Survey Research Methods (2009)
Vol.3, No.3, pp. 113-122
ISSN 1864-3361
http://www.surveymethods.org

The Impact of Instructions on Survey Translation: An Experimental Study

Brian Kleiner
FORS-Swiss Centre for Expertise in the Social Sciences

Yuling Pan
US Census Bureau

Jerelyn Bouic
Westat

The experimental study described in this paper examined the impact of providing special instructions and supporting material to translators. Specifically, it addressed whether Spanish, Chinese, and French translators provided with explanatory text and guidelines were able to produce translations that were more faithful to the intended meaning of English source survey items, as well as that were more culturally appropriate and natural sounding compared to those of translators who received no such guidance. Study findings indicate that while the provision of special instructions and documentation to translators had a considerable impact on their translations, the direction of the impact (positive or negative) *differed across the target languages*, according to scale ratings of professional survey researchers who were native speakers of those languages.
Keywords: survey translation, equivalence, adaption, translation, instructions

Introduction

> *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*

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" (1964:161). 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" (1964:156). 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. For instance, Van Ommeren et al. (1999) provide examples illustrating 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 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

Contact information: Brian Kleiner, FORS-Swiss Centre for Expertise in the Social Sciences, e-mail: brian.kleiner@fors.unil.ch

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).

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" (1998:95). 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.

However, in our view, for most surveys that need translation (except perhaps now for some of the large-scale comparative international surveys), translators are generally *not* 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. Obviously, such interpretative freedom may give rise to mistakes in translation, or at least translations that are not optimal in some way.

While providing translators with materials that clarify the intended meaning of survey items seems reasonable on the surface, there is currently little empirical work that lends support to the utility of this practice. The exploratory study described in this paper 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.

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

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 (see below for details).

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 Exhibits 4 and 5). 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 differences were found in the ratings, then 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 United States 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 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). It was assumed that certain question types (e.g., sensitive questions, questions involving valuative assessments, and conceptually complex questions) would be more difficult to translate, especially without documentation on intended meaning, and would thus result in greater variation in ratings across the instruction groups.

The 18 adopted items were maintained without changes to wording and were placed in a logical order.[1] The National Survey of Children with Special Health Care Needs survey has been thoroughly 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 U.S. Census Bureau's most recent census form.

Recruitment of Translators

The 27 translators were recruited following the 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).[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. 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.

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.

[1] 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.

[2] All translators received instructions and materials individually by email, and there was no joint training.

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.

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

Fifteen professional survey researchers (five for each language in the study) were recruited 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.[4] 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.

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. 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.

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 analysis of variance testing of differences between means to determine whether there were statistically significant differences in the average ratings between the translations following different instructions (overall and for each

[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.

[4] All evaluators received instructions and the evaluation form individually by email. There was no joint training.

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.

item), with Bonferroni post hoc tests using multiple comparisons. 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. Results below a p-value of 0.05 are considered as significant. Analyses included examination of results by language in order to determine whether there were language-specific effects associated with the ratings.

In order to examine the relationship between survey question type and the ratings, survey items were grouped into the following dichotomous types: sensitive versus non-sensitive questions, short versus long questions,[5] subjective versus factual questions, yes/no versus non-yes/no questions, and questions requiring calculation versus those not requiring calculation. The mean ratings of the dichotomous groupings were then compared along each dimension using two-tailed t-tests. 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.

Given the relatively small number of translators and evaluators, it is possible that there may have been individual

[5] Short and long were defined around a mode value based on the number of words in the surveys' questions.

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 faltó (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:	Overall quality of translation:	Overall quality of translation:
	Poor Satisfactory Excellent 1 2 3 4 5 6 7	Poor Satisfactory Excellent 1 2 3 4 5 6 7	Poor Satisfactory Excellent 1 2 3 4 5 6 7
	Faithfulness to intended meaning:	Faithfulness to intended meaning:	Faithfulness to intended meaning:
	Poor Satisfactory Excellent 1 2 3 4 5 6 7	Poor Satisfactory Excellent 1 2 3 4 5 6 7	Poor Satisfactory Excellent 1 2 3 4 5 6 7
	Cultural appropriateness:	Cultural appropriateness:	Cultural appropriateness:
	Poor Satisfactory Excellent 1 2 3 4 5 6 7	Poor Satisfactory Excellent 1 2 3 4 5 6 7	Poor Satisfactory Excellent 1 2 3 4 5 6 7

effects, but these were not addressed in our study analytically. In addition, we did not conduct any linguistic analysis to determine the level of homogeneity or variation across translations within the three instruction groups.

Study Results

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 did not have a statistically significant effect overall on the survey translations for any of the three rated dimensions (Figure 1).

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 (4.80 mean score versus 4.27 mean score) and cultural appropriateness (4.81 versus 4.22).[6]

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

[6] With p=.036 and p=.015, respectively.

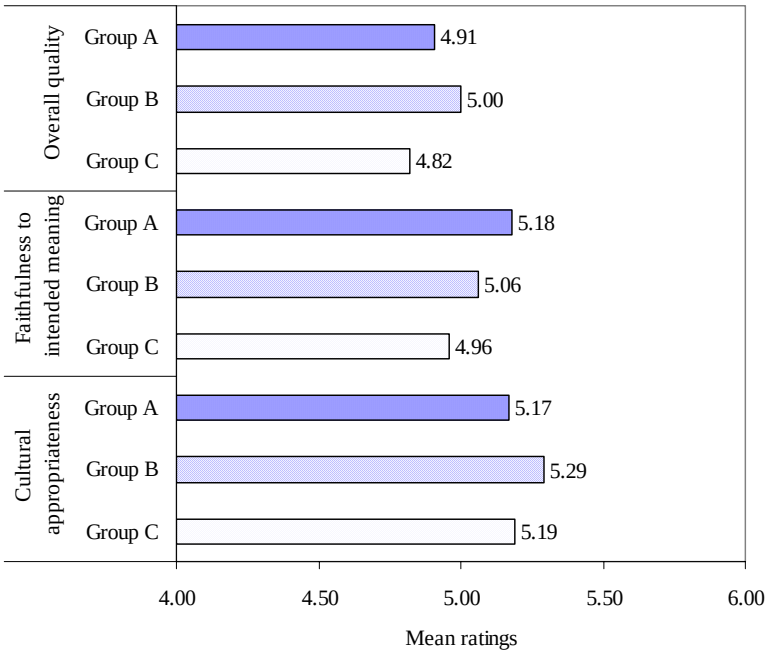

Figure 1. Mean ratings for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup. Group A translators received only the core instructions. Group B translators received the core instructions plus QxQs. Group C translators received the core instructions plus QxQs plus the cultural appropriateness instruction.

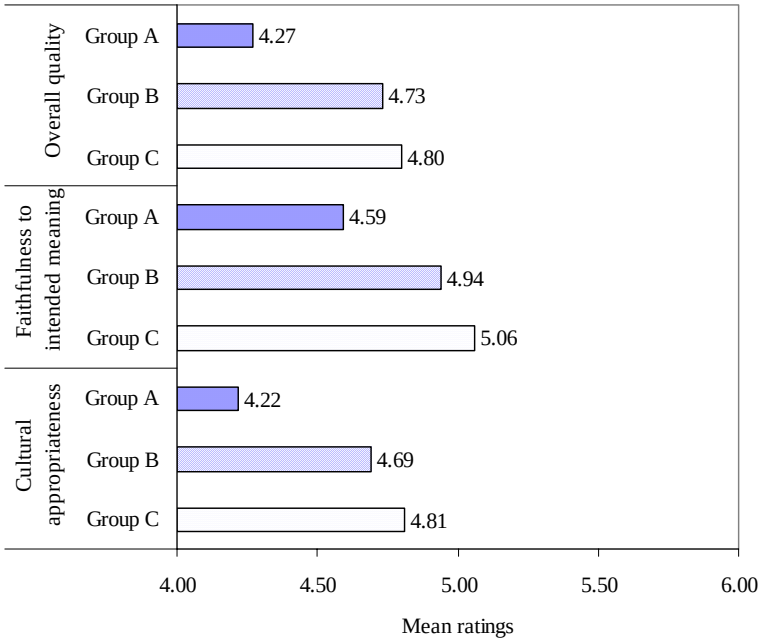

Figure 2. Mean ratings for French translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup. Group A translators received only the core instructions. Group B translators received the core instructions plus QxQs. Group C translators received the core instructions plus QxQs plus the cultural appropriateness instruction.

versus 4.19),[7] 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).[8]

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).[9]

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.[10] A Pearson correlation coefficient test revealed that the ratings assigned to the three dimensions had a positive correlation.[11] 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 statistically significant differences by instructional subgroup for any of the 18 survey items along any of the three dimensions.[12] Nor were there any statistically significant differences in ratings by any of the groupings by question type (i.e., sensitive versus nonsensitive, subjective versus factual, etc.). 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.

Discussion and Implications

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. 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.[13] 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. Clearly 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.

For French, ratings assigned to Groups B and C translations were higher than ratings for group A translations for all three dimensions. 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 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 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.

---

[7] With p=.024.

[8] With p=.003 and p=.002, respectively.

[9] With p=.028.

[10] The higher average ratings assigned by the Chinese evaluators compared to the French and Spanish evaluators may have had more to do with the cultural tendencies or professional backgrounds 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.

[11] With p<.001.

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

[13] While the evaluators within languages were generally very homogenous in their ratings, 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.

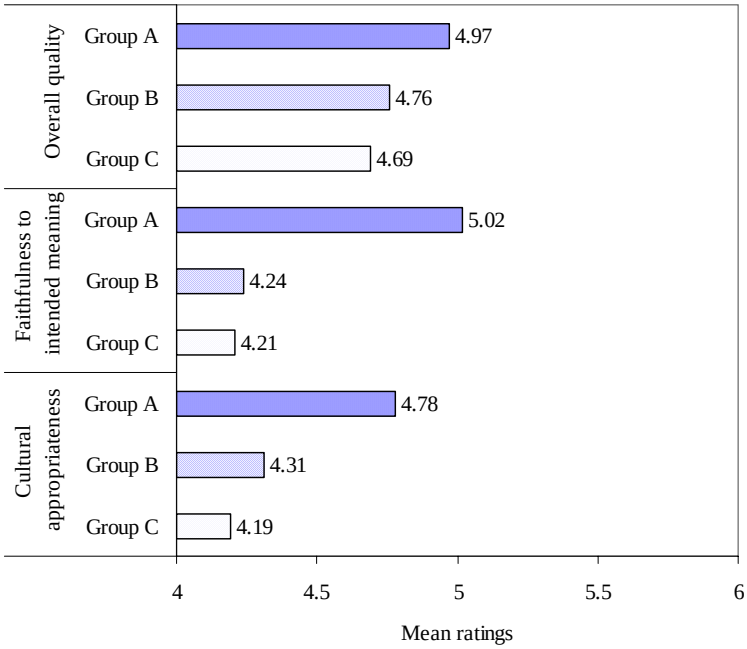

Figure 3. Mean ratings for Spanish translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup. Group A translators received only the core instructions. Group B translators received the core instructions plus QxQs. Group C translators received the core instructions plus QxQs plus the cultural appropriateness instruction.

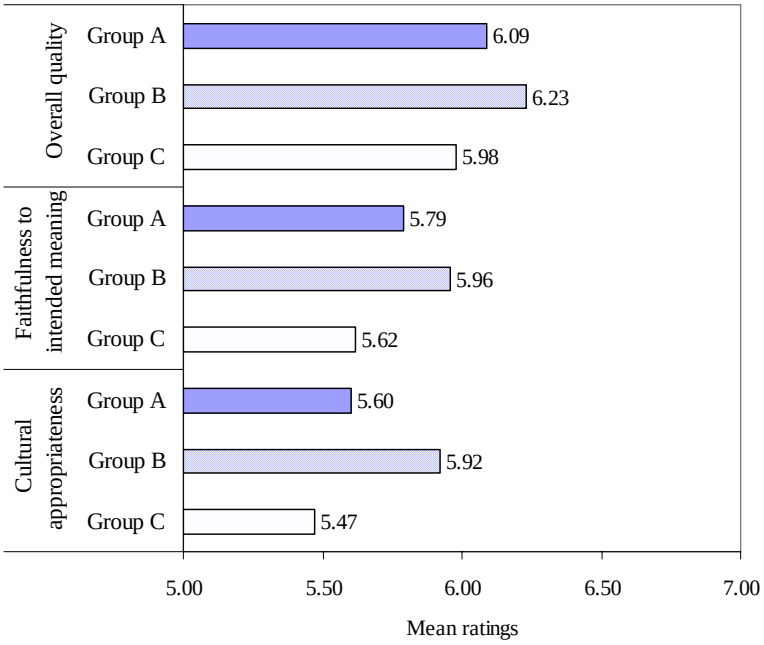

Figure 4. Mean ratings for Chinese translations for overall quality, faithfulness to intended meaning, and cultural appropriateness, by instructional subgroup. Group A translators received only the core instructions. Group B translators received the core instructions plus QxQs. Group C translators received the core instructions plus QxQs plus the cultural appropriateness instruction.

In sum, study findings suggest on the surface that providing QxQs may be effective for some languages, but not for others. However, the average ratings across languages were more likely 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.[14] 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.

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. Certainly, an equivalent stimulus does not necessarily equate to an identical *effect* on a respondent. 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. 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

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. For example, different versions of translations can be adminstered to randomly sampled target populations, and the aggregated estimates can be compared.

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 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.

References

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

Harkness, J. A., & Schoua-Glusberg, A. (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., et al. (1999). Preparing instruments for transcultural research: use of the translation monitoring form with Nepali-speaking Bhutanese refugees. *Transcultural Psychiatry*, *36*(3), 285-301.

[14] One reviewer noted that since the three translated versions of items were placed on the same page, evaluators may have been *ranking* rather than *rating* them. Clearly, comparisons across the versions may have played a role in the assignment of ratings, but we do not believe that this process compromsed the evaluation of the translations along the different dimensions.