is NEVER allowed inside, allowed in SOME places or at SOME times, or allowed ANYWHERE and ANYTIME inside your home?"

Other questions proved problematic because of specific translation or cultural issues. For example, the item asking about sexual orientation was difficult for Latinos because of the word "bisexual," which was sometimes misinterpreted as a synonym for "heterosexual." The same question was problematic for some Korean respondents, particularly older ones, because it is not appropriate to talk about the subject.

[2] The apparent finding that Spanish-language interviews had a notably low rate of unscripted probes is actually due to coder differences in assigning this code.

The differences in agreement among coders appear to be due more to the individuals involved than any language or cultural differences. Two of the Spanish-language coders were replacements for others who did not complete the initial training, and the third did not code as many interviews as others. The two Korean coders completed the most cases, and also worked as interviewers during the same period.[3]

Conclusions and Further Research

This early look at the results of the CHIS 2003 cross-cultural behavior coding suggests that behavior coding is a useful tool for evaluating questionnaires in multiple languages. The coding results identify differences by culture and by language in interviewer and respondent behavior, both at the aggregate and question-specific levels. Inter-coder reliability was relatively good across languages, although somewhat problematic for the Spanish-language interviews. Recruiting and training bilingual coders takes more time and effort than a comparable English-only project, but the extra effort is comparable to or less than that required to field the survey in multiple languages to begin with.

Interpretation of differences in coding results across cultures and languages is complicated by the multiple possible causes of those differences. General and question-specific translation issues are one significant source, as are differences in cultural norms and familiarity with terms and concepts used in survey questions. Differences in training and supervision of interviewers in multiple languages may also contribute to apparent questionnaire problems. Finally, differences in coder performance may confound interpretation of discrepant results.

Future research planned for this study includes:

- Question-specific analysis of coding and debriefing data by and across ethnicity and language;
- Association of question features with problem types;
- Relationship between interviewer and respondent behavior;
- Association between respondent characteristics and interviewer/respondent problems; and
- Association between extraneous comments and data quality.

It would of course be worthwhile to see other studies pursue this topic, using other survey questionnaires and including other languages and ethnic groups.

References

Cannell, C., Fowler, F., and Marquis, K. "The Influence of Interviewer and Respondent Psychological and Behavior Variables on the Reporting in Household Interviews." Vital and Health Statistics: Data Evaluation and Methods Research, Series 2, Number 26. National Center for Health Statistics, 1968.

Edwards, W., Narayanan, V., Fry, S., Catania, J., and Pollack, L. "A Comparison of Two Behavior Coding Systems for Pretesting Questionnaires," Proceedings of the Survey Research Methods Section, American Statistical Association, 2002.

Oksenberg, L., Cannell, C., and Kalton, G. "New Strategies of Pretesting Survey Questions," Journal of Official Statistics, 1991, Vol. 7, No. 3, pp. 349-366.

Zukerberg, A., VonThurn, D, and Moore, J. "Practical Considerations in Sample Size Selection for Behavior Coding Pretests," Proceedings of the Survey Research Methods Section, American Statistical Association, 1995, pp. 1116-1121.

[3] None of their interviews was recorded for behavior coding.

Table 1. Explanation of CHIS 2003 Behavior Coding Categories

Interviewer Behavior – Reading Question

No change	Interviewer read the question exactly as printed.
Minor change	Interviewer slightly changed question but meaning was not affected.
Major change	Interviewer altered the meaning of the question or response task. Includes incomplete reading of question because R interrupted with answer
Verification	Interviewer verified answer to a question from information respondent had previously given.

Interviewer Behavior – Follow-up

Re-read question	Interviewer re-read all or part of question, exactly as written, including re-reading of one or more answer categories in lower case.
Clarification	Interviewer made a statement about question intended to help respondent answer.
Scripted probe	Interviewer asked a follow-up question exactly as shown on the screen, or read one or more answer categories in upper case.
Unscripted probe	Interviewer asked a follow-up question that was neither a scripted probe nor part of the original question.

Respondent Behavior

Adequate answer	Respondent gave answer that met question objective and could be coded using a response category provided.
Interrupts	R interrupted reading of question with answer.
Requests repeat	R asked interviewer to repeat the question or response category.
Requests clarification	R indicated uncertainty of meaning or asked interviewer to define a term or provide additional information in order to understand the question or response categories.
Qualified answer	R gave answer that met question objective but was qualified to indicate uncertainty
Inadequate answer	R gave answer that did not meet question objective and which did not provide an answer which could be coded using response categories provided.
Don't know	R did not know the answer to the question being asked.
Refused	R declined to provide an answer to the question being asked.

Extraneous Comments

By Interviewer	Interviewer initiated a statement or asked a question that was unnecessary for administering the questionnaire or coding the response. Does not include direct responses to extraneous comments initiated by the respondent if no further extraneous comments were added by the interviewer.
By Respondent	R initiated a statement or asked a question that was unnecessary for responding to the study question being asked. Does not include direct responses to extraneous comments initiated by the interviewer if no further extraneous comments were added by the R.

Table 2. Summary of Behavior Codes Assigned by Stratum

Stratum	English	Hispanic English	Hispanic Spanish	Korean English	Korean Korean	All
Number of items coded	7231	6359	6420	6283	5557	31850
Interviewer Behaviors -- Asking Questions						
No Change	6207	5543	5654	5560	1846	24810
	85.8%	87.2%	88.1%	88.5%	33.2%	77.9%
Minor Change	729	611	528	564	1770	4202
	10.1%	9.6%	8.2%	9.0%	31.9%	13.2%
Major Change	231	171	216	125	1795	2538
	3.2%	2.7%	3.4%	2.0%	32.3%	8.0%
Verification	64	34	22	34	146	300
	0.9%	0.5%	0.3%	0.5%	2.6%	0.9%
Interviewer Behaviors -- Follow-up						
Reread Question	239	248	213	300	190	1190
	3.3%	3.9%	3.3%	4.8%	3.4%	3.7%
Clarification	197	185	184	254	388	1208
	2.7%	2.9%	2.9%	4.0%	7.0%	3.8%
Scripted Probe	247	200	608	150	128	1333
	3.4%	3.1%	9.5%	2.4%	2.3%	4.2%
Unscripted Probe	531	411	176	400	749	2267
	7.3%	6.5%	2.7%	6.4%	13.5%	7.1%
All Interviewer Follow-ups	1214	1044	1181	1104	1455	5998
	16.8%	16.4%	18.4%	17.6%	26.2%	18.8%
Respondent Behaviors						
Adequate Answer	6197	5669	5773	5618	3637	26894
	85.7%	89.1%	89.9%	89.4%	65.4%	84.4%
Interrupts with Answer	336	173	118	186	267	1080
	4.6%	2.7%	1.8%	3.0%	4.8%	3.4%
Requests Repeat	84	88	68	91	99	430
	1.2%	1.4%	1.1%	1.4%	1.8%	1.4%
Requests Clarification	270	223	231	349	482	1555
	3.7%	3.5%	3.6%	5.6%	8.7%	4.9%
Qualified Answer	169	120	43	136	76	544
	2.3%	1.9%	0.7%	2.2%	1.4%	1.7%
Inadequate Answer	925	752	1003	647	1559	4886
	12.8%	11.8%	15.6%	10.3%	28.1%	15.3%
Don't Know	131	62	165	99	159	616
	1.8%	1.0%	2.6%	1.6%	2.9%	1.9%
Refused	14	4	8	18	4	48
	0.2%	0.1%	0.1%	0.3%	0.1%	0.2%
Any Respondent Problem	1731	1306	1531	1362	2329	8259
	23.9%	20.5%	23.8%	21.7%	41.9%	25.9%
Extraneous Comments						
By Interviewer	186	120	103	85	143	637
	2.6%	1.9%	1.6%	1.4%	2.6%	2.0%
By Respondent	400	173	234	118	208	1133
	5.5%	2.7%	3.6%	1.9%	3.7%	3.6%

Table 3. Inter-Coder Reliability (Kappa)

Ethnicity *Language*	Other English	Latino English	Latino Spanish	Korean English	Korean Korean
Interviewer Behaviors					
Asking question					
(Unweighted Kappa)	0.42	0.47	0.17	0.45	0.65
(Weighted Kappa)	0.46	0.55	0.18	0.50	0.73
Follow-up					
(Unweighted Kappa)	0.58	0.59	0.52	0.59	0.73
(Weighted Kappa)	0.67	0.64	0.61	0.65	0.74
Extraneous comments	0.22	0.15	0.22	0.27	0.53
Respondent Behaviors					
Adequate answer	0.58	0.49	0.35	0.55	0.78
Interrupts with answer	0.79	0.78	0.47	0.83	0.58
Requests repeat	0.84	0.66	0.33	0.67	0.91
Requests clarification	0.77	0.84	0.49	0.76	0.86
Qualified answer	0.22	0.23	-0.01	0.37	0.00
Inadequate answer	0.55	0.52	0.59	0.58	0.79
Don't know	0.80	0.70	0.66	0.81	0.80
Refused	1.00	N/A	1.00	0.96	1.00
Extraneous comments	0.390	0.276	0.383	0.418	0.322