Constructing Cross-National Surveys

Tom W. Smith

National Opinion Research Center
University of Chicago

January, 2002

GSS Cross-National Report No. 22

This report was prepared for the General Social Survey project directed by James A. Davis, Tom W. Smith, and Peter V. Marsden. The project is supported by the National Science Foundation, SES-0094979.

Introduction

As challenging as developing questions, scales, and entire questionnaires within a monocultural context is, the task becomes considerably more difficult when done in a multi-cultural setting. Overlayering the standard need to create reliable and valid measures are the complications inherent in cross-cultural and cross-national differences in language, customs, and structure. Only by dealing with these challenges on top of the usual instrument design issues can scientifically credible cross-national survey instruments emerge.

The basic goal of cross-national survey research is to construct a questionnaire that is functionally equivalent across the target populations.[1] Questions need not only be valid, but must have comparable validity across nations. But of course the very differences in language, culture, and structure that make cross-national research so analytically valuable, seriously hinder achieving equivalency of measurement. On one hand, the difficulty of establishing comparability is widely acknowledged. But on the other hand, the challenge is more often ignored than met.

A comprehensive review of major books and articles using cross-national research (Bollen, Entwisle, and Alderson, 1994) found:

> Major measurement problems are expected in macrocomparative research, but if one were to judge by current practices, one might be led to a different conclusion. Issues surrounding measurement are usually overlooked....Roughly three quarters of the books and articles do not consider alternative measures or multiple indicators of constructs, whether measures are equally valid in the different countries, or the reliability of measures.

Additionally, among the less than a quarter of comparative research that discussed whether "measures are equally valid in different countries...," sometimes this consisted only of "a passing reference" to the equal validity. Furthermore, they observed that "Few authors...discuss reliability (26%), which should temper our confidence in the quality of measurement and in the results....[and] just 6% of the books and 14% of the articles report the reliability of their measures."

Considering the value of cross-national research, the importance of obtaining comparable measurements, and the frequent failure to take measurement seriously, there is an obvious need for general improvement. This chapter contributes towards that goal by discussing 1) the development of equivalent questions in surveys, focusing on a) the question-asking and b) answer-recording parts, 2) response effects that contribute to measurement error in general and variable error structures across nations, considering in particular social desirability, acquiescence bias, extreme response

---

[1]On the different types of equivalent see Johnson, 1998 and Knoop, 1979.

styles, Don't Knows (DKs) and non-attitudes, neutral and middle options, response order, question order, and mode of administration, and 3) steps to enhance validity and comparability in cross-national surveys, including the form of source questions, translation procedures, and item development and pretesting.

Question Wordings

Question wordings and their translation are "the weakest link" in achieving crossnational equivalence (Kumata and Schramm, 1956) and therefore valid comparisons. Questions have two parts to them, the body of the item where the substance and the stimulus are presented and the response scale where the answers are recorded. Both parts will be examined in turn.

Question-Asking Part

First, one needs to consider substantive meaning and conceptual focus of the question. Here the challenge is to achieve linguistic equivalence across versions of the questionnaire. The first hurtle is to come up with a "good" translation. That means a translation in which the optimum words are used to cover the same concepts as in the original, source version (or in the desirable situation in which the two+ versions are being simultaneously developed that words used in each language are the closest possible matches).[2] However, as important and difficult as this is, it is only the first part of the challenge. For the best possible translation (i.e. the closest possible matching of terms) may not produce equivalency. The problem is not a bad or incorrect translation, but differences intrinsic in the languages. For example, "mental health" in English is translated into "jingshen jiankang" or "xinli jiankang" in Chinese which might more literally be understood as respectively "spiritual health" and "psychological health". In particular, the first Chinese term contains an element of meaning that is essentially absent from the English.

Similarly, consider the use of the English and French terms "liberty" and "libertie/" in the United States and France . In both languages the terms have strong historical ties and are closely associated with each country's formative revolutions (e.g. respectively in such phrases as "life, liberty, and the pursuit of happiness" and "Libertie/! Egalite/! Fraternite/!"). In other languages, they would often translate into terms with no strong, historical and revolutionary connotations.

Finally, even cognates between fairly closely related languages can have substantial differences. For example, for Spanish-speaking immigrants in the United States, "educacion [accent over o]" includes social skills of proper behavior that are essentially missing from the more academic meaning of "education" in English (Greenfield, 1997).

---

[2]See later section on how best to actually do translations.

A related problem is when a concept is easily represented by a word in one language and no word corresponds in another language. For example, a study of Turkish peasants (Frey, 1963) concluded that "there was no nationally understood word, familiar to all peasants, for such concepts as 'problem,' 'prestige,' and 'loyalty'..." Similarly, the Japanese concept of "giri" [having to do with duty, honor, and social obligation] has no "linguistic, operational, or conceptual corollary in Western cultures (Sasaki, 1995)."

It is not only purely language differences that hinder the achievement of functional equivalence. Differences in contemporary conditions and existing structures present problems. First, the different current situation can interact with words that may have equivalent literal meaning to produce questions with different social implications. As Bollen, et al. (1993) note:

> Consider the young woman who has reached her family size goal. In the United states, if you ask such a woman whether it would be a problem if she were to get pregnant, she is likely to say yes. In Costa Rica, she may say no. This is because in Costa Rica, such a question may be perceived as a veiled inquiry about the likely use of abortion rather than a measure of commitment to a family size goal.

Also, structural differences mean that equivalent objects may not exist or that terms used to describe one object in one country describe something else in another country. For example, one can ask about approval of the nation's monarch in Great Britain and the Netherlands, but not in France and Germany. Or to make the distinction clearer, linguistically one could ask about approval of a monarch in all four languages, but due to structural differences in the political systems, such an item would make sense in only Britain and the Netherlands. Likewise, a major American welfare program, the "food stamp program" - which gives qualifying people script that can be used to purchase certain food stuffs at regular commercial stores, has no close equivalent in most other countries. In other cases, questions must ask not about the literal translation, but the functionally equivalent object. For example, most questions asking about the American president would inquire about the German chancellor and the Israeli prime minister and not the German or Israeli president.

Variations in conditions and structures mean that what one asks about and how one asks about objects will differ across societies. This applies to items on concrete behaviors and demographics as well as to attitudinal and psychological measures. For example, one study involving occupations in rural Mali found that to the standard American occupational classifications of how each job relates to data, people, and things there had to be added a fourth dimension of relating to animals (Schooler, et al., 1998). Similarly, items about spouses have to allow for multiple mates in Islamic and most African societies.

Basic demographics can be among the most problematic of

variables. For a few variables, like age and gender, there are relatively simple ways to ask essentially equivalent questions in most societies. But then there are another set of demographics in which the items (both questions and answers) must use country-specific terms. For example, region or area of residence will use units particular to each country in the question (e.g. "states" in the United states, "provinces," in Canada, and "laender" in Germany) and of course the answers will be unique geographic localities (e.g. Indiana, Nova Scotia, and Brandenburg). Likewise, voting and party preference must refer to country-specific candidates and political parties.

Then there are a middle set of items that might be asked in either country-specific or generic, cross-country manners. For example, an generic approach to measuring education might ask something like "How many years of schooling have you completed?" A country-specific approach might ask about the highest degree obtained, the type of school attended, and/or the type of examination passed. The International Social Survey Program (ISSP), for example, follows the latter course, judging that getting precise country-specific information on level and type of education is important. The former produces a simple and superficially equivalent measure, but lumps together people who have been educated in completely different educational tracks within a country. But the later has to struggle with how of compare the unique, country-specific, educational categories across nations.

Another example would be a choice between country-specific occupational classification systems such as the 1990 US Census Classification of Occupations and corresponding national systems in each other country and international systems such as the 1988 International Standard Classification of Occupations.

With such problems of linguistic and structural equivalence added on top of the already notable monolingual challenge to creating valid measures, the need for the standard call for multiple indicators is greatly reenforced. Even with the most careful of translations, it is difficult to compare the distributions of two questions that employ abstract concepts and subjective response categories (Grunert and Muller, 1996 and Smith, 1988). While it is probably possible to ask effectively equivalent questions like "In what year were you born?" and "Did you vote in the last national election?", it is highly doubtful that the response to the query "Are you very happy, pretty happy, or not too happy?" are precisely comparable across languages. In all likelihood the closest linguistic equivalent to "happy" will differ from the English concept in various ways, perhaps conveying different connotations and tapping other related dimensions (e.g. satisfaction), but at a minimum probably expressing a different level of intensity (say where happiness itself would fall on a scale from absolute bliss to total despair). Similarly, the adjectives "very," "pretty," and "not too" are unlikely to have precise equivalents. Even in the situation in which the English adjective "very" is consistently (and correctly) translated into the French "tres", it is not known if their strength is

sufficiently identical to cut the underlying continuum of happiness as the same point.

To illustrate the added need for multiple indicators in a cross-national context, consider an example of a scheme to assess whether the French or Americans have greater psychological well-being:

A. A measure of general happiness
B. A measure of overall satisfaction
C. A scale of measures of domain-specific satisfaction

Franco-American comparisons on any one of these questions or scales would be suspect because of possible language ambiguities. Even the multi-item measure of domain-specific satisfaction would not be sufficient since all items would be built around the shared and repeated use of "satisfaction" which means that any non-equivalence is compounded across items since the error is correlated. Nor would the combination of the domain-specific and overall satisfaction contribute much to solving the problem since any disparity in the meaning of "satisfaction" in the languages would merely be perpetuated. However, switching to asking about how happy/heureux one is adds a question that is distinct from the satisfaction item and avoids obvious problems of correlated, linguistic error from repeated terms. Similarly, the use of the 10-item Bradburn affect-balance scale would have this same advantage since it asks about how often respondents have experienced five positive and five negative emotions using largely different terminology.[3]

If linguistically distinct measures are used, then it is possible to get unambiguous results if the results across items are consistent (e.g. the French leading or trailing the Americans on all measures). With one measure it is impossible to know if any measured differences (or even a measured non-difference) is societal or merely linguistic. With two measures a consistent pattern on both items establishes a clear finding, but if the measures disagree it is possible that one is social and the other linguistic and there is no basis to identify which is which. What is desirable is three linguistically-distinct measures of the same construct.[4] If all three agree, one has a clear and robust finding.

---

[3]MacIntosh (1998) argues the Bradburn affect-balance scale as used in the World Value Study was not comparable across nations, either because of different emotional structures or translation problems. However, the point here is that the Bradburn scale would not replicate measurement error associated with the format or terminology of the other psychological scales.

[4]This does not refer to three, single-item measures, but three linguistically distinct items or scales. In this example, the Bradburn scale has 10 items and domain-specific satisfaction measures usually cover many different areas (e.g. job, finances, family, health, education, etc.).

If two agree and the third shows a different pattern, one has to be more cautious with the results, but there is at least a "preponderance of evidence" towards one substantive interpretation of the cross-national differences. If all three results disagree (positive, negative, and no difference), then one clearly has no firm evidence about cross-national differences and much further work is needed to clarify concepts, improve items and translations, and refield investigations. A similar approach is called "triangulation" (Van de Vijver and Leung, 1997).[5]

Since country and language are totally confounded in most cross-national comparisons, additional leverage is needed to separate out the language differences from country differences (which is what one is trying to measure). That leverage comes by using multiple measures that vary the words used in specific questions. This approach would not help if the linguistic differences were persuasive across languages or if by chance the different terms happened to replicate the same distinctions. But they create a prima facie basis for believing that linguistic artifacts have been minimized, a presumption that can not be made with a single item or set of items using the same key terms.

Trans-language comparisons add to the burden of creating valid measures of constructs and inevitably mean that more items are needed to achieve the same degree of validity across languages than within a language. As Jacobson, Kumata, and Gullahorn (1960) have noted

> However difficult it may be to deal with theoretical issues concerning psychological processes which intervene between observable stimuli and responses in intracultural studies, the cross-cultural research situation magnifies these problems and adds new ones.

Although the general advantages of using multiple indicators are expanded in comparative research, Bollen, Entwisle, and Alderson (1994) found that "multiple indicators appear in only a small minority of the books (18%)...[and] in a similarly modest percentage of journal articles (26%)."

Answer-Recording Part

Equally as important as establishing the equivalency of the concepts and substance in questions is achieving equivalency in the response categories. Several solutions have been offered to increase the equivalency between questions (and ultimately the answers to the questions) in cross-national research.

Non-Verbal Scales

---

[5]On the merits of multiple indicators from cross-national research see Przeworski and Teune, 1966; Jowell, 1998; and Scheuch, 1987.

Numerical or other non-verbal scales are advocated by some (Fowler, 1993). These include such numerical instruments as ratio-level, magnitude measurement scales, 10-point scalometers, feeling thermometers, and frequency counts. Non-numerical, non-verbal scales would include such instruments as ladders, truncated pyramids or stepped-mountains, and figures or symbols often used in psychological tests. Numerical scales are assumed to reduce problems by providing a universally understood set of categories that have precise and similar meanings (e.g. 1, 2, 3 or 2:1, 10:1) and that there is no need to come up with language labels to try and denote the intensity of each response category. Similarly, it is argued that visual questions and response scales using images reduce verbal complexity.

However, non-verbal approaches have their own problems. First, many of the numerical scales are more complex and more difficult for people to use than simple, verbal items. Furthermore, variations in comprehension across countries could easily increase non-comparability. For example, the magnitude measurement method assigns a base value to a reference object and other objects are evaluated or rated by assigning values to them that reflect their ratio to the fixed item (Lodge, et al., 1975, 1976, 1976, 1979, 1981, 1981, 1982; Hougland, Johnson, and Wolf, 1992). For example, robbing a store of merchandise worth $1000 may be selected as the reference crime and assigned a seriousness score of 100. People then are asked to rate the seriousness of other crimes (e.g. jaywalking and homicide) and to rate the seriousness of these as a ratio to the base of 100 pre-assigned to a store robbery. Or alternatively seriousness might be rated by showing a line and asking people to draw shorter or longer lines to express the relative seriousness of other crimes. Sometimes people are asked to use two rating tasks (e.g. both numbers and lines) to rate the subject being studied. A serious problem with this technique is that typically (in the United States) 10-15% of people are confused by this complex, demanding task and are unable to supply meaningful responses. Moreover, it is likely that the level of confusion would vary across countries, perhaps co-varying with levels of numeracy.

Second, numerical scales are not as invariant in meaning and free of error as their simple, straight forward, mathematical nature presupposes. Schwarz and Hippler (1995) have shown that people rate objects quite differently on 10-point scales going from 1 to 10 than on the scalometers going from -5 to -1 and +1 to +5. Also, on the 10-point scalometer, people routinely misunderstand what the mid-point of the scale is (Smith, 1994). In another example, the 101-point feeling thermometer is not actually used as such a refined measurement tool. It is rare for more than about 10-20 values to be chosen by any respondents (mostly 10s and 25s/75s) and some people seem to be influenced by the temperature analogy

and avoid very high ratings as "too hot."[6] Furthermore, no research establishes whether these patterns in the use of numerical scale are consistent across nations.

Third, most societies have lucky and unlucky numbers (e.g. notice how many hotels have no 13th floor). These may influence numerical responses and since the lucky and unlucky numbers vary across societies, the effects will be variable as well.

Fourth, numerical scales, only reduce the use of words as part of the response scale and do not eliminate them. For example, on the 10-point scalometer the dimension on which the objects are being rated (usually liking and disliking) and the operation of the scale itself have to be explained. Likewise, the verbal description of the feeling thermometer scale is quite long.

Finally, alternative numbering or grouping schemes that influence the reporting of frequencies exist. Respondents are often unable (or at least unwilling) to try to provide an exact count of some behavior or possession and will round off in various ways (Tourangeau, Rips, and Rasinski, 2000). As in the feeling thermometer example above, people often round-off to tens (or hundreds, etc.). But in other case they may round of to another number such as 12 reflecting the unit of a dozen or based on the number of months in a year. Again there is no assurance that the cultural practices behind the numerical favorings will be the same across societies.

Related problems occur with non-verbal, non-numerical questions and scales. Visual stimuli are not necessarily equivalent across cultures (Tanzer, Gittler, and Ellis, 1995). For example, the color called "orange" in English is not clearly coded or distinctly labeled in Navajo. Because of this the Navajo do more poorly in matching objects that are "orange" (Jacobson, Kumata, and Gullahorn, 1960). Likewise, the physical ordering of images makes a difference. In Western designed matrix items used in psychological testing the missing element is placed in the bottom right corner (Tanzer, forthcoming)(See Figure 4.1A). This makes perfect sense for people using languages that run from left-to-right and top-to-bottom. However, the matrix is wrongly oriented for Arab respondents who read right-to-left and top-to-bottom. For them the matrix needs to be arranged with the missing element in the lower, left corner (Figure 4.1B).

Similarly, in some low-literacy, multi-lingual African societies, voting is not done with written ballots, but by placing ones vote in different ballot boxes marked by party symbols and sometimes the picture of the party leader. Africans used to this system of voting would presumably be able to follow a self-completion questionnaire item that listed party choices using similar images (and probably unable to successfully complete a

---

[6]On feeling thermometers see Wilcox, Sigelman, and Cook, 1989 and Tourangeau, Rips, and Rasinski, 2000. On the possibility that local climate might be a factor in cross-national research (and possibly interact with the feeling thermometer) see Doob, 1968.

questionnaire that only wrote out the names of parties and candidates). But those from developed countries might well have trouble following what would be for them the unusual pictorial format.

Finally, visual stimuli must be accurately replicated across countries. The 1987 ISSP study on social inequality included a measure of subjective social stratification:

> In our society there are groups which tend to be towards the top and groups which tend to be towards the bottom. Below is a scale that runs from top to bottom. Where would you place yourself on this scale.

There were 10 response categories with 1=Top and 10=Bottom. This item was asked in nine countries (Australia, Austria, Germany (West), Great Britain, Hungary, Italy, the Netherlands, Switzerland, and the United States). All countries showed a majority placing themselves towards the middle (4-7), but the Netherlands clearly was an outlier. The range in the % placing themselves in the middle was 24.0 percentage points from 83.8% in Australia to 59.8% in the Netherlands. Over half the overall difference (12.4 percentage points) was due to the Netherlands. Likewise, at the bottom (8-10) the range was 31.3 percentage points with the Netherlands contributing almost half (13.6 percentage points). While most of the other differences appeared to reflect actual differences in social structure, the Netherlands' distinctive distribution did not fit other measures of Dutch society (e.g. income distributions), nor was the Netherlands so distinctive on other social inequality measures (e.g. subjective class identification) (Smith, 1990).

Translation error was a likely suspect for the Dutch deviation, but a check of the Dutch wording indicated it was equivalent to the English in meaning and appropriate and clear in Dutch. It was then discovered that the visually displayed scale in the Netherlands differed from that employed in the other countries. The intended scale was to have 10 vertically stacked squares (with the highest box labelled "Top" and the lowest labelled "Bottom"). The Dutch scale had 10 stacked boxes, but they were in the shape of a truncated pyramid, with the bottom boxes wider than those in the middle and top. Dutch respondents were apparently attracted to the lower boxes because they were wider and were probably seen as indicating where more people were. This impact of the different arrangement of boxes was latter verified by experiments (Schwarz, Grayson, and Knaeuper, 1998).

Simple Response Scales

A second suggested solution, in a sense the opposite of the numerical approach, is the "keep-it-simple-stupid" approach. For surveys this would mean only using dichotomies. It is argued that yes/no, favor/oppose, and other pairs of antonyms have similar meanings and cutting points across languages. The argument is that

it may be difficult to determine, because of language differences, just where someone is along a continuum, but relatively easy to measure on which side of a mid-point someone is.

But the assumption that dichotomies are simple and equivalent across societies is questionable. Take a legal example. Under Anglo-American law a person may be judged as "guilty" or "not guilty" in a criminal case or "liable" or "not liable" in a civil case. While both are dichotomies, they differ greatly in where the tipping point is. In criminal cases the standard of proof is "beyond a reasonable doubt" and in civil cases it is "by the preponderance of the evidence." In rough probability terms this means a person would be liable in a civil case merely if most (the bare majority) of the evidence points towards that end, but in a criminal case a guilty verdict rests on the evidence overwhelmingly (almost all of it) pointing against the defendant.[7] Thus, all simple dichotomies are not the same with a similarly located tipping point.

Moreover, extending this legal example, the guilty/not guilty dichotomy recognized by English law is itself not followed even by the Scots. Scotland recognizes three verdicts: guilty, not proven, and innocent. Thus, the idea that guilt and innocence is a simple dichotomy does not hold up even within Great Britain.[8] Other simple, dichotomous distinctions may not hold up across languages and/or cultures.

Another drawback of this approach is its loss of precision. Dichotomies can only measure direction of attitudes and not their extremity and they are likely to create skewed distributions. Moreover, it would obviously take several questions using dichotomies to differentiate respondents as well as one item that had 5 or 7 scale points. (Of course, as noted above, on other statistical grounds a single item with multiple responses would usually not produce as valid or reliable a measure as a multiple-item, composite scale.)

Calibrating Response Scales

A third proposed solution is to calibrate the response scales by measuring and standardizing the strength of the verbal labels

---

[7]And following this distinction in the burden of proof, in a criminal jury trial the verdict must be unanimous (typically all 12 jurors voting for guilty or not guilty - with a non-unanimous decision there is a "hung jury" and a mistrial), but in many areas civil cases can be decided with less than unanimity.

[8]If you are unfamiliar with British legal practices, then this example may not be a useful as it is intended to be. But in that case, it serves as a different, equally important example, how institutions, such as the legal system, differs across nations and that such structural and conceptual differences are obstacles to establishing equivalence in cross-national research.

used. There are several ways to measure the strength of response categories along an underlying response continuum. One procedure has respondents rate the strength of terms defining each as a point on the continuum. There are three standard variants of this approach.

First, one can rank the terms from weaker to stronger (or from less to more or along any similar continuum) (Spector, 1976). This of course only indicates their relative position and not the absolute strength or distance between terms.

Second, one can rate each term on a numerical scale (usually with 10 to 21 points)(Wildt and Mazis, 1978; Worcester and Burns, 1975; Myers and Warner, 1968; Cliff, 1959; Jones and Thurstone, 1955; Mosier, 1941; Vidali, 1975; Mittelstaedt, 1971; Bartram and Yelding, 1973; Traenkle, 1987). This measures the absolute strength or distance between each term and thus facilitates the creation of equal, interval scales. An alphabetical scale or unlabelled spaces, rungs, or boxes as in a semantic differential scale (Osgood, Suci, Tannenbaum, 1957) can also be used. The letters or spaces are then transformed into their numerical equivalents (assuming equal intervals between boxes, rungs, etc.).

Finally, magnitude measurement techniques can be used to place each term on a ratio scale (Lodge, et al., 1975, 1976, 1976, 1979, 1981, 1981, 1982; Hougland, Johnson, and Wolf, 1992). As explained earlier, the magnitude measure techniques assign an arbitrary value to a reference term and then respondents rate other terms as ratios to the base term. This allows more precision than the numerical scale approach (since the terms are not constrained by the artificial limits of the bounded number scale).

Among the three variants the middle appears most useful. The ranking method fails to provide the numerical precision that is necessary to calibrate terms across languages. The magnitude measurement technique does this, but is much more difficult to administer and much harder for respondents to do (about 10-15% seem unable to master the procedure). In addition, the extra precision that the magnitude measurement procedure can provide over that achievable using a 21-point scale approach does not appear to be useful.

The direct-rating approach has been used to rate words along various dimensions. Of most interest are those that either rate terms along a general good/bad or positive/negative dimension or which rate the intensity of modifiers (Wildt and Mazis, 1978; Worcester and Burns, 1975; Myers and Warner, 1968; Cliff, 1959; Jones and Thurstone, 1955; Mosier, 1941; Vidali, 1975; Mittelstaedt, 1971; Bartram and Yelding, 1973; Lodge, et al, 1975, 1976, 1979, 1981, 1982; Hougland, Johnson, and Wolf, 1992). Similarly, other studies have rated probability statements (Wallsten, Budescu, Rapoport, Zwick, and Forsyth, 1986; Lichtenstein and Newman, 1967); frequency terms (Spector, 1976; Schaeffer, 1991; O'Muircheartaigh, Gaskell, and Wright, 1993; Strahan and Gerbasi, 1973, Bradburn and Sudman, 1979; Schriesheim and Schriesheim, 1974; Hakel, 1968; Simpson, 1944); and terms used to describe percentages from public opinion surveys (Crespi, 1981

and "RAC...," 1984).

These studies usually show that a) respondents (most often college students) can perform the required ratings tasks,[9] b) ratings and rankings are highly similar across different studies and populations, c) high test/retest reliability is achieved, and d) several different treatments or variations in rating procedures yield comparable results. Thus, the general technique seems robust and reliable.[10]

Another approach for assessing the intensity of scale terms and response categories is to measure the distributions generated by using different response scales (Smith, 1979; MacKuen and Turner, 1984; Laumann, Gagnon, Michael, and Michaels, 1994; Hougland, Johnson, and Wolf, 1992; Orren, 1978; Sigelman, 1990). In an experimental, across-subjects design, one random group is asked to evaluate an object (e.g. presidential popularity or one's personal happiness) with one set of response categories and a second random group evaluates the same object with another set of response categories. Since the stimulus is constant and the assignment is randomized, the number of people attracted to each category will depend on the absolute location of each response category on the underlying continuum and the relative position of each of the scale points adopted. With some modelling around what the two observed distributions suggest are the underlying distribution, one calculates at what point each term is cutting the underlying scale (Clogg, 1982; 1984).

In a within-subjects variation people answer the same question (i.e. presented with the same stimulus) two (or more) times with different response categories being used (Orren, 1978). This differs from a test/retest reliability design in that a) the measurement instrument is not constant (since the response categories differ) and b) the two administrations are essentially consecutive without any intervening time and/or buffer tasks. This provides additional information since it allows the direct comparison of responses, but the initial evaluations may artificially influence responses to the later scales. Respondents may feel constrained to chose the same response on a subsequent administration as used on the first administration in order to

---

[9]While reassuring, other studies show that various measurement artifacts can influence responses to numerical scales (Wilcox, Sigelman, and Cook, 1989; Smith, 1994; Schwarz and Hippler, 1995; Schwarz, Hippler, Deutsch, and Strack, 1985; and Schwarz, Knaeuper, Hippler, Noelle-Neumann, and Clark, 1991). See also, O'Muircheartaigh, Gaskell, and Wright, 1993.

[10]An exception is that vague frequency terms correspond to different absolute values depending on the commonness or rarity of the specified event or behavior. Thus, people who "usually" vote vote once every year or every other year, but people who "usually" dine out dine out more than once a week (Schaeffer, 1991; Bradburn and Sudman, 1979).

appear consistent. They may select responses representing the same scale position (e.g. in the middle) or using the same term despite other significant differences in the response scales.

The advantage of the distributional approaches is that they have respondents do only what they are normally asked to do - answer substantive questions with a simple set of response categories. The disadvantages are that a) it is harder to evaluate a large number of response terms and thus is better suited for assessing a discrete response scale already adopted than for evaluating a large number of terms that might be utilized in constructing optimal response scales,[11] b) results will depend on the precise underlying distribution and the modelling procedures adopted, and c) it creates more work for the analysts since the strength of terms must be indirectly estimated from the distributions rather than directly calculated from respondent ratings.

The direct-rating approach has been used in a study of terms used in response scales in Germany and the United States (Smith, 1997 and Mohler, Smith, and Harkness, 1998) and this pilot study yield very promising results. Many response terms were shown to be highly equivalent in Germany and the US, but some notable systematic differences also appeared. Besides the many technical challenges that the approach demands, the major drawback is that separate methodological studies are needed in each country and language to establish the calibration. This obviously is something that every substantive cross-national study can not undertake. However, in theory once calibrations are determined they could be used by other studies without any extra data collection being needed. Moreover, since the same response scales are used across many different substantive questions, developing a small number of carefully calibrated response scales could be used in a large number of questions.

Considerable effort must be taken to achieve cross-national comparability in questions. On top of the most careful translations possible, care must be taken to make sure that questions actually have the same meaning across societies. Likewise, response scales should be designed to achieve maximum equivalence. In addition to minimizing differences in particular items and response scales, it is essential to employ multi-indicators that vary the particular concept words and response scales so that variables are assessed with linguistically independent measures.

Response Effects

[11]It would be possible to evaluate more terms using more random sub-groups, but in order to maintain the same level of precision this would mean increasing the sample size. Similarly, the same people could be asked many repetitions of a question with different response scales, but this would soon become tedious and later repetitions would probably be distorted by the previous administrations.

Cross-national comparability is also difficult to achieve because of differences in various response effects (Hui and Triandis, 1985; Usunier, 1999). Of course response effects contribute measurement error to all surveys. The special danger in cross-national surveys is that various error components may be correlated with nation and therefore that observed differences represent differences in response effects rather than in substance. Work by Saris (1998) across 13 cultural groups/nations indicates that measurement error is not constant. As he notes, "Even if the same method is used, one can get different results due to differences in the error structure in different countries." Among the more important of the crossnational sources of measurement variation are effects related to social desirability, acquiescence, extremity, no opinion, middle options, response order, context and order, and mode.

Social Desirability

Social desirability effects distort people's responses (DeMaio, 1984; Johnson, et al., 2000; Tourangeau, Rips, and Rasinski, 2000). Image management and self-presentation bias lead respondents to give responses that put themselves in a positive light. The general tendency is to overreport popular opinions and to underreport unpopular or deviant opinions, or even more strongly to underreport undesirable behaviors and overreport socially acceptable activities. A related tendency is to make similar over and underreports, but to gear the response not to general social norms, but to the perceived values of the interviewer.

Social desirability effects appear common across social groups, but often differ in both intensity and particulars. First of all, the pressure to conform varies. Conformist societies would presumably have larger social desirability effects than individualist societies. Moreover, social desirability effects may be related to and compounded by interactions between the characteristics of respondents and interviewers. The race and ethnicity, gender, social class, and age of respondents and interviewers have been shown to interact to alter responses. For example, in the US the largest and most consistently documented interviewer effect is that people express more tolerant inter-group attitudes when being interviewed by a person of another race/ethnicity (Schuman, Steeh, Bobo, and Krysan, 1997). A similar effect has been found in Kazakhstan involving Russian and Kazakh interviewers (Javeline, 1999). Likewise, the social desirability effects are likely to be greater when larger status/power differential exists between the interviewer and the respondent and these differentially are likely to vary across nations. In developing countries, for example, it is more likely the interviewers would be members of an educated elite, while in developed countries interviewers are typically of average status.

Second of all, what is seen as sensitive topics and undesirable behavior varies both across individuals and cultures. As Newby (1998) has noted, cultures differ by what they consider as

private and public information. Also, what is legally possible to ask varies. China now permits much survey research, but many political questions such as those asking about the Communist party are not allowed. However, in most instances, the constraints come from social conventions rather than legal regulations. Societies can differ greatly in the acceptability of various items. Items about alcohol use are much more sensitive in Islamic countries than in Judeo-Christian societies. Likewise, cohabitation was a widely accepted practice in Sweden long before it became socially acceptable in the United States.

To deal with social desirability effects one can frame questions in less threatening manners, train interviewers to be non-judgmental in asking items and responding to answers, and use modes that reduce self-presentation bias (see section on modes below).

Acquiescence Bias

Acquiescence or yea-saying bias is the tendency for respondents to try to be overly compliant and tell the interviewer what the respondent thinks she/he wants to hear (Tourangeau, Rips, and Rasinski, 2000). It is particularly likely to occur on agree/disagree and other items in which there are clear affirming and rejecting responses. Acquiescence leads people to select the affirming response. Research clearly indicates that this bias can be variable across cultures. For example, Church (1987) found yea-saying to be particularly strong in the Philippines. Landsberger and Saavedra (1967) reported similar effects among Spanish speakers in the US and Chile. Javeline (1999), using experiments with reversed coded items in Kazakhstan, found not only a high level of acquiescence overall, but an even greater level among the Kazakhs than the Russians. Van Herk (2000) showed that the Greeks gave more positive responses than those in other European countries.

Acquiescence bias can be reduced in several ways. First, scales can be balanced so that the affirming response half the time is in the direction of the construct and half the time in the opposite direction (e.g. six agree/disagree items on national pride with the patriotic response matching three agree and three disagree responses). Such reversals hopefully forces the respondent to think about the meaning the items and to reply in a substantively meaningful pattern. If this does not occur, then the answers cancel themselves out and the respondent ends up in the middle which at least is a better place for such a substantively uninvolved respondent than at one extreme where the respondent would be if the items were all scored in one direction. Second, formal reversals can be built into an instrument to catch yea-sayers (Javeline, 1999; Bradburn, 1983). Finally, alternative question formats, such as force-choice items, that are less susceptible to acquiescence bias, can be employed (Converse and Presser, 1986; Krosnick, 1999). As Javeline (1999) has observed:

(M)embers of certain ethnic groups - in the name of deference,

> hospitality, or some other cultural norm [agree falsely]...more frequently...(T)he fact that they do must be taken into account in designing questionnaires. We can not change the respondents, so we must change our methods.

Extreme Response Styles

Some people appear to be especially attracted to extreme, end categories (e.g. strongly agree, most important), while others avoid these and tend to favor less extreme responses (e.g. agree, somewhat important). People tend to follow the extreme/non-extreme patterns regardless of the true strength of attitudes towards particular items, so that the choice of categories may represent a response set, rather than a substantive, gradation of opinions. Moreover, this tendency varies across racial and ethnic groups. For example, Black students in the United States are more prone to select extreme responses than White students are (Bachman and O'Malley, 1984). Likewise, Hispanics in the US Navy used extreme categories more than non-Hispanics, although the differences showed up more clearly on a 5-point scale than on a 10-point scale (Hui and Triandis, 1989).

Differences in the propensity to select extreme categories has also been shown in cross-national studies. Asians in general and the Japanese in particular are inclined to avoid extreme responses (Chen, Lee, and Stevenson, 1995; Chun, Campbell, and Yoo, 1974; Hayashi, 1992; Lee and Green, 1991; Onodera, 1999). Whether these differences are tied to cultural differences, such as sub-group norms of cognition, or explained by structural differences in other factors related to extremity preference such as education, age, and income is not known (Greenleaf, 1992; Greenfield, 1997).

Various approaches have been proposed to deal with response styles related to extremity. First, a multi-trait, multi-measurement design can be employed (Van Deck, 2000). For example, the response scales used can be varied (as in the 5- and 10-point scales in the Navy study - Hui and Triandis, 1989) to see if effects occur across measurement instruments.

Second, some have suggested that items should be ranked rather than rated. But while this formally eliminates the possibility of an extreme response effect, it forces respondents to complete a more difficult task, loses measurement differentiation, and assumes that there are no ties across objects (Van Deck, 2000).

Third, another approach argues that linguistic equivalency may have to be sacrificed to achieve functional equivalency on the scale. Since it appears that the Japanese are predisposed to avoid response categories with strong labels, some advocate that the labels for Japanese categories be softened, so that the categories "strongly agree" and "agree" would not be literally translated into Japanese, but rendered to be equivalent to "agree" and "tend to agree" instead. Others suggest however that the problem is a disconnect between translation equivalence and response-scale equivalence. For example, Voss, Stem, Johnson, and Arce (1996) in a study of English, Chinese, and Japanese students found that a

number of terms used in typical survey responses and translated as equivalent were not rated as similar in intensity in quantitative comparisons. These results indicate that part of the problem is that equivalence (at least in terms of survey responses) is not being achieved via standard translation.

In either case the issue is whether non-comparability in one aspect is needed to establish comparability on a more important basis. The general rule that one follows is to do things exactly the "same" across surveys. The challenge is identifying cases in which things that are the "same" really are not and an adjustment is needed to establish equivalency.[12]

Finally, several steps can be done at the analysis stage. One can first check if extremity of responses is similar across countries. Also, one can conduct analyses with items collapsed into dichotomies to see if this appreciably changes conclusions. For example, in an analysis of items on scientific and environmental knowledge on the ISSP, one summary scale using the 12 items merely counted the number of correct responses, while another used the five response categories (definitely true, probably true, can't choose, probably false, and definitely false). The two scales produced similar findings and this suggested that they were robust (Smith, 1996).

DKs and Non-Attitudes

People have different propensities towards offering opinions. DKs are higher among the less educated across countries (Young, 1999), but even with education controlled for, levels of DKs vary by country. Some of this cross-national variation is undoubtedly real, reflecting true differences in the level of opinionation, but some appears to come from different response styles. As Delbanco and colleagues have suggested (1997):

> Attitudes about responding to surveys (e.g. a tendency for individuals to say they do not know an answer or to refuse to answer) may differ across countries.

For example, Americans are more likely to supply personal income information than people in many European countries (Smith, 1991a). Also, there is apparently a greater willingness of people in some countries to guess about questions on scientific knowledge rather than admit ignorance by giving a DK response (Smith, 1996). It is also commonly assumed that DKs will be higher in developing countries both because of their lower levels of education and

---

[12]An example of such an approach is in the graduate-level language examinations of the University of Chicago. The typical passage to be translated from the French is about 700 words, but the German text is about 450 words. The difference in word length is designed to achieve an equivalently difficult translation task within the same allotted amount of time.

because of different social norms (e.g. a reluctance to share private thoughts with strangers).

Considerable debate exists in survey research about whether surveys should encourage or discourage DKs. From Converse' nonattitude perspective, people tend to express opinions even when they do not have any and surveys should use full filters to try and discourage the false expression of opinions by nonattitudes holders (Smith, 1984). However, Krosnick and others argue that explicitly offering DKs does not improve data quality and probably assists satisficing and therefore favor not giving an explicit DK response (Krosnick, 1999). There is little evidence how such tendencies vary across countries.

Neutral/Middle Options

Related to the issue of no opinion responses is whether "neutral", middle options should be offered and what the impact of their inclusion or exclusion is. No-middle-option questions might ask one to "strongly agree, agree, disagree, or strongly disagree," while middle-option versions would ask one to "strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree." Research from several countries finds that by providing the ambivalents with a clear response option, the middle-option scale produces more reliable results (O'Muircheartaigh, Krosnick, Helic, 1998; Smith, 1997).

Response Order

The order of response options can influence the distribution of answers. Under some conditions respondents tend to favor the first offered response (i.e. primacy effects), while in other circumstances people lean towards the last response (i.e. recency effects). Mostly American studies find that when questions "are presented visually, respondents are likely to begin by processing the first response option presented; when the items are presented aurally, respondents are likely to begin processing the final option they heard (Tourangeau, Rips, and Rasinski, 2000, p. 305). The former leading to primacy effects and the latter to recency effects. Response effects have also been found to interact with both cognitive ability and respondent motivation (Sudman, Bradburn, and Schwarz, 1996; Krosnick, 1999). It is unknown how robust these patterns are across cultures and languages.

Question Order

Context and order effects occur when previously asked questions influence responses to later questions (Smith, 1991b). The questions fail to remain independent of each other and the prior stimuli or one's response to same joins with the stimuli from subsequent questions to affect the responses to subsequent items. Many different varieties of context and order effects depend on how the previous questions influence later items. Tourangeau and

Rasinski (1988; Tourangeau, 1999), for example, describe four steps to answering a question: 1) interpretation of question meaning and intent, 2) retrieval from memory of relevant and necessary information, 3) judgment as to how the memories relate to the question, and 4) making a suitable response selection given the offered response categories. For each of these stages, they say there may be carryover or backfire effects to the prior content of the survey. For example, a carryover effect at the interpretation stage might involve a definition supplied in a prior item being used to help understand a later question. A backfire effect at this stage might involve failing to mention events covered by a previous question because they were considered redundant.

Context effects undoubtedly operate across surveys in most countries, but experiments and detailed studies have apparently been done in only a few countries and no coordinated, cross-national, experimental studies seem to have been carried out.

Even when the effects are consistent in their structure and nature, their impact may vary across countries. For example, Schwarz (1999) has shown in Germany that the favorability rating of politicians are influenced by the ratings of political figures that precede them on a list. Asking first about a discredited and widely disliked politician, leads to a higher rating of political leaders in good standing who are asked about immediately after the "bad example". This effect would probably appear in other countries. If so, then in a cross-national survey of political leadership that asked about party leaders in some predetermined order (say head of government, leader of largest opposition party, leader of next largest party, etc.), the ratings of some subsequent figures (e.g. opposition leaders) could easily vary artificially because of true differences in the status of previously mentioned leaders (e.g. the head of government). This in turn could lead to misinterpretations about the absolute popularity of the opposing leadership across nations. Similarly, Sudman, Bradburn, and Schwarz (1996) describes a context experiment that hinges on the fact that beer and not vodka, is a popular, national drink in Germany. It is entirely possible that the opposite effect would appear in Russia due to the reverse positions of the two beverages.

Particularly likely to vary across cultures are conditional context effects. As Smith (1991b) has shown, context effects are often conditional on people's attitudes towards the preceding, context-triggering question. In the US asking first about generally popular government spending program leads people to lower their judgement that their income taxes are too high. However, this effect depends on the popularity of the spending programs. Among those in the most pro-spending group, asking the spending items first <u>lowers</u> the % saying the federal income tax they pay is too high by 25 percentage points. But for those most against government spending, asking the spending items first <u>increases</u> the % saying their taxes are too high by 7 percentage points. People with intermediate support for government spending programs showed intermediate context effects. In this and similar conditional context effects, if the conditional context effects were similar

across countries, the net effect would be similar only if the popularity of government programs was also comparable across countries.

Mode of Administration

Survey responses often differ by mode of administration (e.g. self-completion, in-person, telephone). Many mode effects have been demonstrated. Among the most consistent is that more socially undesirable or sensitive behaviors (e.g. high alcohol consumption, illegal drug use, criminal activity) are reported in self-completion modes than in interviewer-assisted modes (Hudler and Richter, 2001; Tourangeau, Rips, and Rasinski, 2000). Keeping mode constant will not automatically solve the problem since mode may not have a constant impact across countries. For example, not only would showcards with words be of little use in societies with low literacy, but the inappropriate use of these might well artificially create greater differences between the literate and illiterate segments of the population than a survey mode that did not interact with education and literacy so strongly.

In brief, various measurement effects influence survey responses. Sometimes we know that these effects can vary across sub-groups and/or countries. In other cases such variable effects are plausible, but have not been empirically demonstrated. Of course this does not mean that response effects are always or even typically different among different groups and across societies. A number of consistent results have been documented. For example, some social desirability effects have been shown to be similar in Canada, the Netherlands, and the United States (Scherpenzeel and Saris, 1997), telephone surveys produce lower quality data in the same countries (Scherpenzeel and Saris, 1997), and forbid/allow question wording variations have like effects in both Germany and the United States (Hippler and Schwarz, 1986). But variable measurement effects remain a serious concern and one that researchers must continually lookout for.

Enhancing Question Comparability

Several steps can be taken to lower, if not eliminate, the hurdles to equivalence and therefore achieve valid cross-national research. These include: 1) cross-national cooperation over the research design and questionnaire content, 2) adopting a master questionnaire using questions forms more readily suitable for translations, 3) considering a balance of emic and etic items, 4) following optimal translation procedures, 5) careful item development and pretesting, and 6) thorough documentation of all survey practices.

Make Cross-National Research Collaborative

Research imperialism or safari research in which a research team from one culture develops a project and instrument and rigidly

imposes it on other societies should be avoided. As Van de Vijver and Leung (1997) have observed:

> Many studies have been exported from the West to non-Western countries and some of the issues examined in these studies are of little relevance to non-Western cultures.

Instead a collaborative, multi-national approach should be followed (Van de Vijver and Leung, 1997; Jowell, 1998; Schooler, et al., 1998; Szalai, 1993). For example, as Sanders (1994) has written:

> One of its [the ISSP's] greatest strengths is that a country can only be incorporated in the survey if a team of researchers from that country are available...to ensure that the translation of the core questions can be achieved without significantly altering their meaning. The potential problem of cross-national variation in meaning is accordingly minimized.

Another quite different example of the same principle of joint development comes from a study of AIDS/HIV in three pre-literate tribes in rural Mali. The research team consisted of American health and African specialists, Mali health researchers, and local tribal informants. They worked together to design and carry out a survey that performed well for the three target populations being linguistically compatible and culturally appropriate (Schooler, et al., 1998).

Question Form and Content

The goal of comparability and the task of translation are also furthered by adopting certain question-wording practices that facilitate translations. A number of general, question-design rules to ease translation burden and enhance the likelihood of comparability have been proposed. Brislin (1986) in particular has 12 guidelines for making items readily translatable. In abbreviated form they are:

1. Use short simple sentences of less than 16 words. (But items can be of more than one sentence.)[13]
2. Employ the active rather than the passive voice.
3. Repeat nouns instead of using pronouns.
4. Avoid metaphors and colloquialisms.
5. Avoid the subjunctive.
6. Add sentences to provide context to key items. Reword key phrases to provide redundancy.
7. Avoid adverbs and prepositions telling 'where' or 'when'.
8. Avoid possessive forms where possible.
9. Use specific rather than general terms.

---

[13]Scherpenzel and Saris (1997) find long questions to be superior, but do not address the issue of sentence length.

10. Avoid words indicating vagueness regarding some event or thing (e.g. probably, maybe, perhaps).
11. Use wording familiar to the translators.
12. Avoid sentence with two different verbs if the verbs suggest two different actions.

Additional general rules about how to formulate questions and design questionnaires have been offered by others, but usually only within a monocultural context (e.g. Converse and Presser, 1986; Fowler, 1995; Sudman and Bradburn, 1982; Van der Zouwen, 2000).

First, do not use vague quantifiers (e.g. frequently, usually, regularly) since these have highly variable understandings both across respondents and question contexts (Bradburn cited in Miller, Slumczynski, and Schoenberg, 1981).

Second, avoid items with ambiguous or dual meanings (Tanzer, forthcoming). Tanzer (forthcoming) has noted that an anger-in item in the State-Trait Anger Expression Inventory ("I am secretly quite critical of others.") could be understood in two different ways. First of all, as it was basically understood in the United States, as indicating keeping ones anger internal (I keep my criticism of others to myself.) and second of all, as indicating the expression of anger (I talk about or criticize other people behind their backs.) Moreover, in South Tyrol German speakers were more likely to understand the item in the first sense, while Italian speakers leaned towards the latter understanding.

Third, ambiguity also emanates from complex questions with more than one key element. So-called doubled-barrelled questions are particularly problematic, such as "Do you support the admission of Poland and Slovenia into the European Union?" If one favors admitting or excluding both, the suitable response is clear. But if one favors the admission of one and opposes the other, there is no appropriate response (Fowler, 1995; van der Zouwen, 2000).

Fourth, avoid hypothetical and counter-factual questions. For example, asking people their attitudes on sexual discrimination if they were of the opposite gender. People often lack coherent thoughts about many imaged situations and may not even grasp the circumstances being described (Fowler, 1995; van der Zouwen, 2000).

Fifth, use terms that are simple and widely and similarly understood across all segments of the population. One needs to avoid technical and jargon terms and to aim for a level of usage suitable for a general audience. Moreover, one wants to minimize inter-respondent variability in the understanding of terms. Word must not only be understood, but comprehended in a similar manner (Smith, 1989). When needed, provide definitions to clarify the meaning of terms (Converse and Presser, 1986; Fowler, 1995; Tourangeau, Rips, and Rasinski, 2000).

Sixth, use clear and precise time references (Fowler, 1995). For example, "Do you fish?", might be understood to mean "Have you ever gone fishing?" or "Do you currently go fishing?". It would be better to ask something like, "Have you gone fishing during the

last 12 months?".[14]

Finally, some recommend avoiding the particularistic and using questions with a higher level of abstraction (Van Deth, 1999). As Inglehart and Carballo (1997) have argued

> If we had asked questions about nation-specific issues, the cross cultural comparability almost certainly would have broken down. In France, for example, a hot recent issue revolved around whether girls should be allowed to wear scarves over their heads in schools (a reaction against Islamic fundamentalism). This question would have had totally different meanings (or would have seemed meaningless) in many other societies. On the other hand, a question about whether religion is important in one's life is meaningful in virtually every society on earth, including those in which most people say it is not.

But other research indicates that people in general and the less educated in particular have more difficulty with abstract items than with concrete questions (Converse and Presser, 1986). Moreover, even within a culture, abstract and specific items can show quite different results. American studies find that very large majorities endorse free speech as a general right, but that many fewer Americans endorse free speech for particular suspect groups such as Communists and militarists (McClosky and Brill, 1983 and Sullivan, et al., 1982). This danger must be weighted against the problem of coming up with idiographic items.

In addition to following these general rules on the construction of items, it is useful to follow the rule that "more is better". As discussed above, multiple indicators should be used in order to both enhance scale reliability and to reduce the likelihood of linguistic artifacts.

Emic and Etic Questions

In survey research "etic" questions refer to items that have a shared meaning and equivalence across cultures and "emic" questions to items of relevance to one or some sub-set of the cultures under study. Suppose that one wanted cross-national data on political participation in general and contacting government officials in particular. In the US items on displaying bumper stickers on ones car, visiting candidate Web sites, and emailing public officials would be relevant. In most developing countries, these would be rare to meaningless. However, an item on asking a village elder to intervene with the government on ones behalf might be a major avenue of participation in many developing societies,

---

[14]On the many difficulties in asking questions with time references (e.g. telescoping, forgetting curves) and means to deal with same (e.g. bounding and dating aids) see Tourangeau, Rips, and Rasinski, 2000.

but a mode that would have little relevance in developed nations. In such circumstances solutions might include 1) using general items that covered the country-specific activities within broader items, 2) asking people in each nation both the relevant and irrelevant participation items, or 3) asking a core set of common items (e.g. voting in local and national elections, talking to friends about politics), plus separate lists of country-specific political behaviors.[15]

Using general-items is perhaps the least appropriate since the necessary loss of detail is often a heavy price to pay and general items may be too vague and sweeping.

The relevant + irrelevant approach makes sense as long as the number of low relevancy items does not become too great and they are not so irrelevant that they do not make sense or are otherwise inappropriate. For example, the ISSP has successfully used this technique in its study of global environmental change where items on personal car use were asked in all countries, even though ownership levels were quite low in a few countries.

The emic/etic approach is useful as long as the common core is adequate for direct comparisons. For example, a study of obeisance to authority in the United States and Poland had five common items plus three country-specific items in Poland and four in the US (Miller, Slumczynski, and Schoenberg, 1981). This allows both direct cross-national comparisons as well as more valid measurement of the construct within countries (and presumably better measurement of how that constructs works in models).[16]

Likewise, in the developing the Chinese Personality Assessment Inventory researcher found that several constructs that were important parts of Chinese personality did not match any dimension measured on traditional Western scales (e.g. ren quin or relationship orientation) and that to be complete these had to be added to the assessment instrument (Cheung, et al., 1996). As the test developers noted:

> illustrates the importance of a combined emic-etic approach to personality assessment in non-Western cultures...The inclusion of relatively emic constructs are needed to provide a more

---

[15]However, even the identical action, voting in the last national election may not be equivalent. In some countries voting is legally mandatory, so it is not a meaningful measure of voluntary, political activity. In other countries elections are meaningless charades, so voting is not a meaningful measure of participating in a democracy or of making a political choice.

[16]If the core items and the core plus country-specific items formed reliable scales that both showed the same basic relationships in models, then results would be clear and robust. The appearance of different patterns for the core and country-specific items would of course raise questions about cross-national validity.

comprehensive coverage of the personality dimension that are important to the local culture.

In effect, the emic/etic approach indicates that some times one needs to do things differently in order to do things the equivalently (Przeworski and Teune, 1966).

Translation Procedures

Perhaps no aspect of cross-national survey research has been less subjected to systematic, empirical investigation than translation. Certainly there have been notable thoughtful pieces on how to do cross-national survey translations (Brislin, 1970 and 1986; Harkness, 1999 and 2001; Harkness and Schoua-Glusberg, 1998 Prieto, 1992; van de Vijver and Hambleton, 1996; Werner and Campbell, 1970). But what has been lacking are rigorous experiments to test the proposed approaches comparable to Schuman and Presser's work on question wording or Tourangeau and Schwarz's work on order and context. Because of this there has been limited progress in the development of scientifically-based translation. In fact, translation is often wrongly seen as a mere technical step rather than as a central step in the scientific process of designing valid cross-national questions.

The path to optimal translation begins at the design stage. As mentioned above, cross-national instruments should be designed by multi-national teams of researchers who are sensitive to translation issues and take them into consideration during the initial design and development stages. They in general need to keep asking themselves how each concept of interest can be measured in each language and each society under study. Specifically, they should keep in mind the idea of decentering (Werber and Campbell, 1979; Harkness and Schoua-Glusberg, 1998). Decentering is the process by which questions are formulated so they are not anchored in one language, but fit equally well in all applicable languages.

Next, there are various techniques for actually carrying out translations. First, there is the no-translator, translation-on-the-fly approach under which multi-lingual interviewers do their own translations when they encounter a respondent who does not understand the source language questions. This approach obviously has no standardization and no quality control.

Second, there is the single-translator, single-translation approach. No one formally recommends this method, but in fact it is frequently used because it is quick, easy, and inexpensive.

Third, there is the back-translation technique under which 1) the items in the source language are translated to the target language by one translator, 2) then the translation is retranslated back into the source language by a second translator, 3) the researchers then compare the two source language questionnaires, and 4) when they see notable differences in the two, they work with one or both of the translators to adjust the target language of the problematic questions. This is probably the most frequently recommended translation method (Brislin, 1970 and 1986; Harkness,

1999). The decided limitation of this technique is that no assessment is directly made of the adequacy of the target language questions. A poorly worded item that successfully back translates is undetected by this approach.

Fourth, there is the parallel-translation approach under which 1) the items in the source language are translated independently by two translators into the target language, 2) then the two translations are compared, and 3) when found to differ appreciably the two translators meet with those who developed the source language questions to figure out the reason for the variant translations. This might include simple errors (i.e. poor translations) in one version or may result from ambiguity or other uncertainty in the source language that the translators are dealing with in different ways. As in back translation, this is a two-translation, two-translators approach, but with more emphasis on optimizing wording in the target language. It also can be done more quickly than back translation since the two translation can be done simultaneously rather than sequentially.

Finally, there is the committee-translation approach under which a team of translators and researchers discuss the meaning of items in the source language, possible translations in the target language, and the adequacy of the translations in the target language in terms of such matters as level of complexity and naturalness as well as meaning. This approach may use parallel translation in that different members of the team may produce independent translations of items or the team may work on a translation simultaneously and interactively. This approach maximizes interaction between translators and between translators and other members of the research team. It also places the greatest emphasis on writing good questions and not just in translating words (Harkness, 1999; Harkness and Schoua-Glusberg, 1998). As Hudler and Richter (2001) have noted, "Group discussions, focus groups, expert groups are needed to develop cross-cultural or cross-national survey instruments. These qualitative research methods can help to identify inequivalences and inappropriateness of questions..."

The experience of the ISSP is that it is important to invest heavily in careful, team translation. This procedure is almost sure to avoid simple translation errors, produces target language questions that are natural and comprehensible, and maximizes equivalence across items.

Pretesting and Related Questionnaire Development Work

While pretesting and piloting are important in monocultural surveys, their value increases greatly for cross-national survey research. Developmental work must establish that the items and explicit scales meet acceptable technical standards (e.g. of comprehension, reliability, and validity) in each country and are comparable (or of equivalent validity) across countries (Krebs and Schuessler, 1986). As Hudler and Richter (2001) have observed about cross-national research, "it is essential that the instrument is

carefully designed and analysed in a pretest."

Useful developmental and testing procedures include: 1) cognitive interviews using such protocols as think-alouds in which respondents verbalize their mental processing of questions and computer-assisted concurrent evaluations (Bolton, and Bronkhorst, 1996; Krosnick, 1999; Pruefer and Rexroth, 1996; Tourangeau, Rips, and Rasinski, 2000), 2) behavioral coding in which the interviewer-respondent exchanges are recorded (usually on audio, but sometimes with audio-video), coded in detail, and then formally analyzed (Fowler and Cannell, 1996; Pruefer and Rexroth, 1996; Krosnick, 1999), and 3) conventional pretesting, including the use of probing (Converse and Presser, 1986; Fowler, 1995; Hudler and Richter, 2001).

Presser and Blair (1994) have compared the various pretest methods, finding that each has special advantages in identifying problems in questionnaires. An example of the value of such pretesting for inter-cultural studies is the use of cognitive follow-ups by Johnson and colleagues (1997) to find out how respondents in general and especially respondents from different cultures understand items.

Survey instruments may also be tested by 1) concurrent ethnographic analysis in which the results from surveys and ethnographic studies are cross-validated (Gerber, 1999), 2) exemplar analysis in which scales are assessed by asking people to describe what types of events would represent the response options (e.g. what would be an example of someone being completely satisfied with his/her job, somewhat dissatisfied, etc.) (Ostrom and Gannon, 1996), and 3) the quantitative scaling of response terms described above (Mohler, Harkness, and Smith, 1998; Smith, 1997). Through such careful development items with maximum comparability in the meaning of questions and in the response scales can be obtained.

Documentation

Solid documentation is also essential. Jowell (1998) has observed that good documentation and "detailed methodological reports about each participating nation's procedures, methods, and success rates..." are needed. However, as Hermalin, Entwistle, and Myers (1985) have noted "maintenance and documentation are painstaking tasks for which little provision is made..." Their work with the World Fertility Surveys found that some of the surveys no longer existed and that "the documentation for surviving surveys is often confused and incomplete." While all phases of each survey from sampling to data processing need to be carefully recorded,[17]

[17]As Hudler and Richter (2001) note, "What information about a survey is necessary for secondary analysis, that one can work scientifically with already collected data? The answer to this question is: everything."

it is particularly important to include the original questionnaires used in each countries so they can be consulted to understand results (and particularly differences in results) across countries. This practice is followed by the ISSP which included copies of original instruments on its CD-ROM. Moreover, solid documentation is more than just a good practice that facilitates primarily and secondary analysis. It enhances comparability from the start by forcing researchers to be clear about what procedures are being used in each country and how comparable they are.

In bief, the chance of achieving valid cross-national research is increased when there is 1) collaboration from the earliest stages of experts from all involved countries, 2) careful developmental work and pretesting, 3) the use of appropriate translations procedures, 4) items are crafted to ease translations, and 5) complete and clear documentation of all phases of the research.

Conclusion

Surveys are complex endeavors (Sudman, Bradburn, and Schwarz, 1996). They are social encounters between respondents and researchers (either personally represented by an interviewer or impersonally via a questionnaire). Respondents must discern the nature of the interaction facing them and determine their role in the encounter. Also, they are cognitive tasks in which the respondent is asked to comprehend various terms and inquires, search his/her memory for relevant information, and formulate it into the proffered response options.

The great challenge in cross-national survey research is that languages, social conventions, cognitive abilities, and response styles all vary across societies (Fiske, et al., 1998). To obtain valid, equivalent measurement across countries and cultures, measurement error from these sources must be minimized and equalized so that valid, reliable, and consistent substantive information emerges. Achieving this is difficult. The task of obtaining cross-national comparability is so complex and challenging that more effort is needed at all stages from conceptualizing the research question to instrument development to survey analysis. But the benefits from cross-national research fully merit the extra efforts. As the Working Group on the Outlook for Comparative International Social Science Research has noted, " A range of research previously conceived of a 'domestic,' or as concerned with analytical propositions assumed invariant across national boundaries, clearly needs to be reconceptualized in the light of recent comparative/international findings." Unless a comparative perspective is adopted and successfully implemented, "models and theories will continue to be 'domestic' while the phenomena being explained are clearly not (Luce, Smelser, and Gerstein, 1989)."

References

Bachman, J.G. and O'Malley, Patrick, "Black-White Differences in Response Styles," Public Opinion Quarterly, 48 (1984), 491-509.

Bartram, Peter and Yelding, David, "The Development of an Empirical Method of Selecting Phrases Used in Verbal Rating Scales: A Report on a Recent Experiment," Journal of the Market Research Society, 15 (July, 1973), 151-156.

Bollen, Kenneth A.; Entwisle, Barbara; and Alderson, Arthur S., "Macrocomparative Research Methods," Annual Review of Sociology, 19 (1993), 321-351.

Bolton, Ruth N. and Bronkhorst, Tina M., "Questionnaire Pretesting: Computer-Assisted Coding of Concurrent Protocols," in Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, edited by Norbert Schwarz and Seymour Sudman. San Francisco: Jossey-Bass, 1996.

Bradburn, Norman M., "Response Effects," in Handbook of Survey Research, edited by Peter H. Rossi, James D. Wright, and Andy B. Anderson. New York: Academic Press, 1983.

Bradburn, Norman M. and Sudman, Seymour, Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass, 1979.

Brislin, R.W., "Back-Translation for Cross-Cultural Research," Journal of Cross-Cultural Research, 1 (1970), 185-216.

Brislin, R.W., "The Wording and Translation of Research Instruments," in Field Methods in Cross-Cultural Research, edited by W.J. Lonner and J.W. Berry. Newbury Park, CA: Sage, 1986.

Chen, C.; Lee, S-y.; and Stevenson, H.W., Response Style and Cross-cultural Comparisons of Rating Scales Among East Asian and North American Students," Psychological Science, 6 (1995), 170-175.

Cheung, F.M.; Leung, K.; Fan, R.M.; Song, W.Z.; Zhang, J.X.; and Zhang, J.P., "Development of the Chinese Personality Assessment Inventory," Journal of Cross-Cultural Psychology, 27 (1996), 181-199.

Chun, K-T.; Campbell, J.B.; and Yoo, J.H., "Extreme Response Style in Cross-cultural Research: A Reminder," Journal of Cross-Cultural Psychology, 5 (1974), 465-480.

Church, A.T., "Personality Research in a Non-Western Setting: The Philippines," Psychology Bulletin, 102 (1987), 272-292.

Cliff, Norman, "Adverbs as Multipliers," Psychological Review, 66 (January, 1959), 27-44.

Clogg, Clifford C., "Using Association Models in Sociological Research: Some Examples," American Journal of Sociology, 88 (1982), 114-134.

Clogg, Clifford C., "Some Statistical Models for Analyzing Why Surveys Disagree," in Surveying Subjective Phenomena, edited by Charles F. Turner and Elizabeth Martin. Volume 2. New York: Russell Sage, 1984.

Converse, Jean M. and Presser, Stanley, Survey Questions: Handcrafting the Standardized Questionnaire. Beverly Hills: Sage, 1986.

Crespi, Leo P., "Semantic Guidelines to Better Survey Reportage," Office of Research, International Communication Agency, Memorandum, August 11, 1981.

Delbanco, Suzanne, et al., "Public Knowledge and Perceptions about Unplanned Pregnancies in Three Countries," Family Planning Perspectives, 29 (March/April, 1997), 70-75.

DeMaio, Theresa J., "Social Desirability and Survey Measurement: A Review," in Surveying Subjective Phenomena, edited by Charles F. Turner and Elizabeth Martin. Volume 2. New York: Russell Sage, 1984.

Doob, L.W., "Tropical Weather and Attitude Surveys," Public Opinion Quarterly, 32 (1968), 423-430.

Fiske, Alan Page; Kitayama, Shinobu; Markus, Hazel Rose; and Nisbett, Richard E., "The Cultural Matrix of Social Psychology," The Handbook of Social Psychology, edited by Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey. Vol 2. Boston: McGraw Hill, 1998.

Fowler, Floyd J., Jr., Improving Survey Questions: Design and Evaluation. Thousand Oaks: Sage, 1995.

Fowler, Floyd J., Jr., Survey Research Methods. 2nd edition. Newbury Park, CA: Sage, 1993.

Fowler, Floyd J., Jr., and Cannell, Charles F., "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions," in Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, edited by Norbert Schwarz and Seymour Sudman. San Francisco: Jossey-Bass, 1996.

Frey, F. W., "Survey Peasant Attitudes in Turkey," Public Opinion

Quarterly, 27 (1963), 335-355.

Gerber, Eleanor R., "The View from Anthropology: Ethnography and the Cognitive Interview," in Cognition and Survey Research, edited by Monroe G. Sirken, et al. New York: John Wiley & Sons, 1999.

Greenfield, Patricia M., "You Can't Take It With You: Why Ability Assessments Don't Cross Cultures," American Psychologist, 52 (Oct., 1997), 1115-1124.

Greenleaf, Eric A., "Measuring Extreme Response Style," Public Opinion Quarterly, 56 (Autumn, 1992), 328-351.

Grunert, Suzanne C. and Muller, Thomas E., "Measuring Values in International Settings: Are Respondents Thinking 'Real' Life or "Ideal' Life," Journal of International Consumer Marketing, 8 (1996), 169-185.

Hakel, Milton D., "How Often is Often?" American Psychologist, 23 (July, 1968), 533-534.

Harkness, Janet A., "In Pursuit of Quality: Issues for Cross-National Survey Research," International Journal of Social Research Methodology, 2 (1999), 125-140.

Harkness, Janet A., "Questionnaire Development, Adaption, and Assessment for the ESS," Paper presented to the International Conference on Quality in Official Statistics, May, 2001.

Harkness, Janet A. and Schoua-Glusberg, Alisu, "Questionnaires in Translation," in Cross-Cultural Survey Equivalence, edited by Janet Harkness. ZUMA-Nachrichten Spezial No. 3. Mannheim: ZUMA, 1998.

Hayashi, Echikio, "Belief Systems, the Way of Thinking, and Sentiments of Five Nations," Behaviormetrica, 19 (1992), 127-170.

Hermalin, Albert I.; Entwisle, Barbara; and Myers, Lora G., "Some Lessons from the Attempt to Retrieve Early KAP and Fertility Surveys," Population Index, 51 (Summer, 1985), 194-208.

Hippler, Hans-Jurgen and Schwarz, Norbert, "Not Forbidding Isn't Allowing: The Cognitive Basis of the Forbid-Allow Asymmetry," Public Opinion Quarterly, 50 (Spring, 1986), 87-96.

Hougland, James G.; Johnson, Timothy P.; and Wolf, James G., "A Fairly Common Ambiguity: Comparing Rating and Approval Measures of Public Opinion," Sociological Focus, 25 (August, 1992), 257-271.

Hudler, Michaela and Richter, Rudolf, "Theoretical and Methodological Concepts for Future Research and Documentation on Social Reporting in Cross-sectional Surveys," EuReporting Working Paper No. 18. Paul Lazarsfeld-Gessellschaft fuer Sozialforschung, Vienna, 2001.

Hui, C.H. and Triandis, H.C., "Effects of Culture and Response Format on Extreme Response Style," Journal of Cross-Cultural Psychology, 20 (1989), 296-309.

Hui, C. Henry and Triandis, Harry C., "The Instability of Response Sets," Public Opinion Quarterly, 49 (Summer, 1985), 253-260.

Inglehart, Ronald and Carballo, Maria, "Does Latin America Exist> (And is there a Confucian Culture?): A Global Analysis of Cross-Cultural Differences," PS, 30 (March, 1997), 34-46.

Javeline, Debra, "Response Effects in Polite Cultures: A Test of Acquiescence in Kazakhstan," Public Opinion Quarterly, 63 (Spring, 1999), 1-28.

Jacobson, E.; Kumata, H.; and Gullahorn, J.E., "Cross-Cultural Contributions to Attitude Research," Public Opinion Quarterly. 24 (1960), 205-223.

Johnson, Timothy P., "Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research," in Nachrichten Spezial, Cross-Cultural Survey Equivalence, No. 3, edited by Janet A. Harkness, 1998.

Johnson, Timothy P., et al., "Social Cognition and Responses to Survey Questions Among Culturally Diverse Populations," in Survey Measurement and Process Control, edited by Lars Lyberg, et al. New York: John Wiley & Sons, 1997.

Johnson, Timothy P., et al., "The Effects of Cultural Orientations on Survey Response: The Case of Individualism and Collectivism," Paper presented to the International Conference on Logic and Methodology, Cologne, 2000.

Jones, Lyle V. and Thurstone, L.L., "The Psychophysics of Semantics: An Experimental Investigation," Journal of Applied Psychology, 39 (February, 1955), 31-36.

Jowell, Roger, "How Comparative is Comparative Research?" American Behavioral Scientist, 42 (Oct., 1998), 168-177.

Knoop, James C., "Assessing Equivalence of Indicators Cross-National Survey Research: Some Practical Guidelines," International Review of Sport Sociology, 14 (1979), 137-156.

Krebs, Darmar and Schyessler, Karl F., "Zur Konstruktion von

Einstellungsskalen im Internationalen Vergelich," ZUMA-Arbeitsbericht No. 86/01, 1986.

Krosnick, Jon A., "Survey Research," Annual Review of Psychology, 50 (1999), 537-567.

Kuechler, Manfred, "The Survey Method: An Indispensable Tool for Social Science Research Everywhere?" American Behavioral Scientist, 42 (Oct., 1998), 178-200.

Kumata, H. and Schramm, W., "A Pilot Study of Cross-Cultural Meaning," Public Opinion Quarterly, 20 (1956), 229-238.

Landsberger, H.A. and Saavedra, A., "Response Set in Developing Countries," Public Opinion Quarterly, 31 (1967), 214-229.

Laumann, Edward O.; Gagnon, John H.; Michael, Robert T.; and Michaels, Stuart, The Social Organization of Sexuality: Sexual Practices in the United States. Chicago: University of Chicago Press, 1994.

Lee, C. and Green, R.T., "Cross-cultural Examination of the Fishbein Behavioral Intentions Model," Journal of Business Studies, 2 (1991), 289-305.

Lichtenstein, Sarah and Newman, J. Robert, "Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities," Psychon. Sci., 9 (1967), 563-564.

Lodge, Milton, Magnitude Scaling: Quantitative Measurement of Opinions. Beverly Hills: Sage Publications, 1981.

Lodge, Milton; Cross, David; Tursky, Bernard; Tanenhaus, Joseph; and Reeder, Richard, "The Psychophysical Scaling of Political Support in the 'Real World'," Political Methodology, 3 (1976), 159-182.

Lodge, Milton; Cross, David V.; Tursky, Bernard; and Tanenhaus, Joseph, "The Psychological Scaling and Validation of a Political Support Scale," American Journal of Political Science, 19 (November, 1975), 611-649.

Lodge, Milton; Tanenhaus, Joseph; Cross, David; Tursky, Bernard; Foley, Mary Ann; and Foley, Hugh, "The Calibration and Cross-Modal Validation of Ratio Scales of Political Opinion in Survey Research," Social Science Research, 5 (1976), 325-347.

Lodge, Milton and Tursky, Bernard," Comparisons between Category and Magnitude Scaling of Political Opinion Employing SRC/CPS Items," American Political Science Review, 73 (1979), 50-66.

Lodge, Milton and Tursky, Bernard, "On the Magnitude Scaling of

Political Opinion in Survey Research," American Journal of Political Science, 25 (May, 1981), 376-419.

Lodge, Milton and Tursky, Bernard, "The Social-Psychological Scaling of Political Opinion," in Social Attitudes and Psychophysical Measurement, edited by Bernd Wegener. Hillsdale, NJ: Lawrence Erlbaum Associates, 1982.

Luce, R. Duncan; Smelser, Neil J.; and Gerstein, Dean R., eds. Leading Edges in Social and Behavioral Science. New York: Russell Sage, 1989.

MacIntosh, Randall, "A Confirmatory Factor Analysis of the Affect Balance Scale in 38 Nations: A Research Note," Social Psychology Quarterly, 61 (March, 1998), 83-91.

MacKuen, Michael B. and Turner, Charles F., "The Popularity of Presidents, 1963-1980," in Surveying Subjective Phenomena, edited by Charles F. Turner and Elizabeth Martin. Volume 2. New York: Russell Sage, 1984.

McClosky, Herbert and Brill, Alida, Dimensions of Tolerance: What Americans Believe about Civil Liberties. New York: Russell Sage, 1983.

Miller, Joanne; Slomczynski, Kazimierz M.; and Schoenberg, Ronald, "Assessing Comparability of Measurement in Cross-National Sociocultural Settings," Social Psychology Quarterly, 44 (Sept., 1981), 178-191.

Mittelstaedt, Robert A., "Semantic Properties of Selected Evaluative Adjectives: Other Evidence," Journal of Marketing Research, 8 (May, 1971), 236-237.

Mohler, Peter Ph,; Smith, Tom W.; and Harkness, Janet A., "Respondent's Ratings of Expressions from Response Scales: A Two-Country, Two-Language Investigation on Equivalence and Translation," in Nachrichten Spezial, Cross-Cultural Survey Equivalence, No. 3, (1998), edited by Janet A. Harkness.

Mosier, Charles, "A Psychometric Study of Meaning," Journal of Social Psychology, 13 (February, 1941), 123-140.

Myers, James H. and Warner, W. Gregory, "Semantic Properties of Selected Evaluation Adjectives," Journal of Marketing Research, 5 (November, 1968), 409-412.

Newby, Margaret; Amin, Sajeda; Diamond, Ian; and Naved, Ruchira T., "Survey Experience Among Women in Bangladesh," American Behavioral Scientist, 42 (Oct., 1998), 252-275.

O'Muircheartaigh, Colm A.; Gaskell, George D., and Wright, Daniel

B., "The Impact of Intensifiers," Public Opinion Quarterly, 57 (Winter, 1993), 552-565.

O'Muircheartaigh, Colm; Kronsnick, Jon A.; and Helic, Armin, "Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data," Unpublished report, December, 1998.

Onodera, Noriko, personal communication, email, 8/16/1999.

Orren, Gary R., "Presidential Popularity Ratings: Another View," Public Opinion, 1 (May/June, 1978), 35.

Osgood, Charles E.; Suci, George J.; and Tannenbaum, Percy H., The Measurement of Meaning. Urbana, IL: University of Illinois Press, 1957.

Ostrom, Thomas M., "Bipolar Survey Items: An Information Processing Perspective." in Social Information Processing and Survey Methodology, edited by Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman. New York: Springer-Verlag, 1987.

Ostrom, Thomas M., "Exemplar Generation: Assessing How Respondents Give Meaning to Rating Scales," in Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, edited by Norbert Schwarz and Seymour Sudman. San Francisco: Jossey-Bass, 1996.

Presser, Stanley and Blair, J., "Survey Pretesting: Do Different Methods Produce Different Results?" Sociological Methodology, 24 (1994), 73-104.

Prieto, A., "A Method for Translation of Instruments to Other Languages," Adult Education Quarterly, 43 (1992), 1-14.

Pruefer, Jostein and Rexroth, Margit, "Verfahren zur Evaluation von Survey-Fragen: Ein Ueberblick." ZUMA-Arbeitsbericht No. 95/5, 1996.

Przeworski, A. and Teune, H., "Equivalence in Cross-National Research," Public Opinion Quarterly, 30 (1966), 551-568.

"RAC Quantifies the Vast Majority," The Sampler from Response Analyses, 31 (March, 1984), 2.

Sanders, David, "Methodological Considerations in Comparative Cross-National Research," International Social Science Journal, 46 (Dec., 1994), 513-521.

Saris, Willem E., "The Effects of Measurement Error in Cross-Cultural Research," in Nachrichten Spezial, Cross-Cultural Survey Equivalence, No. 3, edited by Janet A. Harkness.

Sasaki, Masamichi, "Research Design of Cross-National Attitude Surveys," Behaviormetrika, 22 (Jan., 1995), 99-114.

Schaeffer, Nora Cate, "Hardly Ever or Constantly? Group Comparisons Using Vague Quantifiers," Public Opinion Quarterly, 55 (Fall, 1991), 395-423.

Scherpenzeel, Annette C. and Saris, Willem E., "The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies," Sociological Methods and Research, 25 (Feb., 1997), 341-383.

Schooler, Carmi, "Cultural and Social-Structural Explanations of Cross-National Psychological Differences," Annual Review of Sociology, 22 (1996), 323-349.

Schooler, Carmi, et al., "Conducting a Complex Sociological Survey in Rural Mali: Three Points of View," American Behavioral Scientist, 42 (Oct., 1998), 252-275.

Scheuch, Erwin A., "Theoretical Implications of Comparative Survey Research: Why the Wheel of Cross-Cultural Methodology Keeps on Being Reinvented," International Sociology, 4 (June, 1989), 147-167.

Schriesheim, Chester and Schriesheim, Janet, "Development and Empirical Verification of New Response Categories to Increase the Validity of Multiple Response Alternatives Questionnaires," Educational and Psychological Measurement, 34 (Winter, 1974), 877-884.

Schuman, Howard; Steeh, Charlotte; Bobo, Lawrence; and Krysan, Maria, Racial Attitudes in America: Trends and Interpretations, Revised Edition. Cambridge, Massachusetts: Harvard University Press, 1997.

Schwarz, Norbert, "Cognitive Research into Survey Measurement: Its Influence on Survey Methodology and Cognitive Theory," in Cognition and Survey Research, edited by Monroe G. Sirken, et al. New York: John Wiley & Sons, 1999.

Schwarz, Norbert, "Self-Reports: How the Questions Shape the Answer," American Psychologist, 54 (Feb., 1999), 93-105.

Schwarz, Norbert; Grayson, Carla E.; and Knaeuper, Baerbel, "Formal Features of Rating Scales and the Interpretation of Question Meaning," International Journal of Public Opinion Research, 10 (1998), 177-183.

Schwarz, Norbert and Hippler, Hans-Jurgen, "The Numeric Values of Rating Scales: A Comparison of Their Impact in Mail Surveys and Telephone Interviews," International Journal of Public

Opinion Research, 7 (1995), 72-74.

Schwarz, Norbert and Hippler, Hans-J., "What Response Scales May Tell Your Respondents: Informative Functions of Response Alternatives," in Social Information Processing and Survey Methodology, edited by Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman. New York: Springer-Verlag, 1987.

Schwarz, Norbert; Hippler, Hans-J.; Deutsch, Brigitte; and Strack, Fritz, "Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments," Public Opinion Quarterly, 49 (Fall, 1985), 388-395.

Schwarz, Norbert; Knaeuper, Baerbel; Parl, Denise, "Aging, Cognition, and Context Effects: How Differential Context Effects Invite Misleading Conclusions about Cohort Differences," Paper presented to the American Association for Public Opinion Research, St. Petersburg Beach, May, 1999.

Schwarz, Norbert; Knaeuper, Baerbel; Hippler, Hans-J.; Noelle-Neumann, Elisabeth; Clark, Leslie, "Rating Scales: Numeric Values May Change the Meaning of Scale Labels," Public Opinion Quarterly, 55 (1991), 570-582.

Sigelman, Lee, "Answering the 1,000,000-Person Question: The Measurement and Meaning of Presidential Popularity," Research in Micropolitics, 3 (1990), 209-226.

Simpson, Ray H., "The Specific Meanings of Certain Terms Indicating Differing Degrees of Frequency," Quarterly Journal of Speech, 30 (October, 1944), 328-330.

Smith, Tom W., "An Analysis of Missing Income Information on the General Social Survey," GSS Methodological Report No. 71. Chicago: NORC, 1991a.

Smith, Tom W., "An Analysis of Response Patterns to the Ten-Point Scalometer," American Statistical Association 1993 Proceedings of the Section on Survey Research Methods. Alexandria, VA: ASA, 1994.

Smith, Tom W., "Environmental and Scientific Knowledge Around the World," GSS Cross-National Report No. 16. Chicago: NORC, 1996.

Smith, Tom W., "Happiness: Time Trends, Seasonal Variations, Intersurvey Differences, and Other Mysteries," Social Psychology Quarterly, 42 (1979), 18-30.

Smith, Tom W., "Little Things Matter: A Sampler of How Differences in Questionnaire Format Can Affect Survey Responses," GSS Methodology Report No. 78. Chicago: NORC, 1993.

Smith, Tom W., "Improving Cross-National Survey Response by Measuring the Intensity of Response Categories," GSS Cross-National Report No. 17. Chicago: NORC, 1997.

Smith, Tom W., "Nonattitudes: A Review and Evaluation," in Surveying Subjective Phenomenon, edited by Charles F. Turner and Elizabeth Martin. New York: Russell Sage, 1984.

Smith, Tom W., "Random Probes of GSS Questions," International Journal of Public Opinion Research, 1 (1989), 305-325.

Smith, Tom W., "Thoughts on the Nature of Context Effects," in Context Effects in Social and Psychological Research, edited by Norbert Schwarz and Seymour Sudman. New York: Springer-Verlag, 1991b.

Smith, Tom W., "The Ups and Downs of Cross-National Survey Research," GSS Cross-National Report No. 8. Chicago, NORC, 1988.

Spector, Paul E., "Choosing Response Categories for Summated Rating Scales," Journal of Applied Psychology, 61 (June, 1976), 374-375.

Strahan, Robert and Gerbasi, Kathleen Carrese, " Semantic Style Variance in Personality Questionnaires,' Journal of Psychology, 85 (September, 1973), 109-118.

Sudman, Seymour and Bradburn, Norman M., Asking Questions: A Practical Guide to Questionnaire Design. San Francisco: Jossey-Bass, 1982.

Sudman, Seymour; Bradburn, Norman M.; and Schwarz, Norbert, Thinking About Answers: The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass, 1996.

Sullivan, John L.; Piereson, James E,; and Marcus, George E., Political Tolerance and American Democracy. Chicago: University of Chicago Press, 1992.

Szalai, Alexander, "The Organization and Execution of Cross-National Survey Research Projects," Historical Social Research, 18 (1993), 139-171.

Tanzer, Norbert K., "Developing Tests for Use in Multiple Languages and Cultures: A Plea for Simultaneous Development,: in Adapting Educational and Psychological Test for Cross-Cultural Assessment, edited by R. Hambleton, P. Merenda, and C.D. Spielberg. Hillsdale, NJ: Lawrence Erlbaum, forthcoming.

Tanzer, N.K.; Gittler, G.; and Ellis, B.B., "Cross-cultural Validation of Item Complexity in a LLTM-calibrated Spatial

Ability Test," European Journal of Psychology Assessment, 11 (1995), 170-183.

Traenkle, Ulrich, "Auswirkungen der Gestaltung der Antworskala suf Quantitative Urteile," Zeitschrift fuer Sozial Psychologie, 18 (1987), 88-99.

Tourangeau, Roger, "Context Effects to Answers to Attitude Questions," in Cognition and Survey Research, edited by Monroe G. Sirken, et al. New York: John Wiley & Sons, 1999.

Tourangeau, Roger and Rasinski, Kenneth A., "Cognitive Processes Underlying Context Effects in Attitude Measurement," Psychology Bulletin, 103 (1988), 299-314.

Tourangeau, Roger; and Rips, Lance J.; and Rasinski, Kenneth, The Psychology of Survey Response. Cambridge: Cambridge University Press, 2000.

Usunier, J.C., Marketing Across Cultures. 3rd edition. New York: Prentice-Hall, 1999.

van der Zouwen, Johannes, "An Assessment of the Difficulty of Questions Used in the ISSP-questionnaires, the Clarity of Their Wording, and the Comparability of the Responses," ZA-Information, No. 46 (2000), 96-114.

van Deth, Jan W., "Equivalence in Comparative Political Research," in Comparative Politics: The Problem of Equivalence, edited by Jan W. van Deth. London: Routledge, 1999.

van de Vijver, Fons and Hambleton, R. K., "Translating Tests: Some Practical Guidelines," European Psychologist, 1 (1996), 89-99.

van de Vijver, Fons and Leung, Kwork, Methods and Data Analysis for Cross-Cultural Research. Thousand Oaks, CA: Sage, 1997.

van Herk, Hester, Equivalence in a Cross-National Context: Methodological and Empirical Issues in Marketing Research. Published Ph.D. dissertation, Catholic University, Brabant, 2000.

Voss, Kevin E.; Stem, Donald E., Jr.; Johnson, Lester W.; and Arce, Constantino, "An Exploration of the Comparability of Semantic Adjectives in Three Languages: A Magnitude Estimation Approach," International Marketing Review, 13 (1996), 44-58.

Vidali, Joseph J., "Context Effects on Scales Evaluatory Adjective Meaning," Journal of the Market Research Society, 17 (January, 1975), 21-25.

Wallsten, Thomas S.; Budescu, David V.; Rapoport, Amnon; Zwick,

Rami; and Forsyth, Barbara, "Measuring the Vague Meanings of Probability Terms," Journal of Experimental Psychology, 115 (December, 1986), 348-365.

Werner, O. and Campbell, D., "Translating, Working through Interpreters, and the Problem of Decentering," in Handbook of Cultural Anthropology, edited by R. Naroll and R. Cohen. New York: American Museum of Natural History, 1970.

Wilcox, Clyde; Sigelman, Lee; and Cook, Elizabeth, "Some Like it Hot: Individual Differences in Responses to Group Feeling Thermometers," Public Opinion Quarterly, 53 (Summer, 1989), 246-257.

Wildt, Albert R. and Mazis, Michael B., "Determinants of Scale Response: Label Versus Position," Journal of Marketing Research, 15 (May, 1978), 261-267.

Wilson, E.C., "Problems of Survey Research in Modernizing Areas," Public Opinion Quarterly, 22 (1958), 230-234.

Worcester, Robert M. and Burns, Timothy R., "A Statistical Examination of the Relative Precision of Verbal Scales," Journal of the Market Research Society, 17 (July, 1975), 181-197.

Young, Clifford A., "What We Know About 'I Don't Know': An Analysis of the Relationship Between 'Don't Know' and Education," Paper presented to the American Association of Public Opinion Research, St. Petersburg Beach, May, 1999.