RESEARCH REPORT SERIES
(Survey Methodology #2007-18)

The Impact of Instructions on Survey Translation: An Experimental Study

Yuling Pan[1]
Brian Kleiner[2]
Jerelyn Bouic[2]

Statistical Research Division[1]
U.S. Census Bureau
Washington, DC 20233

Westat[2]
1650 Research Boulevard
Rockville, Maryland 20850

Report Issued: June 15, 2007

The Impact of Instructions on Survey Translation: An Experimental Study

Final Report

Authors:

Brian Kleiner
Jerelyn Bouic
Westat

Yuling Pan
U.S. Census Bureau

January 2007

Prepared for:

U.S. Census Bureau

Prepared by:

WESTAT
1650 Research Boulevard
Rockville, Maryland 20850
(301) 251-1500

TABLE OF CONTENTS

	Page
Introduction	1
Background and Purpose of the Study	1
Study Approach	4
Source Instrument	4
Recruitment of Translators	5
Instructions Provided to Translators	5
Evaluation of Translations by Survey Researchers	7
Interviews with Evaluators	10
Analysis of Data	10
Study Results	11
Quantitative Findings	11
Qualitative Findings	15
Evaluator Background	15
Level of Difficulty of the Task	15
Assignment of Ratings	16
The Role of QxQs	18
Discussion and Implications	19
Future Research	21
References	23
Appendix A: Source Instrument	A-1
Appendix B: QxQs	B-1

LIST OF TABLES

Table		Page
1	Distribution of 27 translators into nine subgroups	4

LIST OF EXHIBITS

Exhibit		
1	Translator recruitment criteria	5
2	Core instructions provided to translators	6
3	Instructions received by the three instructional subgroups	7
4	Instructions given to evaluators	8
5	Example page from Spanish evaluation form	9
6	Protocol for interviews with evaluators	10

LIST OF FIGURES

Figure		
1	Mean ratings for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup	11
2	Mean ratings for French translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup	12
3	Mean ratings for Spanish translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup	13
4	Mean ratings for Chinese translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup	14

If a translation is to meet the four basic requirements of (1) making sense, (2) conveying the spirit and manner of the original, (3) having a natural and easy form of expression, and (4) producing a similar response, it is obvious that at certain points the conflict between content and form (or meaning and manner) will be acute, and that one or the other must give way.

—Eugene Nida, 1964

Introduction

Minimally directed translations of survey instruments, even by experienced translators, risk being of poor quality, whether it is because translators are not informed about the intended meaning of survey items (and therefore have to bring their own interpretations of meaning), they do not take into consideration cultural and communicative norms, or they are not given guidance about an acceptable degree of adaptation. For any of these reasons, translations may misrepresent the intent of survey items or may sound unnatural or even offensive to survey respondents. Nonetheless, minimally directed survey translation work is still the norm, rather than the exception.

The experimental study described in this report aimed to assess the impact of different types of instructions given to translators for their translation of survey instruments from English into three target languages (Mandarin Chinese, Spanish, and Canadian French). The instructions were meant to guide translators toward a better understanding of what the survey items were intended to mean as well as how to make items sound more natural and culturally appropriate. Study findings indicate that while the provision of such instructions to translators had a significant impact on their translations, the direction of the impact (positive or negative) differed across the target languages, according to ratings of professional survey researchers who were also native speakers of those languages. The differences in ratings are attributed to the beliefs of the survey researchers and their level of commitment to two conflicting general types of equivalence in survey translation: equivalence of stimulus and equivalence of effect.

Background and Purpose of Study

Eugene Nida's distinction between "formal" and "dynamic" equivalence in literary translation and the assertion that the two are inevitably in conflict bear strong relevance to current issues in survey translation theory. Within Nida's framework, formal equivalence focuses on correspondence of both form and content between a source and a target language. Such correspondences may be grammatical, lexical, and/or semantic, such that "the message in the receptor language should match as closely as possible the different elements in the source language" (p161). In contrast, dynamic equivalence focuses less on formal and semantic correspondences and more on establishing an equivalent *effect* on the receiver of the message, such that "the relationship between receptor and message should be substantially the same as that which existed between the original receptors and the message" p156. Rather than attempting to impose the cultural patterns and formal properties of the source-language context, a dynamic translation aims for naturalness of expression and recreates the message with respect to modes of behavior that are meaningful within the context of the receptor's own culture.

While the translation of surveys is a far cry from translation of literature or poetry, Nida's formal/dynamic equivalence distinction provides important background to a current debate within the field of comparative survey research. A guiding principle of survey research is that of standardization, which dictates that survey respondents receive exactly the same stimulus in the same manner: When respondents are all asked the same questions in exactly the same way, this reduces the chance of bias and

measurement error. For the same reason, survey researchers who conduct surveys that cross linguistic and cultural borders generally place a high premium on formal equivalence (in Nida's terms) and aim for standardization and equivalence of stimulus within the translations of source instruments.

Unfortunately, in practice the aim of equivalence of stimulus and of one-to-one formal and semantic correspondence quickly runs into difficulties in the translation of survey items. Van Ommeren et al. (1999) provide several examples that illustrate the need for adaptation from an instrument developed for use with Nepali-speaking Bhutanese refugees in living in Nepal.[1] In one example, an English verb (*exert*) has no corresponding word in Nepali and had to be translated as a larger phrase:

> **Source:** "Have you ever had shortness of breath when you had not been exerting yourself?"
>
> **Translation (into village Nepali):** "Have your ever had shortness of breath when you had not been working, running, or climbing up a steep path?"

In another example, rather than adding text not directly corresponding to the source, text was *omitted* in the translation, because it was not relevant to the ordinary experiences of target respondents:

> **Source:** "Did drinking ever cause you to give up or greatly reduce important activities – like participating in sports, going to school or work, or keeping up with friends or relatives?"
>
> **Translation:** "Did drinking ever cause you to give up or greatly reduce important activities – like going to school or work, or keeping up with friends or relatives?"

In this case, participation in sports is not an important value for Nepali villagers and so this phrase was removed from the question. Here the adaptation was not strictly necessary for purposes of comprehension, but did assist in making the translation more relevant and meaningful within the cultural context of respondents.

Sometimes cultural and linguistic norms require adaptation in a translation and departure from formal equivalence. In a third example, the translation of a sensitive question adds a great deal of text not found in the source item and modifies the question itself to ensure acceptability and cooperation among potential respondents (e.g., women for whom public discussion about sexual activities and interests is strictly taboo):

> **Source:** "During one of those periods [of feeling depressed] was your interest in sex a lot less than usual?"
>
> **Translation:** "Now I am going to ask you a private question. Please do not feel bad about answering as it will remain confidential. During one of those periods [of feeling depressed] was your desire of sleeping with your spouse a lot less than usual?"

These examples show that maintaining equivalence of stimulus in translations of survey items can be a vexing challenge. The adaptation required to promote comprehension, relevance, and cultural appropriateness is often at odds with the researcher's goal of standardization and formal and semantic correspondence between a source and a target survey item. On the other hand, strict adherence to formal

[1] The authors, following Manson (1997), focus on four main ways in which translations must depart from a formal equivalence approach or risk failing to be meaningful to respondents: 1) comprehensibility, 2) acceptability, 3) relevance, and 4) completeness.

equivalence may lead to incomprehensibility, irrelevance, offensiveness, or awkwardness that may critically undermine the translation.

It is in response to these sorts of challenges that some researchers and translation theorists have come to regard formal equivalence as an impractical and undesirable aim, opting instead for an approach based on dynamic equivalence and a greater tolerance for adaptation. Harkness and Schoua-Glusberg (1998) argue that a translation should adequately maintain the measurement properties of the source item provided that it faithfully conveys the *intended meaning* of the source. Such a pragmatics-based approach to translation situates the creation of meaning firmly within the communicative process and the linguistic and culture-specific norms and maxims that guide the conveying and interpretation of intent (Gutt 1991). Harkness (2003) points out that the "Achilles heel" of survey translation is that translators do not ordinarily take into consideration "pragmatic equivalence," which involves ways of employing linguistic forms in particular contexts to communicate meaning and intent. Kleiner and Pan (2006) demonstrate that cross-cultural survey researchers and survey translators not only do not sufficiently consider sentence-level pragmatic equivalence, but further do not ordinarily take into consideration cross-cultural differences in discourse-level norms of language use.

In this light, translation of survey items should primarily involve the communication of an intended interpretation by way of exploitation of the appropriate linguistic and cultural norms of the target community of respondents. This adherence to dynamic equivalence in translation thus requires a shift from equivalence of *stimulus* to equivalence of *effect*. Nida recognized a continuum of standards and practice between the extremes of full devotion to formal or dynamic equivalence in literary translation. In a similar way, survey researchers who work with translations of source instruments fall on different points of a stimulus/effect continuum. One finds among survey researchers, therefore, differing levels of fidelity to strict formal equivalence on the one hand and tolerance to adaptation on the other.

In calling for faithfulness to intended meaning as a guiding principle of survey translation, Harkness and Schoua-Glusberg (1998) also argue for the need to provide translators with documentation so that they can better understand the intended readings and research aims of survey items: "…given that meaning is not fixed and finite, one of the goals of translation must be to convey the intended and most salient reading of a well-written question. The intended meaning of an item should therefore be documented for translators in the source materials they receive for their task." (p95) The authors also point out that translators should be given detailed guidelines and examples regarding an acceptable degree of freedom in adapting a target item.

In practice, however, translators are rarely provided with such documentation and guidelines and so are normally left on their own to divine the intended interpretation of survey items and the extent to which they can adapt them. While providing translators with materials that clarify the intended meaning of survey items seems reasonable on the surface (at least to adherents of this approach to survey translation), there is currently little empirical work that lends support to the utility of this practice. The exploratory study described in this report examines the impact of providing such documentation and detailed guidelines to translators. Specifically, the experimental study that was conducted addressed whether translators given explanatory material and instructions are able to produce translations that are more faithful to the intended meaning of source survey items and more culturally appropriate and natural sounding than translators who receive no such material. The next sections describe the study design. These are followed by sections on study findings and discussion.

Study Approach

The study involved an experimental design to examine the extent to which particular instructions provided to translators had a significant effect on the translation of survey items. In order to assess the impact of different types of instructions, 27 professional translators translated an English source instrument into one of three target languages—Mandarin Chinese, Spanish, and Canadian French, following one of three sets of instructions (see below for details about the instructions). Table 1 shows the distribution of the 27 translators into nine different subgroups. Translators who fulfilled our selection criteria (see below) were randomly assigned to one of three sets of instructional subgroups.

Table 1.—Distribution of 27 translators into nine subgroups

	Chinese	Spanish	French
Instruction set 1 (Group A)	3	3	3
Instruction set 2 (Group B)	3	3	3
Instruction set 3 (Group C)	3	3	3

Once the translations were completed, 15 professional survey researchers who were native speakers of the target languages conducted blind evaluations, with each evaluator examining three translated versions of each survey item, one version for each set of instructions. The evaluation involved rating each translated item on Likert scales along several dimensions (see below). Thus each survey item received three ratings along each dimension from each evaluator.

Analyses involved comparison of ratings along the three dimensions overall and for individual items, following the three sets of instructions. It was presumed that, if significant differences were found in the ratings, this would suggest that certain instructions provided to translators may result in higher quality translations, which in practice would require less followup quality control and revision. The remainder of this section spells out the details of the study design, including discussion of the source instrument, instructions to translators, how translators and evaluators were recruited, how evaluators rated the various translations, and how the ratings were analyzed.

Source Instrument

With the permission of the National Center for Health Statistics (NCHS) of the Centers for Disease Control (CDC), we adopted items from the National Survey of Children with Special Health Care Needs for our source instrument. This "donor" survey was selected because it satisfied several criteria—this was a household telephone survey with sensitive and varied types of questions. The specific items that were adopted were selected based on the need for questions with various structures (e.g., yes/no questions, scales, multiple response categories, questions with prefaces, questions with topic shift indicators, discourse markers) and varying content (e.g., factual questions, opinion questions, sensitive questions about the health conditions of children, demographic questions).

The 18 adopted items were maintained without changes to wording and were placed in a logical order.[2] The National Survey of Children with Special Health Care Needs survey has been thoroughly

[2] Sixteen of the items were actual survey questions, while the first two were drawn from introductory text where the purpose of the survey and telephone call are explained to respondents.

tested and administered several times, and so we felt confident that the source items were of sufficiently high quality. Item 16 of the source instrument (on race/ethnicity) used for the current study was borrowed from the Census Bureau's most recent short form. The source instrument is provided in Appendix A.

Recruitment of Translators

The 27 translators were recruited by Westat following the recruitment criteria shown in Exhibit 1. Most importantly, all participating translators were required to be native speakers of the target languages with at least 5 years of full-time (or equivalent) professional experience in translation. In addition, translators had to have been a resident of the United States or Canada for at least the last 5 years and had to possess at least a Bachelor's degree. To avoid potential bias, translators were not told about the nature or purpose of the current study.

Exhibit 1.—Translator recruitment criteria

1. Native speaker of target language, and ability to translate using standard dialect of the target language, where standard dialect is defined as:
    a. Standard Mandarin Chinese (simplified characters).
    b. Standard Canadian or Quebec French.
    c. Standard Latin American Spanish (any country).
2. Has lived, worked, and received college education in the native country.
3. Resident of the United States or Canada for the last five years, at a minimum (to ensure knowledge of North American society and fluency in English).
4. Five years full-time (or equivalent) experience in translation.
5. Possession of a minimum of a Bachelors degree in their chosen field.
6. Individual (not an agency) to facilitate contacts and interviewing.

Instructions Provided to Translators

All 27 translators were given a core set of instructions to guide their translations (see Exhibit 2). The instructions included a description of the objectives of the survey and the characteristics of the respondents. The core instructions also highlighted important features of the survey interview to be taken into consideration when translating, including that the interview was intended for oral administration.

Translators within the second and third instructional subgroup (hereafter referred to as Groups B and C) were also provided with question-by-question explanations (QxQs) that clarified the intended meaning of each source survey item (see Appendix B). The QxQs followed each source survey item. In addition to the core instructions, translators in Groups B and C were instructed to translate items in a way that was faithful to the intended meaning of the source items, as reflected in the QxQs:

IMPORTANT: We ask that you translate the survey items with respect to their **intended meaning**. Before attempting to translate each item, read carefully the explanation that follows the item in order to get a better sense of what is intended. (The explanations are all in italics – do not translate these.)

It should be noted that the QxQs were developed by study staff, but were reviewed for accuracy by the original survey designer and project director of the National Survey of Children with Special Health Care Needs at the National Center for Health Statistics.

Exhibit 2.—Core instructions provided to translators

INSTRUCTIONS FOR TRANSLATION AND INFORMATION ABOUT THE SURVEY

We ask that you translate the survey with the following information and guidelines in mind:

1) Objectives of the study

This brief household survey will be conducted by telephone and will be used to collect data from parents on aspects of the health care of their young children. The survey targets households with a child 8 years old or younger.

2) Characteristics of the respondents

For the purposes of your translation, please assume that the typical survey respondent will be…

a. A parent of a young child (either mother or father),
b. A native speaker of the target language, between the ages of 18 and 60,
c. Someone now living in the United States, and
d. Someone who can understand the standard dialect that we are asking you to translate into.

3) Education level of respondents

Not all of the respondents will be highly educated. Please try to translate so that the translation can be understood by most people, even if they have not been formally educated.

4) Translation of telephone survey

Since this is a telephone interview, please use the standard spoken form of the language. Avoid using wording or syntax of formal written language.

5) Features of the survey instrument

Please translate everything in the attached source survey. However, do NOT translate text that is CAPITALIZED (i.e., all in CAPS) or italicized.

In addition to being given the QxQs and instructions to translate with respect to intended meaning, Group C received an instruction to take what liberties were necessary to translate items in a way that sounded natural and culturally appropriate, given normal ways of using language and expressing meaning in interaction:

> **IMPORTANT:** There are culturally different ways of using language in interaction. Please translate the survey items so that they sound as natural as possible in the context of a telephone interview. This means that the questions should not sound awkward to the survey respondents, and the questions should be phrased in a culturally appropriate way. Feel free to make whatever changes necessary to accomplish this.

Group C translators were instructed to attempt to achieve this aim *at the same time* as translating in a way that was faithful to the intended meaning of the source items. Exhibit 3 summarizes the three instructional subgroups:

Exhibit 3.—Instructions received by the three instructional subgroups

Group A	Core instructions		
Group B	Core instructions	QxQs and instruction for faithfulness to intended meaning	
Group C	Core instructions	QxQs and instruction for faithfulness to intended meaning	Instruction for cultural appropriateness

Evaluation of Translations by Survey Researchers

Westat recruited 15 professional survey researchers (five for each language in the study) to serve as evaluators. To be selected, researchers needed to have at least several years of experience in designing and conducting surveys.[3] They also had to be native speakers of the target languages and needed to have lived in the U.S. (or Canada for the French evaluators) for at least 5 years. Researchers were paid a small stipend ($100 US) for their time and effort.

Once recruited, the evaluators received instructions (Exhibit 4), as well as an evaluation form. The evaluators were instructed to assess the translated survey items on 7-point Likert scales along three dimensions, namely "overall quality," "faithfulness to intended meaning of the source item," and "naturalness and cultural appropriateness." The evaluators were given definitions for each of the three dimensions. Faithfulness to intended meaning was defined as the extent to which a translation maximizes the chances that survey respondents grasp the meaning intended by the survey designer. Cultural appropriateness was defined as the extent to which the translation sounds natural and is consistent with the cultural and linguistic norms and values of the target population. The meaning of overall quality was intentionally left open to the evaluators themselves, in order to determine the extent to which this dimension correlated with the other two.

Each page of the evaluation form contained three translated versions of each survey item, followed by the QxQs and the scales to be rated (see Exhibit 5). The translations of items from the three instructional subgroups of translators were randomly placed on each page, so that evaluators would not be able to discern or be biased by any patterns of placement. Evaluators were told that the translations had been randomly ordered on each page.

[3] It should be noted that several of the Chinese evaluators specialized in data analysis and had limited experience in survey design, although all of the Chinese evaluators were employed in research settings.

Exhibit 4.—Instructions given to evaluators

Please read the following instructions:

1) After printing out all of the attached documents, read through the three-page source instrument once or twice to become familiar with its flow and overall sense. Also, review the instructions that were given to the translators.

2) On each of the following pages of the "evaluator form," you will see an English source item, followed by three alternative translations of that item. In the row beneath the three translations, you will see italicized text that explains the meaning and intent of the source item. Read carefully the source item, the alternative translations, and the italicized text.

3) After reading and reflecting on the source item, the alternative translations, and the italicized explanation of the meaning and intent of the item, rate each translation on a scale of 1 to 7 along the three dimensions provided (with 1 being "poor" and 7 being "excellent"). The three dimensions are explained below.

Dimension 1: **Overall quality**. Rate the translated items for their "overall quality," according to however **you** happen to understand or define translation quality.

Dimension 2: **Faithfulness to intended meaning**. "Meaning" is not something created solely out of words strung together, but rather something that comes from words used appropriately in context between people with communicative purposes. What we "mean" to say (or "intend") often goes beyond the actual words that we use. For the purposes of your task, "faithfulness to intended meaning" refers to the extent to which a survey item translation maximizes the likelihood that the survey respondent will grasp the same meaning *intended* by the survey designer. To help you understand the intended meaning of each survey item, be sure to read the italicized text that follows the translations.

Dimension 3: **Cultural appropriateness**. A survey translation that does not take into consideration the cultural values and norms of the target population may risk sounding awkward or even offensive. For example, asking a direct question such as "How old are you?" may be acceptable in one culture, but may be acceptable only if asked in a more indirect and polite way in another culture (e.g., "May I please ask the year in which you were born?"). The dimension of cultural appropriateness has to do with the extent to which the translated item sounds "natural" and is consistent with the cultural and linguistic norms and values of the target population.

Some other things to consider:

First, please note that there were nine translators of the English source items. The translations of items from these nine translators are mixed in *random order* throughout the evaluator's instrument, and they may appear in any column.

Second, all of the nine translators were native speakers of the target language and had at least five years of full-time experience in professional translation work. In addition, all nine translators had at least a bachelor's degree and were highly fluent in English.

Finally, keep in mind that the survey is meant to be administered orally by telephone. Thus, your ratings of the individual translated items should be sensitive to the fact that they were translated under the assumption that they would be *spoken* by a telephone interviewer within a real time interaction.

Exhibit 5.—Example page from Spanish evaluation form

Question 1			
Researchers' explanation	*For this item, we want to know the approximate number of days in the past year that the respondent's child did not go to school because he/she was sick or had an injury. The illness might have been a head cold or flu or related to an ongoing health problem or condition of the child. An injury is a physical problem resulting from some kind of accident. The question is concerned only with the child's absence from school within the previous 12 months, working backward from the time of the survey interview. (For example, if the survey was conducted on February 1, 2006, then the past 12 months would be from February 1, 2005 to February 1, 2006.)*		
Original English	1) During the past 12 months, that is since (12 mo. ref. date), about how many days did (CHILD) miss school because of illness or injury? _____ _____ _____ DAYS [RANGE 0-240] CHECK HERE IF CHILD NOT YET IN SCHOOL: ☐		
Translations	Durante los 12 meses pasados, o sea desde (fecha de referencia de hace 12 meses), ¿más o menos cuántos días tuvo que faltar a la escuela (CHILD) porque estaba enfermo(a) o lastimado(a)?	Durante los últimos 12 meses, es decir, desde (fecha de referencia de hace 12 meses), ¿cuántos días ha faltado (CHILD) a la escuela debido a una enfermedad o lesión?	Durante los 12 meses previos, es decir desde (fecha de referencia de hace 12 meses), ¿aproximadamente cuántos días faltó (CHILD) a la escuela debido a enfermedad o lesiones?
Evaluator's Rating	Overall quality of translation: Poor Satisfactory Excellent 1 2 3 4 5 6 7	Overall quality of translation: Poor Satisfactory Excellent 1 2 3 4 5 6 7	Overall quality of translation: Poor Satisfactory Excellent 1 2 3 4 5 6 7
	Faithfulness to intended meaning: Poor Satisfactory Excellent 1 2 3 4 5 6 7	Faithfulness to intended meaning: Poor Satisfactory Excellent 1 2 3 4 5 6 7	Faithfulness to intended meaning: Poor Satisfactory Excellent 1 2 3 4 5 6 7
	Cultural appropriateness: Poor Satisfactory Excellent 1 2 3 4 5 6 7	Cultural appropriateness: Poor Satisfactory Excellent 1 2 3 4 5 6 7	Cultural appropriateness: Poor Satisfactory Excellent 1 2 3 4 5 6 7

Interviews with Evaluators

After completed evaluation forms were received by study staff, brief 15-20 minute individual telephone interviews were conducted with the evaluators**,** following a prepared protocol (Exhibit 6). The object of the interviews was to obtain detailed feedback from the evaluators on the evaluation task, and to discuss some general issues regarding beliefs and practices in survey translation. Interviews included a discussion of challenges faced in assigning ratings for the three dimensions and criteria employed by evaluators to assess overall quality. The qualitative data collected from the interviews were intended to supplement and shed light on the quantitative data provided in the ratings from evaluators.

Exhibit 6.—Protocol for interviews with evaluators

1) Background information about respondent – title and position, experience in survey design, etc.

2) Overall impressions

- How easy was the task?
- How much time did you spend on the task?
- What challenges did you face in assigning ratings for the different dimensions?
- Were the instructions clear? Where they helpful?
- How did you assign ratings for overall quality? (i.e., how did you judge overall quality?)
- Did overall quality for you include criteria besides faithfulness to intended meaning and cultural appropriateness?
- Did you find that the italicized text helped you to understand the "intended meaning" of the items?
- Do you believe that providing QxQs to translators is helpful?
- How difficult was it to assess the cultural appropriateness of the translations?
- What sorts of things were you looking at when judging the cultural appropriateness of translated items?

3) Impressions and thoughts on individual items

After reviewing the ratings for individual items, probe on why ratings were done as they were, in cases where something stands out. For example, ask why an item was rated high for faithfulness to intended meaning, but low for cultural appropriateness, or vice versa. Or ask why an item was given a low rating for overall quality, but high ratings for the other two dimensions.

Analysis of Data

Analysis of the data consisted of obtaining mean scores both overall and on an item by item basis for translations following the different sets of instructions. This was followed by statistical testing (primarily ANOVAs) to determine whether there were statistically significant differences in the average ratings between the translations following different instructions (overall and for each item). Such analyses allowed us to determine whether, for example, translations that followed the QxQs were rated higher for "faithfulness to intended meaning," than translations with no QxQs, or whether translations with instructions to ensure naturalness and cultural appropriateness were rated higher along this dimension.

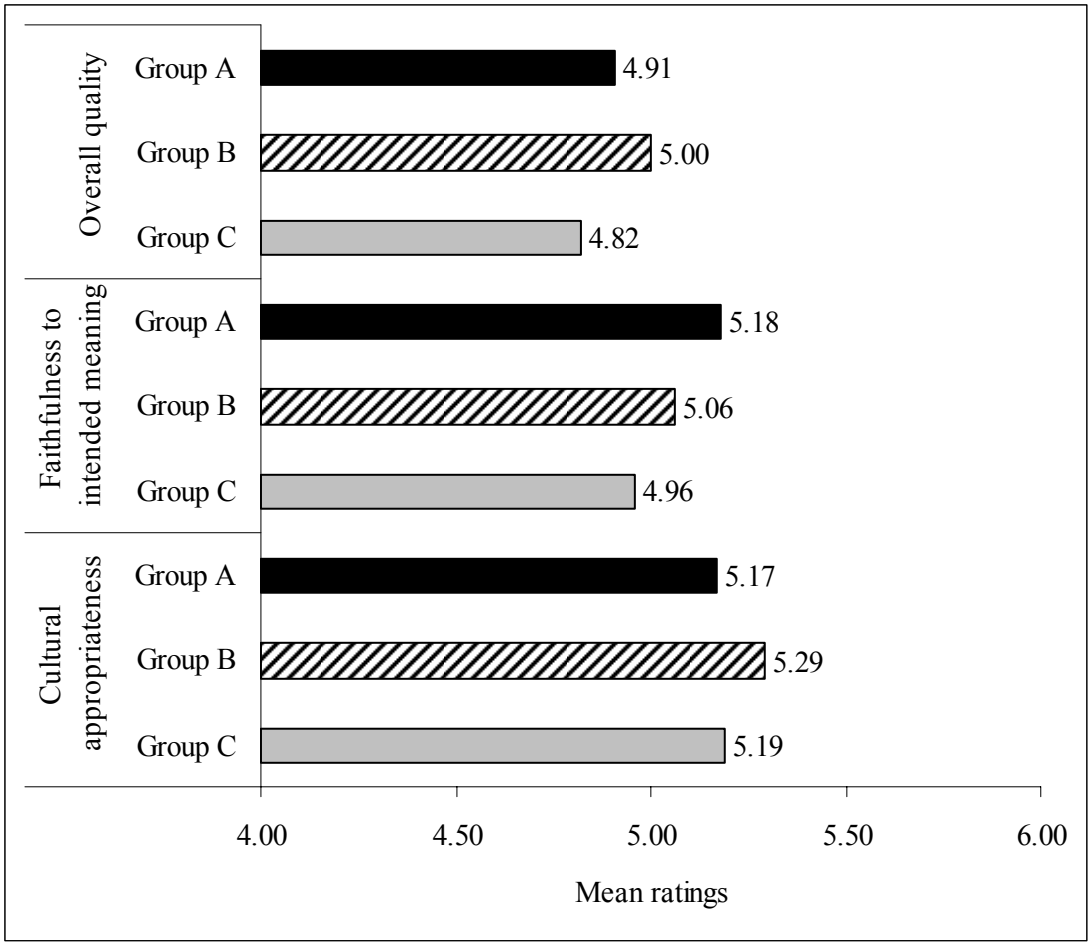
Analyses included examination of results by language in order to determine whether there were language-specific effects associated with the ratings. We also employed tests of correlation to determine whether, for example, higher ratings for faithfulness to intended meaning were negatively or positively correlated with ratings for cultural appropriateness. Finally, we analyzed the qualitative data collected in the telephone interviews with evaluators, with a focus on how the interview data help to account for the various quantitative findings.

Study Results

Quantitative Findings

While evaluators assigned slightly lower ratings to translations for overall quality than for faithfulness to intended meaning and for cultural appropriateness, the different sets of instructions appear not to have had a significant effect on the survey translations for any of the three rated dimensions (Figure 1). This overall lack of difference was confirmed by statistical tests of ANOVA.

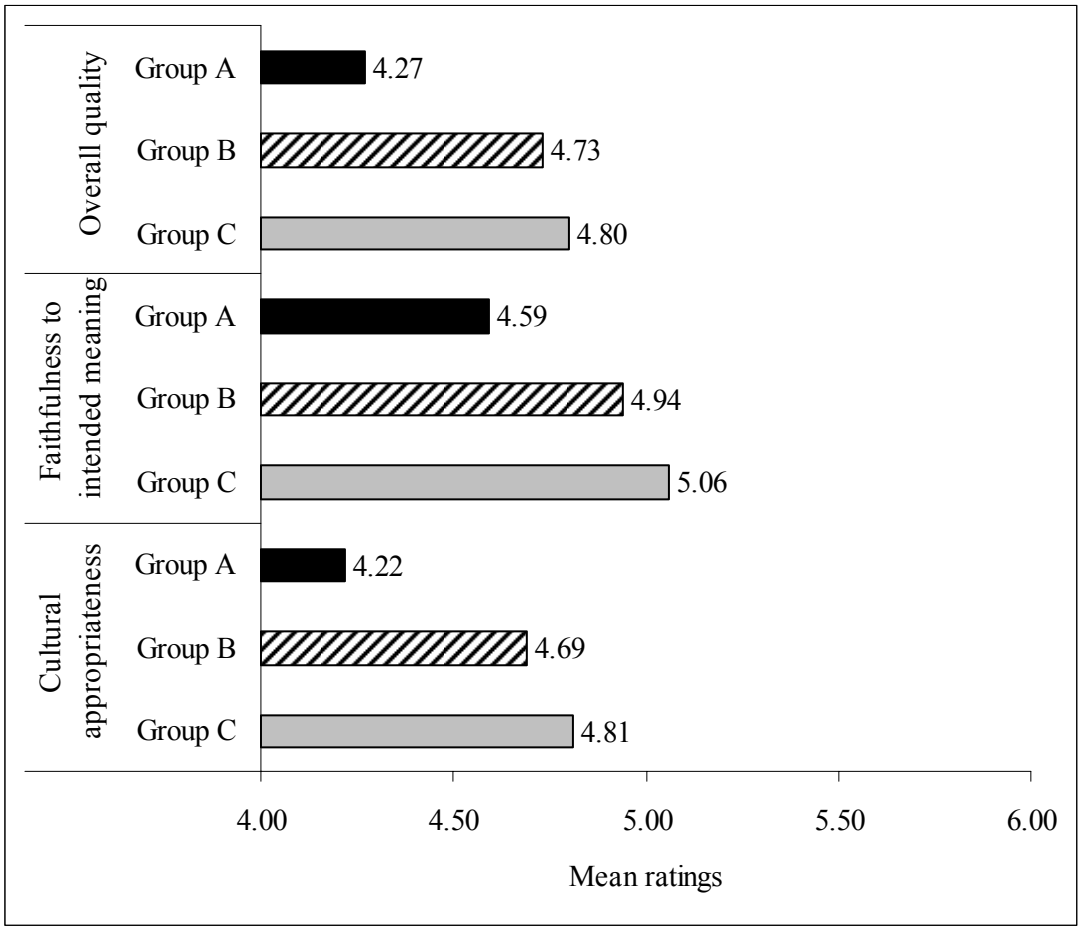
Figure 1.—Mean ratings for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup



Note: Group A translators received only the core instructions. Group B translators received the core instructions plus QxQs. Group C translators received the core instructions plus QxQs plus the cultural appropriateness instruction.

However, some intriguing patterns emerge upon examination of the results *by language*. In the case of the French translations, ratings assigned to Group B and C translations were higher than ratings for Group A translations for all three dimensions (Figure 2). A test of ANOVA reveals that the ratings for Group C were significantly higher than those for Group A for both overall quality (4.81 mean score versus 4.22 mean score) and cultural appropriateness (4.80 versus 4.27).

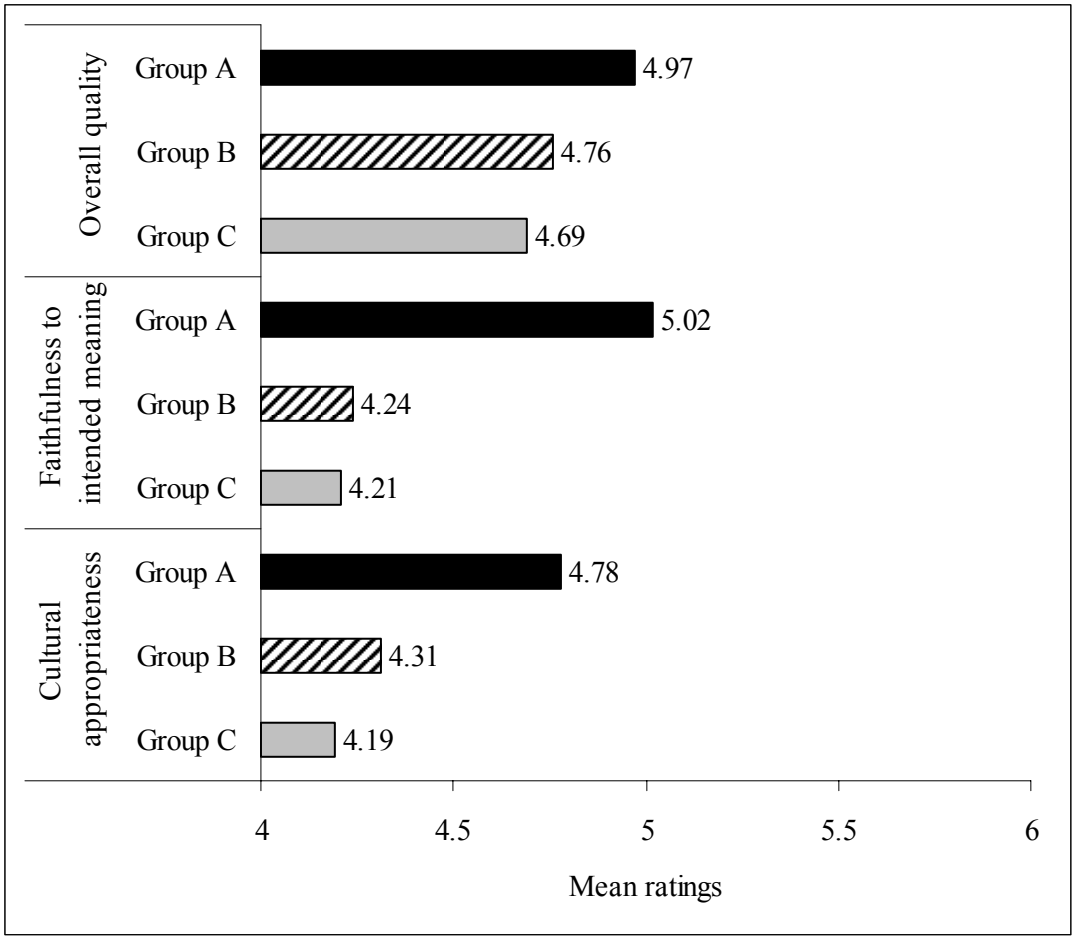
Figure 2.—Mean ratings for French translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup



Note: Group A translators received only the core instructions. Group B translators received the core instructions plus QxQs. Group C translators received the core instructions plus QxQs plus the cultural appropriateness instruction.

In the case of the Spanish translations, the pattern was reversed—Group A translations outscored Groups B and C translations across all three dimensions, most notably for faithfulness to intended meaning, and less decisively for cultural appropriateness (Figure 3). An ANOVA test showed that scores for Group A translations were statistically significantly higher than Group C scores for overall quality (4.78 versus 4.19), and that Group A scores were higher than Group B and C scores for faithfulness to intended meaning (5.02 versus 4.24 and 4.21, respectively).

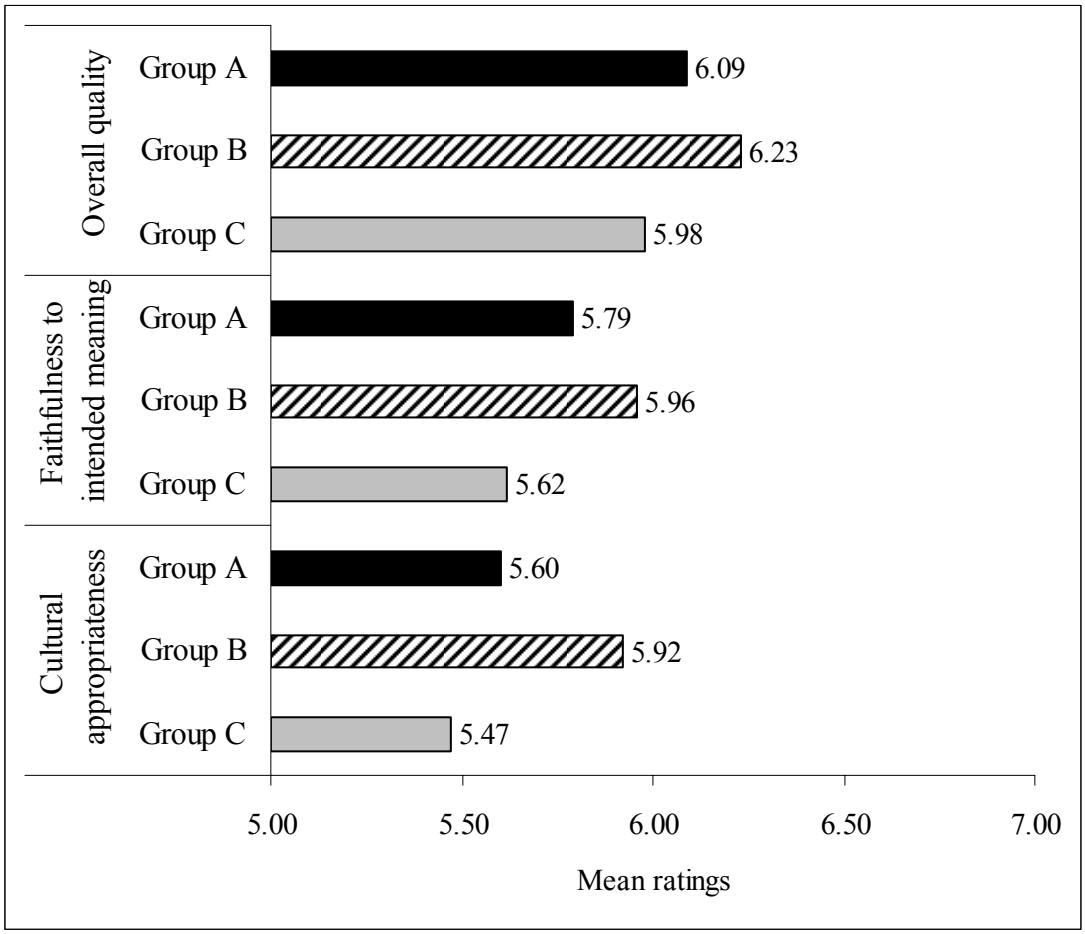
Figure 3.—Mean ratings for Spanish translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup



Note: Group A translators received only the core instructions. Group B translators received the core instructions plus QxQs. Group C translators received the core instructions plus QxQs plus the cultural appropriateness instruction.

In the case of the Chinese translations, the ratings for the Group B translations were higher than ratings for the other groups across the three dimensions (Figure 4). In addition, for each of the three dimensions, the ratings for the Group C translations lagged behind those for Groups A and B. A test of ANOVA indicates that the cultural appropriateness ratings for Group B were statistically significantly higher than for Group C (5.92 mean score versus 5.47 mean score).

Figure 4.—Mean ratings for Chinese translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup



Note: Group A translators received only the core instructions. Group B translators received the core instructions plus QxQs. Group C translators received the core instructions plus QxQs plus the cultural appropriateness instruction.

Comparison of ratings across languages shows that the Chinese evaluators generally gave higher ratings than the French and Spanish evaluators to translations along the three dimensions.[4] Tests of correlation reveal that the ratings assigned to the three dimensions had a positive correlation. That is, for example, higher ratings along one dimension tended to co-occur with higher ratings along another dimension. Finally, examination of mean scores (ANOVA) for individual survey items revealed no

[4] The higher average ratings assigned by the Chinese evaluators compared to the French and Spanish evaluators may have more to do with the cultural tendencies of the evaluators than the quality of the translations. On the other hand, as noted by Nida 1964, translations between more closely related languages may result in greater problems in translation than translation between distantly related languages due to superficial similarities.

statistically significant differences by instructional subgroup for any of the 18 survey items along any of the three dimensions.[5] Given the relatively small number of ratings per survey item, it was not possible to examine individual survey items in a similar way by language.

Qualitative Findings

The information collected from telephone interviews with the evaluators after they had completed their ratings helps to shed light on some of the quantitative findings. Specifically, feedback from the interviews reveals how evaluators conducted their assignment of ratings and may account in part for the language-specific findings for the various dimensions. The interviews also brought out various beliefs held by evaluators regarding survey translation and suggest that these played a significant role in how they assigned ratings for the three dimensions.

Evaluator Background

The evaluators recruited for this study were required to be professional researchers with some experience in survey design. In fact, while all of those recruited were professional researchers, their level and types of experience involving survey research varied. Nine of the evaluators had many years of experience in developing and evaluating surveys in both English and their native language (i.e., Mandarin Chinese, Spanish, or Canadian French). Four had less experience in designing surveys and more in other aspects of survey research, such as survey operations (e.g., training field staff, developing materials, overseeing collection procedures), interviewing, programming questionnaires for CATI, data analysis, and survey translation. Two of the evaluators had no direct experience with either survey design or survey translation.

Information from the interviews revealed that the Spanish and French evaluators had more direct experience in survey design than the Chinese evaluators. In fact, while all of the French evaluators and four of the five Spanish evaluators had years of experience in developing surveys and overseeing and evaluating survey translations, *none* of the Chinese evaluators was primarily responsible for survey design and translation. Three Chinese evaluators primarily conducted analyses of survey data, and all but one reported having done some translation work, but not necessarily for surveys.

It is likely that the backgrounds of the evaluators may have influenced how they conducted their ratings of the translated survey items in the study. Interview data make clear that the evaluators came from very different backgrounds. Besides being from different cultures, the evaluators were from a wide range of educational backgrounds, including anthropology, political science, public policy, journalism, and computational linguistics. They were also employed at various types of professional research organizations, including universities, federal agencies, private research firms, and polling companies.

Level of Difficulty of the Task

Interview data suggest that the task assigned to the evaluators was carried out with care and with respect to the instructions provided. When asked about the level of difficulty of the task, 10 of the 15 evaluators said that it was not difficult. All five of the Chinese evaluators said that the task was not difficult, while three of the Spanish and two of the French evaluators said that the task was not difficult.

[5] Significant differences might have been detected with a larger sample size.

Of those who said that the task was difficult, three explained that it required considerable concentration and attention to detail. One said that the difficulty was in assigning ratings to scales that were not helpful or clearly distinct (see below), and another said that in some cases it was difficult to choose an acceptable translation.

Evaluators spent on average about 2 hours on the task, and none spent less than 1.5 or more than 3 hours on the task. All 15 evaluators said that the instructions provided to them were clear and helpful. Several commented that the evaluation instrument was well designed and easy to follow. One evaluator said that the evaluation form should have included a fourth scale—clarity and succinctness. Another said that we should have included a fourth scale called "mode appropriateness."

Assignment of Ratings

Interviews with the evaluators revealed that, while all understood the instructions, definitions, and purpose of the task, most employed the definitions to assign ratings in the light of their own beliefs and notions about features and standards of survey translation.

Evaluators were unanimous in saying that the translations were of good quality overall. There were, however, a few weaknesses pointed out by the evaluators, which appear to have been language specific. First, three of the five Chinese evaluators noted that the translations were sometimes too wordy and lengthy. This may be linked to the fact that the translations were sometimes too "formal" according to these evaluators, and therefore used language more appropriate to written types of discourse. While most of the translations were technically accurate and reflected the intended meaning of the source items, some came across as too wordy and awkward. It is interesting to note that while the Spanish and French evaluators clearly prized brevity in translation, none reported wordiness as a significant problem.

Second, three of the Spanish evaluators complained that the translators sometimes took too many liberties in adapting the translations of source items. Specifically, they felt that some translations departed too much from the wording and concepts of the source items, and therefore somehow compromised equivalence of meaning. Several commented that material borrowed from the QxQs had been imported into the translation for clarification, but that this was problematic, since the source item did not include this material. However, this issue was not raised by the Chinese or French evaluators.

As for the task of assigning ratings, some comments from evaluators shed light on how the task was carried out, as well as on some of the challenges encountered. The process of assigning ratings appears to have been done in a similar fashion across all of the evaluations. Most evaluators reported assigning ratings after having read the source item and QxQs and after reading and comparing the alternative translations on each page of the evaluation form. This indicates that the ratings assigned to items along the different dimensions were not absolute in nature, but rather were relative and due to a comparative process.

Many evaluators commented on the relationship among the various dimensions. Some said that the faithfulness to intended meaning and the cultural appropriateness dimensions were not truly independent but were overlapping. One Spanish evaluator said that during the evaluation process he realized to his surprise that faithfulness to intended meaning and cultural appropriateness often did not go hand in hand. He found that some translations tended to be very culturally appropriate but did not convey the original meaning and intent. He wondered whether the translators had felt that by simplifying the questions (leading to departures from the original meaning) they were adapting them culturally. He said that the definition of "cultural appropriateness" provided to translators would definitely influence how such adaptations finally took shape.

On the other hand, a French evaluator said that he did not believe that there is necessarily a conflict between faithfulness to intended meaning and cultural appropriateness. According to this evaluator, these should really go hand in hand. A good translation should capture the intended meaning and concepts of the source item, but should be recast in the appropriate words, phrases, expressions, and structures of the target language.

One Chinese and one Spanish evaluator said that the overall quality scale should have followed the other two scales, rather than preceding them on each page. One evaluator noted that assigning ratings was especially difficult for the longer translations. The following sections focus on findings about each of the three dimensions.

Faithfulness to intended meaning. It is clear that most evaluators believed faithfulness to intended meaning to be a key component of a quality survey translation. Feedback from interviews with evaluators suggests that faithfulness to intended meaning was the easiest dimension of the three to assess. Moreover, interview findings indicate that different evaluators had slightly different notions of this dimension, and that this varied by language.

Generally, the Spanish evaluators appear to have been more focused on equivalence of concepts—that is, that translations should include all *and only* the concepts contained in the source item. Further, they employed their beliefs about conceptual equivalence rather strictly in their ratings of faithfulness to intended meaning. On the other hand, the French evaluators seem to have held less rigid views about the role of equivalence, contenting themselves with translations that captured the intended meaning, even if concepts not in the source made their way into the translations. The Chinese evaluators fell somewhere in between the French and Spanish evaluators, with several noting that the need for equivalent concepts and terms may have necessitated some departures from the original text.

Cultural appropriateness. Most evaluators acknowledged the importance of cultural appropriateness as a requisite feature of survey translation. All of the Chinese evaluators said that it was not difficult to assess the cultural appropriateness of the translations. One Spanish evaluator said it wasn't difficult at all. A second Spanish evaluator said that this was the most difficult dimension, because she could not rate the cultural appropriateness of any dialect but her own (Peruvian/South American Spanish). A third Spanish evaluator indicated that she rated all translations high on cultural appropriateness because the English formulation was already culturally appropriate. One French evaluator noted that cultural appropriateness could include many different things, and so assigning ratings was "perplexing." Another French evaluator said that he rarely focuses on cultural appropriateness when he has translations done for his surveys in Canada, because the translators are always native speakers of Canadian French, and so they know "automatically" how to make their translations natural sounding and appropriate.

While most evaluators followed the definition provided in assigning ratings for cultural appropriateness, it appears that each focused on different aspects of this dimension. Many paid most attention to the "smoothness" or "naturalness" of the translation, and whether or not it sounded like a translation. They appear to have employed a very intuitive approach, looking simply for whether the translation read well, "sounded right," and was not too awkward. Several evaluators focused primarily on the extent to which the translation might be offensive to respondents. One evaluator said she looked for the appropriate "tone" in her assignment of ratings.

Still others reported emphasizing the level of formality of the translation (given the respondents and type of discourse) and the appropriateness of the translation for a telephone (oral) conversation. One of these evaluators said that many of the translated items were "too high class." Other evaluators focused most on whether the translation used language that would be accessible and comprehensible to

respondents of different ethnicities, regions, and educational levels. One of the French evaluators noted that there is an inherent tension between translating with words that are precise and efficient for conveying meaning and at the same time accessible to most people.

Overall quality. The definition of overall quality provided to evaluators was purposefully left open to them. This was done for several reasons. First, we did not want to force a definition on evaluators that they might not share. Second, doing so would help us to determine whether, according to professional survey practitioners, there are salient features of survey translation quality besides faithfulness to intended meaning and cultural appropriateness. Third, it would allow us to examine the extent to which both faithfulness to intended meaning and cultural appropriateness are correlated with overall quality.

Some of the evaluators appreciated the lack of a stated definition of overall quality, while others were uncomfortable with this. Several evaluators said that overall quality was the most difficult scale to rate. One of the Chinese evaluators noted that the assignment of ratings for overall quality was an overly subjective measure and that different evaluators will bring with them different criteria for assigning ratings. One French evaluator said that his ratings of overall quality were based on "a feeling" that the translation was right. A Spanish evaluator said she imagined how comfortable she would feel if she had to present the translation to a client. Another Spanish evaluator said that the translations were very close in terms of overall quality, and so finding differences between them was challenging.

It is clear from interview data that many evaluators followed a rather intuitive and impressionistic approach to assigning ratings for overall quality. Nonetheless, when pressed, most evaluators were able to articulate several criteria that they used to assess this dimension. Most considered either faithfulness to intended meaning or some aspect of cultural appropriateness when assigning ratings for overall quality. At least eight of the 15 evaluators said that, for them, overall quality included elements of *both* faithfulness to intended meaning and cultural appropriateness. A surprising number (almost half) also referred to "clarity," "brevity," and "ease of comprehension" as important features of overall quality.

The Role of QxQs

Asked whether the QxQs helped them to understand the intended meaning of the survey items, eight of the 15 evaluators responded that they were helpful, with no qualification. One of these evaluators commented that he would probably have inferred the same intent of the questions as the QxQs indicated, but that even he, as a researcher, may have missed a few things that the QxQs clarified. One said that the QxQs were helpful for certain questions, but less so for others. Another mentioned that reference to the QxQs led her to change her ratings on several occasions.

Three evaluators said that the QxQs were helpful but not really that critical, since the intended meaning of the survey items was already clear to them. One said that he was experienced enough as a survey designer that he knew the intended meaning of the questions just by reading the questions themselves. In addition, he said that the factual nature of most of the source questions made it easier to grasp their intended meaning. He noted that the source questions were very well written and clear, and so there was less of a need (for him) to examine the QxQs.

One evaluator was ambivalent about the QxQs, saying that they were somewhat helpful but a distraction at the same time. Another said that she only read the QxQs very rapidly and that they were not that helpful to her.

Evaluators were also asked about their beliefs regarding the benefits of providing QxQs to translators. While some said they would wholeheartedly endorse this practice, others said they would

support it only under certain conditions, and others said that it does more harm than good. Seven of the 15 evaluators believed that providing QxQs to translators was a good practice under any circumstances. One of the seven noted that, while translators may "have the linguistics down," they may not have a good grasp of the research intent of questions. He said that the QxQs provide guidance to translators who may not have a lot of background in research, especially in terms of methodology and question response options.

One evaluator said that QxQs may be very helpful to translators, but they should be provided selectively and not for every question. The risk is that too many QxQs would become routine and lose their effectiveness. It would be better to provide QxQs only for those questions that are more subject to multiple interpretations on topics that might not be very familiar to translators and respondents. He noted that more experienced translators (i.e., experienced in translating surveys) would have less need for QxQs, and that less experienced translators would benefit more.

A second evaluator said that QxQs are especially helpful for source questions that have weaknesses, that is, that are unclear or ambiguous. He noted that well-constructed questions should not require QxQs. A third said that QxQs should be provided only for particular survey topics and particular items. The risk is that too much information could cause confusion and make translators think too hard about their translations. Similarly, another evaluator said that he thinks that providing QxQs is not necessary, because it adds too much to an already challenging task: "It might be a little overwhelming." Another evaluator said that more helpful than QxQs would be a set of specialized vocabulary and explanation of technical terms.

One evaluator said that she would not give QxQs to the translator at the outset of the translation process. She commented that giving translators extra information may sometimes be counterproductive, because it might open the door for the translator to question the way the item was designed in the first place. Instead, she would let the translator work on the translation and then, once complete, would give the translator the QxQs to help him or her evaluate whether or not the translated questions matched up with the intended meaning. She indicated that the translation process is already hard enough without adding this extra burden. Also, as an afterthought, she said it might be more helpful not to give the translator QxQs at all, but to give them to the last person to evaluate a translation (i.e., a coordinator or adjudicator).

Discussion and Implications

Examination of ratings for the three dimensions found no overall differences by instructional subgroup. However, it appears that the flat overall findings were masking some interesting and conflicting patterns among the three languages. For Chinese, the ratings for the Group B translations were on average higher than ratings for Groups A and C translations across the three dimensions. For each of the three dimensions the ratings for the Group C translations were lower on average than those for Groups A and B. For French, ratings assigned to Groups B and C translations were higher than ratings for group A translations for all three dimensions. On the other hand, for Spanish, Group A translations outscored Groups B and C translations across all three dimensions. Tests of correlation showed that the ratings assigned to the three dimensions were positively correlated.

Findings from the telephone interviews with evaluators suggest some possible explanations for the language-specific differences in the ratings. First, it is clear that the Spanish evaluators were generally opposed to the kind of adaptation carried out by the Group B and C translators, providing lower ratings for the translations of these groups for all three dimensions. Groups B and C were indeed given a degree

of latitude in their translations, where they could diverge from a close or literal translation, as long as the translations were faithful to the intended meaning of the source item (and were culturally appropriate at the same time for Group C). Several Spanish evaluators made note of the use of QxQ material in some of the translations. For these evaluators such departures are inappropriate, because they add text and concepts that do not exist in the source item, thus compromising equivalence. We believe that the QxQs are indeed responsible for the Spanish rating results, since Groups B and C both received QxQs, and their ratings patterned together in contrast to those of Group A.

It is interesting to note that one of the Spanish evaluators provided ratings that were less in line with those of the other four Spanish evaluators. Specifically, his ratings for Group A translations were not rated higher on average than his ratings for Group B and C translations. In fact, his ratings of Group C translations were slightly higher than for the other groups. He mentioned that one reason he may have rated the Group C translations higher than the other evaluators is because of the type of research he conducts (program evaluation) and his organization's pragmatic approach to translation. Since his organization attempts to build trust and cooperation with respondents in order to keep them involved in its research studies, it places greater emphasis on promoting comprehension and cooperation and less on ensuring strict equivalence between source items and their translations.

We might conclude that, for Spanish, the provision of QxQs to translators did more harm than good, at least from the perspective of this particular group of native Spanish-speaking professional survey researchers. In addition, the cultural appropriateness instruction had no measurable positive or negative effect on the translations of Group C. It is important to emphasize that the ratings did not occur in a vacuum and that the evaluators brought to bear their own particular beliefs and presuppositions regarding survey translation. Certainly the Spanish evaluators tended to place a high premium on "close" translation, enforcing the standard that there should be little conceptual departure from source items.

The French evaluators were clearly not averse to adaptation and divergence, as long as the translations captured the intended meaning of the source items. On the contrary, these evaluators actually rewarded such adaptation with higher ratings, including ratings for overall quality. Several of the Canadian French evaluators noted in their interviews their tolerance for small departures from the source item in translations, provided that the intended meaning would be comprehended by respondents.

It should be pointed out that the ratings for Group C French translations were not significantly higher on average than those for Group B (or A) for the cultural appropriateness dimension. With respect to the Canadian French evaluations, therefore, one might conclude that the provision of QxQs to translators (in Groups B and C) had a significantly positive impact on the translations, even for the cultural appropriateness dimension. However, as with the Spanish evaluation, the cultural appropriateness instruction provided to Group C did not have a measurable effect in either a positive or negative direction.

For the Chinese, providing the QxQs to Group B appears to have had a slightly positive impact on the ratings for all three dimensions, although the additional instruction to Group C seems to have *negatively* affected the ratings for that group for all three dimensions. Several of the Chinese evaluators noted the importance for them of concise translations. They indicated that the lengthy translations might actually interfere with comprehension and data quality. It is possible that the Group C translators took the liberty of adding language to make their translations more culturally appropriate, but that this was generally not received well by the Chinese evaluators. On the other hand, the Group B translators may not have added much text towards ensuring faithfulness to the intended meaning and so may not have been penalized by the evaluators.

There is a natural tension and potential for conflict between the need for equivalence of stimulus (prized by the trained professional survey researcher) and the equally important need for respondent

comprehension and cooperation. For example, adding text to a translated item that does not appear in the source item for the purpose of clarifying a term or idea that may be difficult in the target language may violate equivalence of stimulus but promote respondent comprehension. Clearly, an equivalent stimulus does not necessarily equate to an identical *effect* on a respondent.

Study findings suggest that the French and Chinese evaluators were willing to allow some degree of adaptation in translations of survey items, as long as the translations were faithful to the intended meaning of the source items.[6] On the other hand, the Spanish evaluators were opposed to such adaptation and deemed the Group B and C translations to be on average of lower quality along the three dimensions. This suggests on the surface that providing QxQs may be effective for some languages, but not for others. However, the average ratings across languages may have been more a reflection of the beliefs, backgrounds, and experiences of the evaluators, and so we should not necessarily conclude that the adapted versions of the translations were either better or worse in terms of their quality and the level of measurement error they could generate.

The study findings also suggest that the cultural appropriateness instruction given to Group C translators did not have any effect for French and Spanish, and even had a small negative impact on the Chinese ratings (for all three dimensions). This indicates that providing such an instruction (at least in the form employed in this study) will not have a considerable effect on translators and may even be detrimental. It is possible that the cultural appropriateness instruction written in a different way might have had a more positive impact on the translations.

While principles of survey research dictate equivalence of stimulus, there appears to be a shift (and growing consensus among researchers) toward equivalence of effect, as reflected in the call for faithfulness to intended meaning in translation (references). This conflict between equivalence of stimulus and effect is paralleled in the adopt/adapt debate, and in the general literature on translation (e.g., Nida's distinction between "formal and dynamic equivalence"). Survey researchers who deal with translation have different beliefs about the primacy of equivalence of stimulus or effect, and this may have led to the language-specific ratings in our study.

It is evident from the evaluator ratings that providing special instructions to translators in the form of QxQs *will* have an effect on the resulting translations. The question is, is the effect desirable under all conditions for all languages? Our exploratory research suggests that the issue is more complex than assumed, and that researchers should consider carefully in advance whether providing such instructions is necessary, given the nature of the survey, as well as their own beliefs about adaptation and the primacy of equivalence of stimulus or equivalence of effect.

Future Research

In addition to demonstrating that different instructions to translators may have variable effects on survey translations, our research suggests that survey researchers vary in terms of where they fall on the equivalence of stimulus/effect continuum. Some researchers, following strictly the guiding principle of standardization, tend to dislike adaptation and evaluate survey translations with a preference for equivalence of stimulus. Others are more tolerant of adaptation and evaluate translations with a preference for equivalence of effect.

[6] It must be assumed that adaptation for the Group B Chinese translations did not necessarily involve the lengthening of translations.

While the research described in this paper points to the preferences and beliefs of survey researchers, it did not weigh in on the issue of how translations that fall along the equivalence of stimulus/effect continuum are *received and understood by actual respondents*. Future research should address empirically how respondents respond to survey questions that have been translated according to differing instructions and guidelines.

Such an empirical study should be designed to address several interrelated questions. First, do survey questions translated following an equivalence of stimulus approach (less adaptation) lead to significantly greater problems in respondent comprehension or cooperation than questions translated following an equivalence of effect approach (more adaptation)? If not, then translations that minimize adaptation and maximize equivalence of stimulus may be preferable, since they may be more likely to maintain the measurement properties of source items. If, on the other hand, there are greater problems of comprehension and cooperation, then researchers may need to reconsider the value of full allegiance to equivalence of stimulus in translation.

Second, do questions translated according to equivalence of effect risk departing from the measurement properties of source questions to such an extent that the resulting data are no longer comparable? If so, then researchers may need to be more skeptical about adaptation and equivalence of effect approaches to translation. If not, then perhaps equivalence of effect may need to be treated as the primary goal of survey translation, with adaptation serving as the means to achieving this.

We believe that empirical research with an appeal to actual respondents is the best hope for resolving the debate about adaptation in survey translation theory. For in the end, the ultimate measure of the various approaches to survey translation lies in the quality of the data received (however quality is defined). Within the same framework, other future research could focus on the positive or negative impact of QxQs on the translations of individual survey question types. In addition, given the exploratory nature of the work described in this report, replication on a larger scale and perhaps with other languages would help to establish the validity of the findings.

References

Gutt, E-A. (1991). *Translation and Relevance*. Cambridge, Massachusetts: Basil Blackwell.

Kleiner, B., and Pan, Y. (2006). Cross-cultural communication and the telephone survey interview. *Conducting Cross-National and Cross-Cultural Surveys: Papers from the 2005 Meeting of the International Workshop on Comparative Design and Implementation (CSDI). ZUMA-Nachrichten Spezial*. 12:81-90.

Harkness, J. (2003). Questionnaire translation. In Harkness, J. A., F. J. R. van de Vijver, and P. P. Mohler. (eds), *Cross-Cultural Survey Methods*. pp35-56. Hoboken, NJ: Wiley-Interscience.

Harkness, J. A. and A. Schoua-Glusberg. (1998). Questionnaires in translation. *Cross-Cultural Survey Equivalence.ZUMA-Nachrichten Spezial*. 3:87-128.

Nida, E. (1964). *Toward a Science of Translating,* Leiden: E.J. Brill.

Van Ommeren, M., Sharma, B., Thapa, S., Makaju, R., Prasain, D., Bhattarai, R., and De Jong, J. (1999). Preparing instruments for transcultural research: use of the translation monitoring form with Nepali-speaking Bhutanese refugees. *Transcultural Psychiatry* 36, 3:285-301.

APPENDIX A

SOURCE INSTRUMENT

Hello, my name is {INTERVIEWER NAME}. I am calling on behalf of the Centers for Disease Control and Prevention. We're conducting a nationwide immunization study to find out how many children under 4 years of age are receiving all of the recommended vaccinations for childhood diseases. Your telephone number has been selected at random to be included in the study. The questions I have will take only a few minutes.

Before we continue, I'd like you to know that your participation in this research is voluntary. You can skip any questions you don't want to answer, or end the interview without penalty. Your answers will be kept strictly private, in accordance with the Public Health Service Act. I can provide you with the specific legal citation if you like. It guarantees that any answers that identify you or your family will not be shared with anyone other than the agency doing this survey. Depending on the health characteristics of your children, these questions take between 5 and 25 minutes, but for most families, it's around 10 minutes. In order to evaluate my performance, my supervisor may record and listen as I ask the questions. I'd like to continue now unless you have any questions.

2) Is (CHILD) limited or prevented in any way in (his or her) ability to do the things most children of the same age can do?

   (1) YES
   (2) NO

3) Does (CHILD) have any kind of emotional, developmental, or behavioral problem for which (he/she) needs treatment or counseling?

   (1) YES
   (2) NO

4) During the past 12 months, that is since (12 MO. REF. DATE), about how many days did (CHILD) miss school because of illness or injury?

   _____ _____ _____ DAYS [RANGE 0-240]

   (996) DON'T KNOW

5) Did (CHILD) receive all the routine preventive care that (he/she) needed?

   (1) YES
   (2) NO (SKIP TO Q5)
   (6) DON'T KNOW

6) Why did (CHILD) not get the routine preventive care (he/she) needed?

   (1) COST TOO MUCH
   (2) HEALTH PLAN PROBLEM
   (3) NOT AVAILABLE IN AREA/TRANSPORT PROBLEMS
   (3) NOT CONVENIENT TIMES
   (4) DOCTOR DID NOT KNOW HOW TO TREAT OR PROVIDE CARE
   (5) OTHER

7) During the past 12 months, was there any time when you or other family members needed mental health care or counseling related to (CHILD)'s medical, behavioral, or other health conditions?

(1) YES
(2) NO (SKIP TO Q8)

8) How well do you think (CHILD)'s doctors and other health care providers communicate with each other about (CHILD)'s care? Would you say their communication is:

(1) Excellent,
(2) Very Good,
(3) Good,
(4) Fair, or
(5) Poor?
(6) COMMUNICATION NOT NEEDED
(96) DON'T KNOW

9) Now I have a few questions about health insurance and health care coverage for your child. At this time, is (CHILD) covered by health insurance that is provided through an employer or union or obtained directly from an insurance company?

(1) YES
(2) NO (SKIP TO Q10)

10) Does (CHILD)'s health insurance offer benefits or cover services that meet (his/her) needs? Would you say:

(1) Never,
(2) Sometimes,
(3) Usually, or
(4) Always?
(6) DON'T KNOW

11) The next question is about the amount of money paid during the past 12 months for (CHILD)'s medical care. Please do not include health insurance premiums or costs that were or will be reimbursed by insurance or another source. But do include out-of pocket payments for all types of health-related needs such as medications, special foods, adaptive clothing, durable equipment, home modifications, and any kind of therapy. During the past 12 months, would you say that the family paid more than $500, $250-$500, less than $250, or nothing for (CHILD)'s medical care?

(1) MORE THAN $500
(2) $250-$500
(3) LESS THAN $250
(4) NOTHING, $0
(6) DON'T KNOW

12) How many hours per week do you or other family members spend arranging or coordinating (CHILD)'s care? By this I mean making appointments, making sure that care providers are exchanging information, and following up on (CHILD)'s care needs.

_____ _____ _____ HOURS PER WEEK [0-168]

13) Now I have some questions about your household. Please tell me how many people live in this household, including all children and anyone who normally lives here even if they are not here now, like someone who is away traveling or in a hospital.

_____ PERSONS [RANGE 01-30]
(96) DON'T KNOW

14) What was the total combined income of your household in (FILL LAST CALENDAR YEAR), including income from all sources including wages, salaries, unemployment payments, public assistance, Social Security or retirement benefits, help from relatives and so forth? Can you tell me that amount before taxes?

RECORD INCOME $_____
(96) DON'T KNOW

15) At any time during the past 12 months, even for one month, did anyone in this household receive any cash assistance from a state or county welfare program?

(1) YES
(2) NO
(7) REFUSED

16) Is (CHILD) of Hispanic, Latino, or Spanish origin?

(1) YES
(2) NO
(7) REFUSED

17) What is (CHILD)'s race? Is it…

(1) White,
(2) Black,
(3) Asian,
(4) Native Hawaiian or other Pacific Islander,
(5) Or some other race?

Those are all the questions I have. I'd like to thank you for the time and effort you've spent answering these questions. [TERMINATE]

APPENDIX B

QxQs

Hello, my name is {INTERVIEWER NAME}. I am calling on behalf of the Centers for Disease Control and Prevention. We're conducting a nationwide immunization study to find out how many children under 8 years of age are receiving all of the recommended vaccinations for childhood diseases. Your telephone number has been selected at random to be included in the study. The questions I have will take only a few minutes.

Much of this paragraph is aimed at establishing a basis of trust and honesty so that the potential respondent will not be afraid to participate in the interview. We tell the potential respondent who we are and why we are calling immediately in order to eliminate any mistrust or fear that the person on the other end of the phone is going to try to sell them something or coerce them in some way. Otherwise, the potential respondent could hang up, which would be a lost opportunity for the study. It is also important to tell them that their phone number was selected randomly, so they are not wondering who gave us their number, whether they are being investigated, and in order to rule out any other possible fears the respondent may have about being contacted by a complete stranger.

Before we continue, I'd like you to know that your participation in this research is voluntary. You can skip any questions you don't want to answer, or end the interview without penalty. Your answers will be kept strictly private, in accordance with the Public Health Service Act. I can provide you with the specific legal citation if you like. It guarantees that any answers that identify you or your family will not be shared with anyone other than the agency doing this survey. Depending on the health characteristics of your children, these questions take between 3 and 10 minutes, but for most families, it's around 5 minutes. In order to evaluate my performance, my supervisor may record and listen as I ask the questions. I'd like to continue now unless you have any questions.

This introduction describes the purpose and sponsor of the telephone survey to the potential respondent. It also provides important information about several features of the survey (e.g., that participation is voluntary, that the information collected is confidential, that the survey interview averages about 5 minutes, and that it will be recorded). Again, much of this paragraph is aimed at creating trust in the study and the purposes of the study. We also want people to know they are under no obligation to participate. We don't want them to feel coerced, as that might affect the honesty with which they reply to our questions. We also ensure them that their privacy will be respected, so that they will feel free to answer completely honestly (and because we are required by law to do so).

1) During the past 12 months, that is since (12 mo. ref. date), about how many days did (CHILD) miss school because of illness or injury?

   _____ _____ _____ DAYS [RANGE 0-240]

   CHECK HERE IF CHILD NOT YET IN SCHOOL: ☐

For this item, we want to know the approximate number of days in the past year that the respondent's child did not go to school because he/she was sick or had an injury. The illness might have been a head cold or flu or related to an ongoing health problem or condition of the child. An injury is a physical problem resulting from some kind of accident. The question is concerned only with the child's absence from school within the previous 12 months, working backward from the time of the survey interview. (For example, if the survey was conducted on February 1, 2006, then the past 12 months would be from February 1, 2005 to February 1, 2006.)

2) During the past 12 months, did (CHILD) receive all the routine preventive care that (he/she) needed?

(1) YES (SKIP TO Q6)
(2) NO

This question aims to find out whether or not, in the respondent's opinion, his or her child received adequate routine preventive care to avoid illness, injury, or a worsening preexisting health problem within the previous 12 month period. "Preventive care" is defined as measures taken in advance by health care providers that emphasize prevention, early detection, and early treatment of illness, injury, or long term health problems. "Routine" here should be taken to mean scheduled as a regular check-up (rather than for an occurrence of sickness).

3) Why did (CHILD) not get the routine preventive care (he/she) needed?

(1) COST TOO MUCH
(2) HEALTH PLAN PROBLEM
(3) NOT AVAILABLE IN AREA/TRANSPORT PROBLEMS
(3) NOT CONVENIENT TIMES
(4) DOCTOR DID NOT KNOW HOW TO TREAT OR PROVIDE CARE
(5) OTHER

This question is addressed to respondents who answered "no" to the previous question. Respondents are asked for the reason or reasons why the routine preventive care needed by the child was not received. This question is designed to identify what exactly kept the child from getting the care (e.g., too expensive, problem with health plan).

4) Is (CHILD) limited or prevented in any way in [his/her] ability to do the things most children of the same age can do?

(1) YES
(2) NO

For this item, we want to know if the respondent's child has trouble doing the same activities that "normal" children at the same age can do with little difficulty. These things might be relatively simple, like brushing teeth or putting on clothes, or they might be more complicated, like playing games or singing songs. Also, the problems might be physical or mental. The answer to this question would be "yes" if the respondent's child is "limited" (i.e., is able to do these things, but not as well, or is "prevented," meaning that his/her condition makes doing these things impossible.

5) Does (CHILD) have any kind of emotional, developmental, or behavioral problem for which (he/she) needs treatment or counseling?

(1) YES
(2) NO

Here we want to know whether the respondent's child currently needs to receive treatment or counseling for any problem that could be emotional, developmental, or behavioral. "Treatment" includes physical therapy or medication, whereas "counseling" involves interaction and talking with trained specialists. "Emotional" problems are where extreme feelings (e.g., fear, anger, anxiety) may be harmful and uncontrollable. "Developmental" problems are where children have not attained the natural physical and/or cognitive abilities normally attained by children at the same age. "Behavioral" problems are where children's actions may be destructive to themselves or to others.

6) How well do you think (CHILD)'s doctors and other health care providers communicate with each other about (CHILD)'s care? Would you say their communication is:

(1) Excellent,
(2) Very Good,
(3) Good,
(4) Fair, or
(5) Poor?
(6) COMMUNICATION NOT NEEDED

For this item, we ask the respondent to give us his/her opinion about the how well the child's doctors and other health care providers exchange important information about the child's care. For the purposes of this item, we assume that more than one doctor and health care provider (for example, nurses, dentists, or family therapy counselors) care for the child. Here, we are interested in the quality of the communication (for example, its accuracy and effectiveness) and not the frequency. The communication might be by telephone, letter, email, face-to-face, etc.

Possible responses fall along a 5-point scale, where "excellent" is the highest rating, "poor" is the lowest rating, and "good" is an average rating between the two extremes. On this scale, "very good" is exactly in the middle between "excellent" and "good," as "fair" is exactly in the middle between "good" and "poor."

7) During the past 12 months, was there any time when you or other family members needed mental health care or counseling related to (CHILD)'s medical, behavioral, or other health conditions?

(1) YES
(2) NO (SKIP TO Q8)

This question aims to find out whether or not, at any time within the previous 12 months, the respondent or any other family members (inside or outside of the home) had to seek mental health care or counseling as a result of the child's condition. This means that the child's condition (either physical or psychological) was severe enough to necessitate treatment on the part of the adults caring for the child. Here, we are interested in cases where the respondent or family member actually received mental health care or counseling, and where the help received was from a professional trained in understanding human behavior, emotions, and how the mind works; this professional could be a psychologist, counselor, social worker, psychiatrist, etc.

8) Now I have a few questions about health insurance and health care coverage for your child. At this time, is (CHILD) covered by health insurance that is provided through an employer or union or obtained directly from an insurance company?

(1) YES
(2) NO (SKIP TO Q10)

Here we want to find out simply whether or not the child currently has any health insurance ("health care coverage" is the same thing). This health insurance could be provided (in full or in part paid) through a parents' employer, through a union, or it could be paid for by the parents directly to a health insurance company. Question 8 serves to allow those without health insurance for their children to skip around question 9.

9) Does (CHILD)'s health insurance offer benefits or cover services that meet (his/her) needs? Would you say:

(1) Never,
(2) Sometimes,
(3) Usually, or
(4) Always?
(6) DON'T KNOW

This question intends to find out the extent to which the actual health insurance policies that parents have for their children pay for the costs of the care needed for their children. Here, the child's "needs" could be anything related to his/her physical or mental condition requiring treatment, including care for illness or injury. The categories for this question include "never," "sometimes," "usually," and "always." "Sometimes" means on occasion, but not the majority of the time. "Usually" means the majority of the time, but not in every case.

10) The next question is about the amount of money paid during the past 12 months for (CHILD)'s medical care. Please do not include health insurance premiums or costs that were or will be reimbursed by insurance or another source. But do include out-of pocket payments for all types of health-related needs such as medications, special foods, adaptive clothing, durable equipment, home modifications, and any kind of therapy. During the past 12 months, would you say that the family paid more than $500, $250-$500, less than $250, or nothing for (CHILD)'s medical care?

(1) MORE THAN $500
(2) $250-$500
(3) LESS THAN $250
(4) NOTHING, $0
(6) DON'T KNOW

The purpose of this question is to determine the overall amount of money paid by families in the previous year (previous 12 months) for all medical care needed by their children. This should ***not*** *include money paid by families to health insurance companies for health care policies (called "premiums"). This should also not include costs paid for by the parents, but which were or will be repaid to the parents by a health insurance company or some other source.*

The total amount paid for by parents should include all the items listed in the question that have to do with "health-related needs," broadly defined. "Adaptive clothing" refers to clothing specially designed for people with physical or developmental disabilities, such as diapers. "Durable equipment" includes

devices to assist people with physical or developmental disabilities for improved mobility, such as wheelchairs. The amount paid should include costs summed across the respondents' family, and should fall into the listed categories ("more than $500," "$250-$500," etc.)

11) How many hours per week do you or other family members spend arranging or coordinating (CHILD)'s care? By this I mean making appointments, making sure that care providers are exchanging information, and following up on (CHILD)'s care needs.

_____ _____ _____ HOURS PER WEEK [0-168]

This question asks for the overall average number of hours spent each week by the respondent and other family members doing activities relating to the child's health care. These activities may include things like making appointments with doctors, counselors, therapists, etc., coordinating communication between different care providers, and seeking advice and further consultation ("following up"). For this question, "arranging or coordinating" care has to do mainly with time spent planning and communicating with health care providers. It should not include time spent traveling to and from appointments, shopping, etc.

12) Now I have some questions about your household. Please tell me how many people live in this household, including all children and anyone who normally lives here even if they are not here now, like someone who is away traveling or in a hospital.

_____ PERSONS [RANGE 01-30]
(96) DON'T KNOW

For this question, we want to know how many people normally live in the respondent's household (where "household" is defined as a group of people occupying a single dwelling). This includes children and all adults who live in the household at least half of the time, even if they are not living there at the time of the interview.

13) What was the total combined income of your household in 2005, including income from all sources including wages, salaries, unemployment payments, public assistance, Social Security or retirement benefits, help from relatives and so forth? Can you tell me that amount before taxes?

RECORD INCOME $_____
(96) DON'T KNOW

Here we want to find out the respondent's household's total combined income for the calendar year 2005 (i.e., from January to December). The total provided should combine income from all members of the household from all sources (e.g., wages, salaries, unemployment payments, etc.), including financial assistance from relatives or other people or institutions. It is common in surveys to classify data according to the economic level of respondents, in order to help the researchers identify trends that may be specific to a given economic level.

"Wages" include compensation to workers by the hour, day, or week for work performed. "Salaries," which is a similar concept, includes fixed compensation for services on a regular basis (usually over a long term). Public assistance is family or individual financial assistance provided by the federal, state or local government. For this question, we want to know the gross total amount, that is, the total amount before taxes are taken out. Please note that this question is ***not*** *asking for two separate totals (i.e., one before and one after taxes).*

14) At any time during the past 12 months, even for one month, did anyone in this household receive any cash assistance from a state or county welfare program?

(1) YES
(2) NO

This question aims to determine whether anyone in the household received any financial assistance from a state or county welfare program within the previous 12 months (counting backwards from the time of the interview). Generally people who receive financial assistance from a government agency are poor and have difficulty paying bills and providing for the basic needs of their families. Here we are interested in "cash" assistance, meaning money given (in the form of a check or electronic deposit) to recipients that is not expected to be repaid to the state or country in the future.

15) Is (CHILD) of Hispanic, Latino, or Spanish origin?

(1) YES
(2) NO

This question asks for whether or not the child is considered by the respondent to be of Hispanic, Latino, or Spanish origin. Here were are interested in whether the child is a person of Mexican, Puerto Rican, Cuban, Central or South American, or other Spanish culture or origin, regardless of race.

16) What is (CHILD)'s race? Is it…

a. White,
b. Black,
c. Asian,
d. Native Hawaiian or other Pacific Islander,
e. Or some other race?

With this question, we ask for the "race" of the child, specifically, whether the child is considered by the respondent to be "White," "Black," "Asian," "Native Hawaiian or other Pacific Islander," or "Some other race." It is common in surveys to classify data according to the race of respondents, in order to help the researchers identify trends that may be specific to a given racial subgroup.

Those are all the questions I have. I'd like to thank you for the time and effort you've spent answering these questions. [TERMINATE]