Chapter 24

Surveying Across Nations and Cultures

Tom W. Smith

24.1. Introduction

As Durkheim (1938, p. 139) noted in 1895, "Comparative sociology is not a particular branch of sociology; it is sociology itself, in so far as it ceases to be purely descriptive and aspires to account for facts." This of course applies to survey research and the social sciences as a whole. While comparative social science is of crucial and central importance, it is also especially difficult. As challenging as developing questions, scales, and entire questionnaires within a monocultural context is, the task is considerably more difficult in multicultural settings. Above and beyond the standard need to create reliable and valid measures are the complications inherent in cross-cultural and cross-national differences in language, culture, and structure. Only by dealing with these extra challenges can scientifically credible cross-national survey instruments emerge (Smith, 2003, 2004).

The basic goal of cross-national surveys is collecting data that are functionally equivalent across populations.[1] Measurements need not only be valid, but must have comparable validity across nations. But the very differences in language, culture, and structure that make cross-national research so analytically valuable hinder achieving measurement equivalency. As Hoffmeyer-Zlotnik and Wolf (2003) note, cross-national research requires "profound knowledge of the different national concepts, cultural and organizational structure behind the variables, and the national indicators used to measure the variables of interest."

The challenge can be illustrated by a simple model of comparative research that considers four dichotomous dimensions: (1) language — same/different,

[1] For different types of equivalence see Johnson, 1998; see also Billiet & Philippens, 2003; Hahn, Bode, Du, 2006; Knoop, 1979).

Handbook of Survey Research, Second Edition

ISBN: 978-1-84855-224-1

(2) culture — same/different, (3) structure — same/different, and (4) nation — intra/inter. Even this simplified exposition produces 16 variations, ranging from intranational, monolingual surveys involving no appreciable differences in culture or structure to international surveys in more than one language, involving appreciable differences in both culture and structure.

Language might seem the greatest barrier to comparable research since it is the mechanism for collecting data and the existence of different languages means that measurements differ. Many basic differences exist across languages that hinder simple linguistic equivalence. But structural differences can also present impediments. For example, if governmental systems differ across countries, one cannot ask about the same offices or procedures. Also, legal restrictions on surveys vary. China now permits much survey research, but many political questions about the Communist party are forbidden. If plural marriage is allowed in some cultures and forbidden in others, the marital status item must vary. Cultural differences can also hinder comparability, affecting the basic mechanics of doing surveys such as the gender of interviewers, acquiescence bias, comprehension levels, and conceptualization of constructs (e.g., whether "democracy" has similar meaning across societies). Language, structure, culture, and their interactions all have to be considered.

Most of the discussion of achieving functional equivalence in cross-national, survey research in this chapter is framed by the total-survey-error paradigm introduced in the next section (Biemer, CITE). Then, several topics that cut across this perspective are considered: (1) emic and etic questions and (2) intra/international comparisons.

24.2. Total Survey Error

For cross-national comparability, functional equivalency is needed at all stages a survey: (1) overall study design, including the target population, type of surv (e.g., panel, cross-sectional, experimental; mode); (2) sampling (e.g., sample fra coverage, respondent-selection procedure); (3) response rate and nonresponse b (4) interviewer recruitment, training, supervision, and validation; (5) instrum development (e.g., pretesting, wording, response scales, order/context, item scali (6) translation, (7) data capture and processing (e.g., data entry, coding, transfer back-ups); (8) documentation and archiving; and (9) analysis (e.g., statistical presentational procedures; write-ups) (Smith, 2005).

Doing creditable surveys is difficult even for a single survey on a homo target population. Simply stated, total survey error is "the difference be [a statistic's] actual [true] value for the full target population and the value est from the survey" (Lessler, 1984). The total-survey-error approach stresse (1) surveys are complex mechanisms with many parts and that both rand nonrandom error can come from each part and (2) the different components with one another in complicated ways (Smith, 2005).

When more than one survey is involved, one must rigorously apply the total-survey-error approach to each survey. The more surveys involved, the more challenging this task becomes. Moreover, to insure that results are comparable, one needs to also compare the error structure of each survey to that of all others.

The task is further complicated when the multiple surveys are cross-cultural or cross-national (Heath, Fisher, & Smith, 2005). Differences in language, culture, and structure across target populations mean that "identical" surveys cannot be administered and that the error structure of each survey is likely to vary. Besides the extra difficulties that emerge from innate differences between target populations, additional variation arises from organizational and procedural differences across different survey-research organizations. Cross-national surveys are typically done by different organizations in each country with their own house styles, protocols, and interviewing staffs. Basic societal differences interact with these organizational differences to create further variation. What is needed in cross-national research is achieving functional equivalence across surveys.

24.2.1. *Study Design*

Basic aspects of study design include defining the target population, the type of survey, and the data-collection mode. In cross-national surveys, the seemingly simple goal of adopting a similar study design across countries is often hampered by structural, cultural, and organizational differences. For example, in most developed countries, households and housing units are typical units of analysis and the corresponding target population consists of household residents. In many developing countries, residences are often organized around extended families and familial compounds are major residential units. In other countries many people live in work camps or single-sex dormitories rather than in households.

Survey-research organizations in different countries have vastly different experiences. Type of survey is often influenced by organizational differences across countries. Panel surveys and experiments are relatively rare in some countries, especially in developing nations. Some organizations are unfamiliar with complex instruments such as factorial vignettes, so close and detailed coordination is needed if they are to be correctly executed.

Survey responses often differ by mode of administration, which often is not held constant in cross-national surveys (Nicoletti & Peracchi, 2005). The International Social Survey Program (ISSP), for example, permits postal surveys in addition to the in-person surveys used in most countries. Such differences often occur because of organizational constraints. In Canada, for example, national in-person surveys are rare and no university-based, survey-research organization conducts them. Conversely, because of low telephone penetration rates, many survey organizations in developing countries do not do telephone surveys.

To illustrate the problem of mode and coverage, a random digit-dialed survey of phone numbers in countries 1, 2, and 3 might produce similar samples of

residents of households with landline phones, but yield very biased samples of the respective household populations if country 1 had a high incidence of cell-phone only households, country 2 had high penetration of landlines, and country 3 had low penetration of both landline and cell phones.

Mode matters because many mode effects exist. Among the most consistent is that more socially undesirable or sensitive behaviors (e.g., high alcohol consumption, criminal activity) are reported in self-completion modes than in interviewer-assisted modes (Hudler & Richter, 2001; Tourangeau, Rips, & Rasinski, 2000). But merely keeping mode constant will not automatically solve the problem since mode may not have a constant impact across countries. For example, using show-cards with words in low-literacy societies could create greater differences between the literate and illiterate than a mode that did not interact with education and literacy so strongly.

24.2.2. *Sampling*

Drawing comparable samples across nations is particularly difficult. First, the information available for drawing a sample varies greatly across countries. Many European countries have accessible population registers that list either all households or households and individual household members. In the United States, no such lists exist, but there are good small-area Census statistics for drawing multistage area probability samples. In many developing countries neither exists. Second, legal access to potential sampling frames varies from country to country. Germany, for example, forbids use of its population register, which leads many survey organizations to adopt random-route procedures instead (Smid & Hess, 2003). Finally, there are various practical constraints. For example, in Italy, the electoral register provides a good frame of households and is legally available, but local offices are often reluctant to provide access to it.

Fortunately these major variations are not debilitating impediments. Kish (1994) noted that "sample designs may be chosen flexibly and there is no need for similarity of sample design. Flexibility of choice is particularly advisable for multinational comparisons, because the sampling resources differ greatly between countries. All this flexibility assumes probability selection methods ...." That is, as long as full probability samples are appropriately drawn from sources that cover the comparable target populations, the sources and methods of selecting the samples may vary since they will result in random samples of equivalent target populations in each country (Lynn, Haeder, Gabler, & Laaksonen, 2004).

Other elements of sample design that often vary across countries and can contribute to measurement differences are the method of within-household respondent selection (Lynn et al., 2004), the use of substitution (Smith, 2006c), and coverage of the target group in the sampling frame (Heath et al., 2005). For example, population-register samples include nonhousehold residents, which especially affects the coverage of older respondents.

24.2.3. *Response Rates*

Response rates vary appreciably across countries (Couper & De Leeuw, 2003; Heath et al., 2005; Nicoletti & Peracchi, 2005). Even with considerable effort to standardize designs and resources and to reach certain target minimums, outcomes vary widely. For example, in the first round of the European Social Survey (ESS) with a target minimum response rate of 70% and considerable standardization, response rates varied from 33% in Switzerland to 80% in Greece (Stoop, 2005; Lynn et al., 2004). Such differences result in part from variation in the survey climate (i.e., general national predispositions regarding privacy and information sharing, and specific cultural norms regarding surveys). Additionally, both the level and structure of nonresponse bias varies.

24.2.4. *Interviewers: Recruitment, Training, and Supervision*

Differences in interviewing are a major source of variation in cross-national surveys. First, the type of people recruited as interviewers varies. For example, in less developed countries interviewers tend to be much better educated than respondents, while in developed countries they tend to be average in education. In the United States, field interviewers are overwhelmingly women, but in other countries (e.g., Muslim societies), interviewers are mostly men. On the ISSP (Smith, 2007b), 56% of countries used no full-time, professional interviewers, while 22% had a majority in this category. Similarly, 37% used no student interviewers, while 19% had students as a majority of interviewers. Second, across countries interviewer training ranges from a couple hours to several days (Smith, 2007b). Third, general supervision ranges from minimal contact and no observation to supervisors routinely accompanying interviewers. Finally, validation practices vary greatly. In some countries, extensive recontacting verifies a proportion of all interviewers' cases. Other countries do not collect respondent identifiers to protect confidentiality, so no recontact validation is possible. Under these circumstances, some organizations have developed clever validation techniques such as asking respondents to write out responses to open-ended questions (subsequently comparing respondent and interviewer handwriting), or including obscure factual items whose correct answers are revealed by subsequent questions (allowing interviewers, but not respondents, to learn the right responses to these items).

24.2.5. *Instrument Content and Development*

Instrument development has several components. First are the parts of the questionnaire: the overall content and substance of a study, conceptualization of constructs, scales to operationalize constructs, individual questions to form scales, and response options that capture answers to individual questions. Second is the

pretesting and development work to fashion these elements. Finally, there are many measurement effects (e.g., social desirability, acquiescence, extremity, no opinion, middle options, and context/order) that affect responses. Question wordings and their adaptation are both the key to achieving functional equivalence across surveys and "the weakest link" (Kumata & Schramm, 1956). Items must not only be reliable and valid, but have comparable reliability and validity across societies. This second, comparative step is, of course, complicated by differences in language, structure, and culture.

Questions have two parts, the body of the item presenting the substance and stimulus, and the response-scale recording the answer. These two components are considered separately.

24.2.5.1. Question-asking part First, there is the substantive meaning and conceptual focus of a question. The challenge is to achieve functional equivalence across versions of the questionnaire. One need is an optimal translation.[2] As important and difficult as this is, however, even an optimal translation may not produce equivalency.

The meaning of cognates between fairly closely related languages can differ substantially. For example, the concept equality/egalité is understood differently in America, English-speaking Canada, and French-speaking Canada (Cloutier, 1976). Likewise, for Spanish-speaking immigrants in the United States, "educación" includes social skills of proper behavior missing from the meaning of "education" in English (Greenfield, 1997).

A related problem occurs when a concept is easily represented by a word in one language and no word corresponds in another language. For example, 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). Conversely, Jowell (1998) relates that Japanese researchers in a religion study indicated that there was no appropriate word or phrase in Japanese that closely matched a Western, monotheistic concept of God.

Besides language incompatibility, differences in conditions and structures also hinder functional equivalence. First, situational differences can interact with words that may have equivalent literal meanings but different social implications. As Bollen, Entwisle, and Alderson (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.

2. See later section on translations.

Also, structural differences mean that equivalent objects may not exist or that terms used to describe an object in one country describe something else in another country. For example, the American food-stamp program, which gives qualifying people script to purchase certain food, has no close equivalent in most other countries. Variations in conditions and structures mean that the objects one asks about and how one asks about them differ across societies. This applies to behaviors and demographics as well as to attitudinal and psychological measures. The demographic differences across countries are probably more readily apparent than attitudinal and psychological differences. The sociodemographic variation often relates to well-documented differences in laws, institutions, and other readily apparent societal features. Also, preexisting national data describing each country's sociodemographic profile will typically exist while no relevant information exists on many attitudinal constructs under investigation.

The differences are likely to be at least as large for demographics as for other variables (Braun & Mohler, 2003). For example, a study in Mali added a dimension on relating to animals to the standard American occupational classifications of how jobs relate to data, people, and things (Schooler, Diakite, Vogel, Mounkoro, & Caplan, 1998). More generally, considerable work has been devoted to enhancing the cross-national comparability of the International Standard Classification of Occupations (Elias, 1997).

Demographics can be among the least compatible of variables. Some demographics must use country-specific terms for both questions and answers. For example, region of residence uses country-specific units (e.g., "states" in the United States, "provinces," in Canada, "länder" in Germany) and of course the answers are unique geographic localities. Likewise, voting and party preference must refer to country-specific candidates and political parties.

Some demographics might be asked in either country-specific or generic, cross-country manners. For example, a generic education question might ask, "How many years of schooling have you completed?" A country-specific approach might ask about the highest degree obtained, the type of school attended, or the examination passed. The ISSP, for example, follows the latter course. The generic question produces a simple, superficially equivalent measure, but combines people educated in different educational tracks within a country. But with the country-specific approach, one has to analyze unique, country-specific, educational categories across nations.

Adding problems of linguistic and structural equivalence to the already notable monolingual challenge of creating valid measures increases the need for multiple indicators. Even with careful translations, it is difficult to compare the distributions of two questions that employ abstract concepts and subjective response categories (Grunert & Muller, 1996; Smith, 1988). It is doubtful that responses to the query "Are you very happy, pretty happy, or not too happy?" are precisely comparable across languages. Most likely 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. Similarly, the adjectives "very," "pretty," and "not too" are unlikely to have precise equivalents. Even if, for example, the

English adjective "very" is consistently (and correctly) translated into the French "tres", it is unknown if "very" and "tres" cut the underlying happiness continuum at the same point.

Cross-national research needs multiple indicators that use different concept words and response scales, both to cover a construct and to separate language effects from substantive differences (Smith, 2003).

24.2.5.2. Answer-recording part Achieving equivalency in response categories is as important as establishing the equivalency of the concepts and substance of questions. Several solutions have been offered to increase cross-national equivalency between responses to questions. Among these are nonverbal scales, simple response scales, and calibrating response options.

24.2.5.2.1. Nonverbal scales Some advocate numerical or other nonverbal scales (Fowler, 1993). These include such numerical instruments as ratio-level, magnitude-measurement scales, scalometers, feeling thermometers, and frequency counts. Nonnumerical, nonverbal scales include such instruments as ladders, stepped mountains, and figures or symbols often used in psychological tests. Numerical scales are assumed to reduce problems by providing an universally understood set of categories that have precise and similar meanings (e.g., 1, 2, 3 or 2:1) so that language labels are not needed. Similarly, visual questions and response scales using images are thought to reduce verbal complexity.

However, nonverbal approaches have their own problems. First, many numerical scales are more complex and difficult than verbal items. For example, the magnitude-measurement method assigns a base value to a reference object and asks respondents to rate other objects by assigning values that reflect their ratio to the reference item (Lodge & Tursky, 1981, 1982; Hougland, Johnson, & Wolf, 1992). In the United States, this complex task typically confuses 10–15% of people, who cannot supply meaningful responses. Such confusion may vary across countries, perhaps covarying with levels of numeracy.

Second, numerical scales are not as invariant in meaning and error free as their mathematical nature presupposes. Schwarz and Hippler (1995) found that people rate objects quite differently on 10-point scales going from 1 to 10 than on scalometers going from -5 to -1 and $+1$ to $+5$ (Smith, 1994). Another example is that the 101-point feeling thermometer is not actually used as such a refined measurement tool (Wilcox, Sigelman, & Cook, 1989; Tourangeau et al., 2000).

Third, most societies have lucky and unlucky numbers (e.g., notice how many U hotels have no 13th floor), which may influence numerical responses. Since lucky a unlucky numbers vary across societies, their effects also differ.

Fourth, numerical scales only reduce the use of words in response scales, n eliminate them. For example, a 10-point scalometer has to describe the dimension which objects are being rated (usually liking/disliking) and the scale's operation

Finally, alternative numbering or grouping schemes influence the reporting frequencies. Respondents are often unable or unwilling to provide an exact co

and estimate in various ways (Tourangeau et al., 2000). These heuristics can vary across societies.

Related problems occur with nonverbal, nonnumerical questions and scales. Visual stimuli are not necessarily equivalent across cultures (Tanzer, Gittler, & Ellis, 1995). For example, in Western-designed matrix items used in psychological testing, the missing element is placed in the bottom right corner (Tanzer, 2005). This works for people using languages running 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 missing element needs to be in the lower, left corner.

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. A majority of respondents placed themselves toward the middle (4–7) in all countries, but the Netherlands clearly was an outlier with by far the fewest in the middle (Smith, 1993).

Translation error was suspected for the Dutch deviation, but a check of the Dutch wording indicated it was equivalent to the English and appropriate and clear in Dutch. The visual display in the Netherlands differed from that employed elsewhere, however. The scale was to have 10 vertically stacked squares. The Dutch scale had 10 stacked boxes, but they formed 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 (Schwarz, Grayson, & Knaeuper, 1998).

24.2.5.2.2. Simple response scales A second suggested solution, in a sense the opposite of the numerical approach, is to keep responses simple by using dichotomies. Advocates argue that yes/no, agree/disagree, and other antonyms have similar meanings and cutting points across languages. While language differences may make difficult to determine where someone is along a continuum, it may be relatively easy measure where someone is relative to a mid-point.

The assumption that dichotomies are simple and equivalent across societies is questionable, however. For example, "agree/disagree" in English can be translated to German in various ways with different measurement consequences (Mohler, Smith, & Harkness, 1998). Also, languages may disagree on the appropriateness of intermediate categories between dichotomies. For example, a "maybe" response may be encouraged or discouraged by a language in addition to its equivalent of "yes/no." Another drawback of this approach is loss of precision. Dichotomies measure only direction, not extremity, and are likely to create skewed distributions.

24.2.5.2.3. Calibrating response scales A third proposed solution calibrates response scales by measuring and standardizing the strength of the labels used. One procedure asks respondents to rate the strength of terms, defining each as a point on a continuum (Smith, 1997). This measures absolute strength and the distance between terms and facilitates the creation of equal-interval scales.

Studies show that (a) respondents can perform the required numerical-scaling tasks, (b) ratings and rankings are highly similar across different studies and populations, (c) high test/retest reliability occurs, and (d) different treatments or variations in rating procedures yield comparable results. Thus, the general technique seems robust and reliable.

The direct-rating approach was used to study terms used in response scales in Germany and the United States, and was later replicated in Japan (Smith, 1997; Mohler et al., 1998; Smith, Mohler, Harkness, & Onodero, 2005). Many response terms were highly equivalent in Germany and the United States, but some notable systematic differences also appeared. Japanese results were also largely in line with the German and American patterns, but there was less agreement.

Besides the technical challenges that this approach poses, its major drawback is that separate methodological studies are needed in each country and language to establish the calibration. Not every cross-national study can undertake these. However, in theory once calibrations are determined, they could be used by other studies without extra data collection. Moreover, since the same response scales are used across many different substantive questions, a small number of carefully calibrated response scales could be used in many questions.

A final approach uses anchoring vignettes to establish comparability across measures (Banks, Kapteyn, Smith, & van Soest, 2004; Bago d'Uva, Van Doorslaer, Lindeboom, & O'Donnell, 2006; King, Murray, Salomon, & Tandon, 2004; Salomon, Tandon, & Murray, 2004). Respondents evaluate and rate short vignettes describing a person's situation regarding the construct of interest. For example, a vignette may describe a person's health status and morbidity and ask respondents to rate the person's health as "excellent, very good, good, fair, or poor." Since the vignette person's objective, health-related conditions are fixed and identical across respondents, differences in ratings are deemed to reflect how the scale is understood and utilized by respondents. When comparing two groups (such as respondents from two surveys in two countries), mean differences in responses to such vignettes can anchor people's ratings of their own health and thus make those ratings more comparable across surveys and subgroups. As with the response-scale, calibration approach, the anchoring-vignette approach does not have to be asked of all respondents on each survey. Instead, adjustment factors obtained in one study might be used in other

The anchoring-vignette approach rests on several assumptions. First, response consistency assumes that respondents use scales to rate people in vignettes in the same way that they use scales to rate their own situation. Second, vignette equivalence assumes the objective situations in vignettes are perceived by respondents across groups in the same way. While not implausible, neither of these assumptions has been seriously tested.

24.2.6. Response Effects

Differences in response effects can also be barriers to achieving cross-national comparability (Hui & Triandis, 1985; Usunier, 1999). The special danger in cross-national surveys is that error components may be correlated with nation such that observed differences reflect response effects rather than substantive differences. 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." Important cross-national sources of measurement variation include effects related to social desirability, acquiescence, extremity, no opinion, middle options, and context/order.

24.2.6.1. Social desirability Social-desirability effects distort people's responses (DeMaio, 1984; Johnson, Harkness, Mohler, van de Vijver, & Ozcan, 2000; Tourangeau et al., 2000). Image management and self-presentation bias lead respondents to portray themselves positively — overreporting popular opinions and actions and underreporting unpopular or deviant attitudes and behaviors.

Social-desirability effects appear common across social groups, but often differ in both intensity and particulars. First, the pressure to conform varies. Such effects are presumably larger in collectivist and conformist societies than in individualist ones (Johnson & Van de Vijver, 2003; Lalwani, Shavitt, & Johnson, 2006). This also applies to immigrant groups within societies. Thus, more collectivist Asian Americans show larger social-desirability effects than more individualist European Americans (Lalwani et al., 2006). In addition, social-desirability effects may interact with characteristics of respondents and interviewers such as race/ethnicity, gender, social class, and age. For example, a well-documented interviewer effect is that people express more intergroup tolerance when being interviewed by someone of another race/ethnicity (Schuman, Steeh, Bobo, & Krysan, 1997; Javeline, 1999).

Likewise, social-desirability effects are likely to be greater when status/power differentials between interviewers and respondents — which are likely to vary across nations — are larger. In developing countries, for example, interviewers tend to be members of educated elites, while in developed countries interviewers are typically of average status.

Moreover, sensitive topics and undesirable behaviors vary both across individuals and cultures (Newby, Amin, Diamond, & Naved, 1998). For example, items about alcohol are much more sensitive in Islamic countries than in Judeo-Christian societies.

To deal with social-desirability effects, one can frame questions in less threatening manners, train interviewers to be nonjudgmental in asking items and responding to answers, and use modes that reduce self-presentation bias.

24.2.6.2. Other measurement effects All other major response effects relating to acquiescence, response extremity, no opinion/nonattitudes, middle options, and context/order also can show variability across countries and social groups. Steps need to be taken to detect and minimize such effects (Smith, 2004).

24.2.7. *Translation*

Translations are needed whenever two or more languages are used by notable segments of the target population. This is most frequent in cross-national studies, but intranational, multilingual surveys are also common. Translations are required in (1) well-recognized multilingual countries such Belgium, Canada, and Switzerland, (2) countries with large immigrant populations such as the United States and Canada, and (3) surveys focusing on immigrants or indigenous, linguistic minorities. For example, in the United States, a recent health survey was conducted in 12 languages (Olson, Osborn, Blumberg, & Brady, 2003) and NORC's New Immigrant Study had full translations in 8 languages and was administered in over 80 languages (Doerr, 2007, personal communication).

Thoughtful pieces on how to do cross-national survey translations exist (Brislin, 1970, 1986; ESS, 2006; Harkness, 1999, 2001; Harkness, Pennell, & Schoua-Glusberg, 2004; Harkness & Schoua-Glusberg, 1998; Prieto, 1992; van de Vijver & Hambleton, 1996), but rigorous experiments to test the proposed approaches are lacking. Because of this the development of scientifically based translation has languished.

24.2.7.1. Translation and instrument development Translation is often wrongly seen as a mere technical step rather than as central to the scientific process of designing valid cross-national questions. Translation must be an integral part of the study design and not an isolated activity (Bullinger, 1995; Harkness, 2006; Pasick et al., 1996). As Pasick and colleagues (1996) describe the designing of a multilingual study, translation is an integrated and interactive part of an eight-step process: (1) conceptual development of topic; (2) inventorying existing items; (3) development of new questions; (4) question assessment through translation; (5) constructing full draft questionnaires; (6) concurrent pretesting across all languages; (7) item revision; and (8) final pretesting and revisions. The keys are that translation is part (a) a larger process of item development and testing and (b) a multistage, interacti process where changes in source and target-language wordings occur at varie points in the design process.

Achieving optimal translation begins at the design stage. Cross-natio instruments should be designed by multinational teams of researchers who sensitive to translation issues and take them into consideration during the design development stages (Bullinger, 1995; Pasick et al., 1996). They need to consider each concept of interest can be measured in each language and society under

In actual practice, the most common model in survey translation is the source-to-target language approach. The survey content is developed, pretested finalized in a source language and then translated into one or more target lang Both the concepts and the operationalization (wordings) are fixed in the language. In the strictest and most common version of this model, no chan made to the source wordings on the basis of the translations. Moreover, when more than one target language, little or no comparison is made across differe

languages. In effect, each source-to-target-language translation is a separate procedure.

A modified version is the iterative, source-to-target language approach. A master source questionnaire is translated into one or more target languages, but if the translation highlights shortcomings or ambiguities in the source language, then revisions of the source questionnaire may be made to correct deficiencies in the source question and thereby clarify the translation in the target language. If there are several target languages and such a feedback loop is engaged for each one, this might lead to multiple changes, some from only a single translation and some from multiple indications related to the same issue. Of course, whenever any change is made in the source language, a new round of translation is needed for all target languages, so unlike the fixed, source-to-target-language approach, the process becomes iterative.

In a truly collaborative approach, a master questionnaire is jointly developed in two or more languages. Here there is no source language and target language(s): all have equal status as primary languages for instrument development and data collection. Investigators agree to cover certain topics with certain general types of measures. Then items are developed, either by multilinguals working simultaneously in each language or by monolinguals working separately in each of the relevant languages. Once items are jointly developed, they are translated into each of the relevant languages, items for pretesting are agreed upon, and pretests are carried out in each of the languages. By comparing the translations and evaluations of the pretests, question content and wordings are revised across most or all languages, followed by further translation and pretesting of the revisions. This process continues through several iterations until a common set of functionally equivalent items with suitable wordings in each relevant language is agreed upon. This is called "decentering" — a process of formulating questions so they are not anchored in one language, but fit equally well in all applicable languages (Carlson, 2000; Eremenco, Cella, & Arnold, 2005; McGorry, 2000; Pasick et al., 1996; Potaka & Cochrane, 2002; Werner & Campbell, 1970).

Problems of translation in general and decentering in particular multiply as the number of languages involved increases and as the linguistic and cultural differences across languages widen. The example of a study involving six languages illustrates the extra complexity of the truly collaborative approach. There, the source-target approach has five paths (source to each of the five target languages), or ten information flows if each path is two-way. A collaborative design involves 15 paths among the six languages, or 30 allowing for two-way flows. Thus, the complexity of the task triples.

A source-target-language approach generally means that items and scales are optimized in the source language and will be less reliable and/or valid in the target languages (Skevington, 2002). This occurs for several reasons. First, there will be no translation errors in the source language. Second, even if outright translation errors are avoided, wordings in the target languages are likely to be less natural and comprehensible than in the source language. Third, the concepts and their operationalizations are likely to be less meaningful and culturally relevant in the target languages than in the source language. Finally, pretesting and other

development work are likely to have been exclusively conducted in the source language and the items selected for inclusion will be those that worked best in that language, usually in one country. Thus, this approach places the source language in a privileged position in terms of the wording of individual items, what items form a particular scale, and the overall content of the questionnaire. This limits flexibility for producing functional equivalence with the source content in target languages, since attributes of items in those languages — such as naturalness, level of difficulty, and familiarity with response-option scales — receive relatively little attention. The approach channels the content toward what is relevant and reliable in the source language, while culture-specific components associated with the target languages and societies are not considered.

24.2.7.2. Translation procedures Various techniques for carrying out translations exist, of which five are distinguished here.

First is the translation-on-the-fly approach under which multilingual interviewers do their own translations when respondents do not understand the source language. This approach lacks standardization and quality control.

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

Third is the back-translation technique under which (1) questions in the source language are translated to a target language by one translator, (2) 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) work with one or both translators to adjust the target language of the problematic questions when notable differences in the source questionnaires appear. This is probably the most frequently recommended translation method (Brislin, 1970, 1986; Cantor et al., 2005; Harkness, 1999). A decided limitation of this technique is that it does not directly assess the adequacy of the target-language questions (Blais & Gidengil, 1993). A poorly worded target-language item that successfully back translates goes undetected.

In a centered back-translation approach, source-language version 1 (SV1) is translated into target-language version 1 (TV1) and then TV1 is translated into source-language version 2 (SV2). Then the following rules are applied: (1) If SV1 = SV2, accept TV1. (2) If SV1 is equivalent to SV2, accept TV1.[3] (3) If SV1 is not equal/equivalent to SV2, reject TV1 or SV2. (3a) Review the SV1 to TV1, TV1 to SV2 translation procedures to locate the reason for nonequivalence, and repeat one or both translations for affected items. If SV1 to TV1 is deemed problematic, this means translating SV1 to TV2 and then TV2 to SV3 (i.e., any source-to-target language translation generates a new back translation). SV1 and SV3 are compared, with possible outcomes similar to the first iter

3. When the back translation is identical to the source wording, they are equal. When the back tr differs from the source in some trivial manner, they are equivalent.

If the translation from TV1 to SV2 is deemed to be problematic, then SV3 is created and SV1 and SV3 are compared, as was previously done with SV1 and SV2. Under a strict source-language-centered approach, SV1 is never changed.

In the fourth, parallel-translation approach, (1) questions in the source language are translated independently by two translators into the target language, (2) the two translations are compared, and (3) the two translators meet with those who developed the source-language questions to determine the reason for the variant translations when they differ appreciably (Bullinger, 1995; Eremenco et al., 2005). Reasons can include simple errors (i.e., poor translations) in one version or ambiguities or other uncertainties in the source language. Like back translation, this approach involves two translations and two translators, but places more emphasis on optimizing wording in the target language. It can be done more quickly than back translation since the two translations are done simultaneously rather than sequentially.

In the fifth, committee-translation approach, a team of translators and researchers discusses the meaning of items in the source language, possible translations in the target language, and the adequacy of the translations in the target language, considering such matters as level of complexity and naturalness as well as meaning (Carlson, 2000; McGorry, 2000). Under this approach different members of the team may produce independent, parallel translations of items or the team may work simultaneously and interactively on a translation. This approach maximizes interaction between translators and between translators and other members of the research team. It places the greatest emphasis on writing good questions, not just on translating words (Harkness, 1999; Harkness & Schoua-Glusberg, 1998). Finally, these translation approaches can be combined (McGorry, 2000; Bullinger, 1995; Eremenco et al., 2005).

24.2.7.3. Aspects of survey translation Survey translators need many skills beyond high competency in the source and target languages. They need to understand the cultures in which the surveys are being administered, survey methodology in general and question construction in particular, and the survey's substance (Carlson, 2000). Translators used to translating documents or doing simultaneous oral translations need special training to be adequate survey translators (Harkness, 2006).

Translations must consider not only language, but also nonlinguistic adaptation across societies. The SF-36 scale exemplifies the need for such "cultural translation." In English it asks about "activities you might do in a typical day" and whether one's health is a limitation in doing them. One item asks about "moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf." This item produced more reports of limitations in China than in the United States. It was thought that this was in part because both bowling and golfing are uncommon in China and considered as difficult activities. Mopping the floor and practicing Tai-Chi were used as complementary examples, but these activities may not be culturally and linguistically equivalent (Li, Wang, & Shen, 2003).

Even within languages, "translations" across countries are often needed (McGorry, 2000). As George Bernard Shaw remarked, "England and America are

two countries separated by a common language." Similarly, Latin American and Iberian Spanish have not only pronunciation differences, but considerable variation in vocabulary. Even within the Spanish-speaking Americas, major differences occur. In one survey, a Spanish word was understood to mean "hit" in one country, but "spank" in another.

24.2.7.4. **Translation and quantitative evaluations** While careful translation procedures are essential for developing equivalent items, they are insufficient alone. Quantitative methods should evaluate qualitative translation procedures. Several approaches exist for quantitatively assessing items and translations. First is the direct evaluation of items. For example, Bullinger (1995) describes a study in which two raters independently judged the difficulty of wordings in the source language, two other raters evaluated the quality of translated items, and two more raters assessed the back-translated items. This allowed both qualitative and quantitative evaluation of the translations, evaluations as to whether the items were comparably understandable, and inter-rater reliability checks on the quantitative ratings.

Second, quantitative ratings of the terms used in response options can determine whether scale points are equivalent.

Third, statistical tests can assess the comparability of cross-national results (Ellis, Minsel, & Becker, 1989; MacIntosh, 1998). While usually applied after data collection at the analysis phase, they should be employed at the development stage. In particular, item-response theory (IRT) has been used to measure equivalency and even assess whether differences are due to translation errors or cultural differences (Eremenco et al., 2005; Hahn et al., 2005, 2006; Lin, Chen, & Chiu, 2005). For example, in a French–German comparison of psychological scales, 10% of the items tested as nonequivalent (Ellis et al., 1989). Excluding the nonequivalent items from scales resulted in one major change in substantive interpretation. Germans rated lower than the French on self-actualization using all items, but no nation differences appeared when only equivalent items were used.[4]

IRT testing has some drawbacks, however. Some consider it too exacting, preferring other techniques such as confirmatory or exploratory factor analysis (Ellis et al., 1989; MacIntosh, 1998; Ryan, Chan, Ployhart, & Slade, 1999) or internal consistency analysis (Eremenco et al., 2005). Such analyses require pilot studies with 200 + respondents per language or country for reliable results, which exceeds the pretesting resources in many studies (Eremenco et al., 2005).

These quantitative evaluation approaches can be combined. Items might be evaluated on various dimensions related both to language (e.g., clarity, difficulty) and substance (e.g., extremity, relevancy). These ratings could then be com-

4. Nonequivalent items should not merely be discarded. As Ellis, Minsel, & Becker (1989) note, "nonequivalent items ... should be examined separately for potential clues of real cross-cultural dif-

across languages (as in the rating of response options) and correlated with results from pilot data using IRT or other techniques.

The various quantitative techniques should be used together with qualitative techniques (Carlson, 2000; Eremenco et al., 2005). For example, in the German-American study of response options (Mohler et al., 1998), equivalent English and German terms for answer scales were developed by translators and then respondents rated the strength of the terms on the underlying dimensions (agreement/disagreement and importance). In almost all cases, the mean ratings of the German and English terms were the same, thereby validating translation equivalency (e.g., finding that "strongly agree" and its German translation were both rated similarly on a 21-point scale that ran from total and complete agreement to total and complete disagreement). In another German-American study using IRT testing (Ellis et al., 1989), an American verbal-reasoning question was found not to be equivalent in German. Evaluation of its wording revealed that the difference occurred because poodles are not considered as retrievers in England and America, but poodles were originally bred as waterfowl retrievers in Germany and they are regarded as part of the latter set. In both cases, qualitative assessment and quantitative measurement yielded consistent judgments about the equivalency of response options or items.

24.2.7.5. **Translation and bilinguals** Some have proposed that translation equivalence can be established by administering items in two languages to bilingual respondents. Bilinguals understand and process language differently than monolinguals do, however (Blais & Gidengil, 1993; Ellis et al., 1989; Lin et al., 2005). Despite this serious impediment, useful evaluations can be gained by comparing results within societies, but across languages (Carlson, 2000). In a test of whether French Canadians were less supportive of democracy than English Canadians, Blais and Gidengil (1993) found that within and outside of Quebec both English and French Canadians interviewed in French were less supportive of elections than English and French Canadians interviewed in English. Their statistical analysis showed that language, rather than culture, explained the differences in support for democracy. A study of the SF-36 with Chinese and English bilinguals and Chinese monolinguals found that bilingualism did not influence responses, controlling for age, education, and other factors on which the groups differed (Thumboo et al., 2002).

24.2.7.6. **Translation and experimental designs** Rigorous empirical testing to better document the strengths and weaknesses of translation in general or to assess the effectiveness of the various translation approaches has yet to be done. One research design would separately translate a series of questions into one or more target languages using back translation, committee translation, and other approaches. The results would be compared to each other and evaluated by a team of language and survey experts.

A more desirable research design would pretest each of the translations and comparatively evaluate them using the pretesting techniques described below to see which produced fewer problems in the target language(s). Better still would be

fielding the different versions to permit comparison of results from the different translation approaches in terms of data quality (e.g., scale reliability).

A third research design would conduct experiments using bilinguals, who could be classified into four groups depending on whether they are native or nonnative speakers of the two languages (i.e., native in both; native in source, nonnative/studied in target; nonnative/studied source, native in target; and nonnative/studied both). In addition, their competency in each language could be formally rated. Then using a between-subjects design, they could be randomly assigned to the source or target language for cognitive interviews to see if items produced similar substantive understanding, comprehension, and measurement errors across languages. Similarly, rather than cognitive interviews, regular interviews could be administered randomly across languages for the same bilingual groups to see if the same quantitative results emerged. Alternatively, a within-subjects design could be employed with the order of languages also randomized, perhaps with buffer questions between the two parallel sets of items.

Nonexperimental studies can be done using surveys administered in two or more languages. For example, the 2006 General Social Survey was done in Spanish and English with respondents selecting their strongest language (Smith, 2007a). Respondents also indicated their language ability in the language not selected. Among Hispanics this identified four groups: English monolinguals, bilinguals interviewed in English, bilinguals interviewed in Spanish, and Spanish monolinguals. Comparison of responses across these four groups focused on cases in which there were no differences in distributions between the two English groups, no differences between the two Spanish groups, and differences across the bilingual groups using English and Spanish. Controls for assimilation and sociodemographic variables were introduced to see if they explained apparent language differences. While not as strong as the foregoing experimental designs, this method avoids special data collection and uses larger and more representative samples associated with final studies rather than smaller and less generalizable samples utilized in pretests and methodological experiments. It can detect translation problems only after final data collection, however.

Optimal translations are essential for achieving item and scale equivalency. Researchers should (1) make translations an integrated part of the development of studies, (2) utilize the best approaches such as committee and combined translation and (3) use quantitative as well as qualitative methods to evaluate translations.

24.2.8. *Pretesting and Related Questionnaire Development Work*

Developmental work must establish that items and scales meet acceptable techn standards (e.g., of comprehension, reliability, and validity) in each country are comparable across countries (Krebs & Schuessler, 1986; Pasick et al., 1 Pretesting is also an important component in the translation process. Hudler Richter (2001) observe that "it is essential that the instrument is carefully de

and analyzed in a pretest" in cross-national research. Moreover, pretesting should be "a team effort with multiple disciplines and preferably multiple cultures represented" (Pasick et al., 1996).

Devoting more time and effort to pretesting leads to better instruments (Bullinger, 1995; Perneger, Leplege, & Etter, 1999). Useful developmental and pretesting procedures include: (1) cognitive interviews using such protocols as think-alouds (Bolton & Bronkhorst, 1996; Gerber & Wellens, 1997, 1998; Johnson et al., 1997; Levine, Gonzalez, Weidmer, & Gallagher, 2004; Tourangeau et al., 2000), (2) behavioral coding with the interviewer–respondent exchanges recorded, coded in detail, and then analyzed (Fowler & Cannell, 1996; Johnson & Bowman, 2003; Pruefer & Rexroth, 1996; Krosnick, 1999), and (3) conventional pretesting (Converse & Presser, 1986; Fowler, 1995; Hudler & Richter, 2001).

Two major obstacles to effective developmental work in cross-national surveys exist: (1) the dearth of methodological studies of the various pretesting approaches and (2) a general underutilization of pretesting.

First, few studies have systematically compared pretesting methods (Presser & Blair, 1994; Willis & Schechter, 1997). Presser and Blair's comparison of pretest methods (conventional, cognitive, and behavioral coding plus expert panels) found considerable differences in the number and nature of problems revealed by the different approaches, indicating that multiple methods should be used. There are no similarly rigorous cross-national comparisons of pretesting.

Studies of cognitive pretesting in such countries as Australia, Belgium, Taiwan, and the United States (Foddy, 1998; Nuyts, Waege, Loosvelts, & Bulliet, 1997; Tien, 1999) have all found this approach valuable, but the optimal combination of cross-national, pretesting approaches has not been established. In addition, a few studies suggest that all pretesting techniques may not be equally effective in all languages (Levine et al., 2004; Pan, 2004). It is unclear whether this is due to intrinsic traits of languages, differing nonlinguistic social traits, or differences in the skill and experience of different pretesters.

Second, most cross-national studies fail to devote adequate time and resources to pretesting. A review of pretesting procedures used in various cross-national surveys found that resources for pretesting were usually severely limited. Pretests are usually too small to allow more than a qualitative assessment of whether items are working. Pretests are also often limited to atypical, convenience samples, like college students. Additionally, most studies use only conventional pretesting. Cognitive pretesting, behavioral coding, think-alouds, and other advanced techniques are rare. Perhaps the most serious problem is that pretests are sometimes not allowed to play their important role in developing items. For example, while the World Fertility Survey used larger pretests than usual (almost all with 100 + cases) and even audio-taped many interviews (a good, but rare, procedure), its content was basically fixed in advance and revised little based on the pretests (Cleland & Scott, 1987, pp. 32–33, 384).

More methodological studies of pretesting are needed. In advance of such studies, a few general guidelines based on what appear to be the best current practices are: (1) multipretesting procedures should be carried out across countries and languages

with results evaluated by researchers expert in (a) the cultures and languages being investigated, (b) the substantive domains being studied, and (c) survey-research methodology; (2) pretesting and translating should be integrated and interactive processes; (3) pretesting needs to be cross-national; and (4) the developmental process takes much more time and resources than for single-country, monolingual studies and usually should involve multiple rounds of pretesting and larger samples.

24.2.9. *Data Capture and Processing*

Quality-assurance systems are needed and data must be carefully handled, cleaned, and checked. Systems and procedures need to be consistent across surveys. In some countries extensive cleaning is routinely conducted so that all structural disagreements across variables and improbable values for a single variable are checked and corrected or recoded as missing. In other countries, disagreements are seen as a real part of the data and little or no cleaning is conducted. Hence, as with everything else, consistency in data processing cannot be assumed, but must be planned and verified.

Open-ended questions especially have cross-national implications (Heath et al., 2005). First, respondents may vary across cultures in their willingness and ability to provide full and complete open-ended responses. Second, interviewers may have difficulty in probing for complete answers and in their accuracy in recording verbatim responses. Better interviewers with special training are needed when open-ended questions are heavily utilized. Third, open-ended material must be consistently coded according to a common coding frame. Codes must both be understood and utilized the same way across surveys and coders. This requires detailed codi protocols, coordinated training, and, ideally, consultation among the codi supervisors across countries. Fourth, codes must be universal enough to allo comparability, but sensitive enough to capture important country-specific details a distinctions. Finally, the full verbatims need to be retained and then translated i languages used in analysis to allow authors to fully utilize the material and use quo as appropriate.

24.2.10. *Documentation and Archiving*

Heath et al., 2005 observed that "the documentation for cross-national s research needs to be especially thorough, but is rarely available." All phases of survey from sampling to data processing need to be carefully recorded (Hu Richter, 2001; Uher & Mohler, 2003). It is particularly important to inclu original questionnaires, so that users can consult them to understand result countries. More meta- and paradata should be added to the data and cod The ESS, for example, includes such paradata as case-level, record information and such metadata as content analysis of the national medi et al., 2004).

Verma (2002) noted that "microdata distribution in particular requires the closest attention" and there should be "economical, liberal, and easy access to microdata by researchers." While no data archive specializes in cross-national data, several have extensive comparative collections, including the Inter-university Consortium for Political and Social Research (www.icpsr.umich.edu) and the Roper Center for Public Opinion Research (www.ropercenter.uconn.edu) in the United States and the Norwegian Social Science Data Services (www.nsd.uib.no), the Central Archive for Empirical Social Science Research in Cologne (www.gesis.org/en/za), and the Economic and Social Data Service at Essex (www.esds.ac.uk)in Europe. In addition, major projects such as the Comparative Study of Electoral Systems, ESS, ISSP, and World Values Survey maintain on-line archives.

24.2.11. *Analysis*

Several aspects of cross-national analysis are especially important. First, analysts must be familiar with all of the data sets being analyzed and with each of the cultures covered by the comparative data. Interpretative errors will increase in direct proportion to cultural ignorance. Essential knowledge includes, but is not restricted to, understanding current conditions, historical and developmental patterns, cultural norms and values, and structural and legal differences. Especially when many nations are involved, this increases the desirability of involving multiple researchers who collectively have expertise covering the substantive topic, survey and analytical methodology, and the various countries and cultures involved.

Second, considerable effort should be devoted to looking for artifactual causes for any large and surprising findings. If a difference is both large and unexpected, it likely results from measurement variation rather than from real differences.

Third, while multilevel analysis is valuable even with single surveys in one country, it is natural in cross-national research where simultaneous individual and country-level analysis should be routine (Jusko & Shively, 2005). Cross-national analysis should also add neighborhood and community-level analysis to the individual and national levels.

Finally, advanced analytical procedures such as structural equation models and multitrait/multimethod techniques can be adapted for cross-national analysis (Billiet, 2003; Saris, 2003).

3. Emic and Etic Questions

"etic" questions are items with a shared meaning and equivalence across cultures, while "emic" questions are items relevant to some subset of the cultures under study (Skevington, 2002). Suppose that one wanted cross-national data on political participation in general and contacting government officials in particular. In the United States items on displaying bumper stickers, visiting candidate Web sites, and

emailing public officials would be relevant. In most developing countries, these would be meaningless. Conversely, an item about asking a village elder to intervene with the government might be important in developing societies, but have little relevance in developed nations.

In such circumstances, solutions include (1) using general questions that cover country-specific activities within broader items, (2) asking people in each nation both the relevant and irrelevant items, or (3) an emic/etic approach, asking a core set of common items (e.g., voting in local and national elections, talking to friends about politics), plus separate country-specific items.[5]

Using general items is perhaps least appropriate. The necessary loss of detail is usually extensive and general items may be too vague and sweeping.

The relevant and irrelevant approach can succeed if the number of low relevancy items is not too great and those items are not nonsensical or otherwise inappropriate. For example, the ISSP successfully used this approach to study environmental change. Items on personal car use were asked in all countries, even though ownership levels were quite low in some countries.

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

Likewise, in developing the Chinese Personality Assessment Inventory, researchers found that important parts of Chinese personality (e.g., ren quin or relationship orientation) did not match any dimension on standard, Western scales and needed to be added (Cheung et al., 1996). The emic/etic approach indicates that sometimes one needs to do things differently in order to do them equivalently (Przeworski & Teune, 1966).

Moreover, maximizing comparability across countries and languages might lowe average reliability in each relevant country/language. The elimination of emic iten from a scale made up of exclusively etic items might weaken the best scale of ea country or language. That is, by focusing on only the shared or common aspec of a construct, the measurement of that construct might become less comple less reliable, and more biased by the exclusion of culture-specific items. The us items that converge across languages and countries toward developing a comm comparable set of questions might overrepresent shared elements across societies thereby underestimate cross-cultural variation and uniqueness.

5. However, even identical actions, such as voting in the last national election may not be equiv some countries voting is legally mandatory, so it is not a meaningful measure of voluntary, activity. In other countries elections are meaningless charades, so voting is not a meaningful m participating in a democracy.

Measure A in country 1 (A1) is likely to be most equivalent to measure A in country 2 (A2) when they are jointly developed with the goal of functional equivalence ($A1_j$ and $A2_j$). Such joint measures may be less optimal in each country than culture-specific measures of the same constructs (say $A1_x$ and $A2_y$). $A1_j$ and $A2_j$ may provide more comparable measurement than $A1_x$ and $A2_y$, if the former pair reduces variation in the error structures across countries (or if it lowers the correlation of measurement error with country), but $A1_j$ and $A2_j$ will each have more measurement error than $A1_x$ and $A2_y$. In some cases this is worthwhile, but in other cases it will not be. In a study of political participation, a jointly developed measure on voting in the last national election and writing a letter to a public official might produce comparable but limited and biased coverage of political participation, if it excluded using political blogs in country 1 and talking to village elders in country 2.

As valuable and essential as standardization is, it can be taken too far, especially when it is applied rigidly and formulaicly without regard for the underlying goal of functional equivalence. As Hamilton and Barton (2000) observe for a cross-national adult literacy scale, the common-ground approach "involves identifying a common cultural core of test items which elicit a similar pattern of response across all cultures and language groups ... [and] any literacy practice not recognized beyond a particular cultural group cannot be used to generate items for the cross-cultural study since this would constitute cultural bias."

For example, an international study on sports might focus on soccer as the sport of global interest and find lower participation and viewership in the United States than in most other countries. This would present an accurate view of soccer, but a very distorted view of the role of sports in general, neglecting top US sports such as football, baseball, basketball, and stock-car racing. These could be added, but then the UK and other Commonwealth nations would be disadvantaged unless cricket and rugby also were included. Similar arguments extend to including sumo wrestling in Japan, curling in Canada, etc. Accepting only shared items across countries argues against this inclusive practice, however.

4.4. Intra/International Comparisons

ypically intranational, subgroup differences will be smaller and easier to deal with an cross-national differences. Differences in language, culture, and structure are bly to be smaller within than between nations. First, in many countries the vast jority of respondents is monolingual, so language is often not an issue. When ltiple national languages do exist (e.g., Canada, Belgium, Switzerland), there are y bilingual speakers and interviewers and well-established concordances across languages. Across nations, language differences would be more the rule and y comparisons would be across very different languages (e.g., Finnish and Thai). nd, structural differences are likely to be minimal within countries since nations lly have one legal and governmental system, a national economy, and other larities across subnational groups that are not typical across cross-national

groups. Finally, cultural differences would on average be smaller within nations, with intranational groups more likely to have a shared history, unified educational system, common mass media, etc. than groups in different nations. For example, as different as Blacks are from Whites in the United States and as Francophones are from Anglophones in Canada, they do share a large number of commonalities such as citizenship, the same laws, and geographic proximity.

Additionally, surveys of intranational subgroups would typically be integrated, with one organization following one study design and set of protocols. This eliminates measurement variation due to organizational differences. Also, intranational surveys will be used to dealing with the subject groups and the group-specific issues involved. For example, while achieving functional equivalency across French and English versions remains a challenge in Canada, survey researchers there are very experienced in designing comparable questionnaires in these two national languages and for these subnational populations, much more so than American and French organizations would be in conducting a one-time, cross-national study in the United States and France.

Although intranational, cross-cultural differences are both smaller and more easily manageable, especially when a common language is utilized (e.g., as between Blacks and Whites in the United States or the Scots and English in the United Kingdom), the differences that do occur are underexamined and underappreciated. As a result, they are often not adequately dealt with. For example US studies show different measurement error for Blacks and Whites (Johnson Bowman, 2003) and find differences in item comprehension across ethn racial groups even controlling for education (Johnson, Kulesa, ISR LLC, Cho Shavitt, 2005).

Subgroup differences within a country are most often thought of in term variation across linguistic, ethnic, religious, and racial groups. These are see cultural subgroups that are likely to live in some isolation from others, to share among themselves and to differ from nongroup members in terms of key attri and norms. This is most apparent among immigrant communities where diffe in language skills, socialization, legal status, and other matters exist and there is often geographic and social separation from the majority-group Indigenous minorities such as American Indians in the United States, L Sweden, and Basques in Spain are similar examples.

Other subgroup differences unrelated to such cultural subgroups may al valid, subgroup comparisons. Differences in vocabulary and comprehensi exist across educational and class groups, cohorts, regions, and genders ground needs to be found in constructing items so that subgroup variation interfere with collecting comparable information. Shared validity across is sought. In particular, interrespondent variation in understanding and should be minimized. What unavoidable variation remains should be k and unrelated to basic sociodemographics.

Within-language surveys search for linguistic common ground, a lingua franca, to achieve widespread and equivalent understanding acro population (Smith, 1988). This includes using words and syntax co

to all segments of society, including those with limited education and subgroup members who are outside the cultural mainstream; explaining or defining terms when needed; and avoiding slang and fad phrases. This must be verified by careful pretesting and not assumed because items were crafted by experienced survey designers.

24.5. Conclusion

Comparative survey research faces the great challenge that languages, social conventions, cognitive abilities, and response styles vary across societies. This means that the phenomenon under study and the means of studying it are confounded (Fiske, Kitayama, Markus, & Nisbett, 1998). Achieving valid, equivalent measurement across cultures requires the total-survey-error approach. Survey-design variation and measurement error need to be minimized and equalized to obtain valid, reliable, and consistent substantive data. Achieving this is neither simple nor easy. Obtaining cross-national comparability is so complex and challenging that greater effort is needed at every stage, from conceptualizing the research question, to instrument development to data analysis. But the substantive gains from cross-national research fully justify all of the extra efforts.

Fortunately, comparativists and survey-research methodologists are not only recognizing the challenges presented by cross-national research, but taking concrete steps to answer those challenges by establishing professional associations (e.g., forming the European Survey Research Association in 2005), conferences (e.g., the International Workshop on Comparative Survey Design and Implementation in 2002), and international standards for survey research (e.g., by the International Organization for Standardization in 2006) (Lynn, 2003; Verma, 2002; Lievesley, 2001; Smith, 2001, forthcoming, 2006a; 2006b; www.iso.org).

References

d'Uva, T., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2006). *Does reporting heterogeneity bias the measurement of health disparities.* Tingergen Institute Discussion Paper, TI 2006-033/3.

, Kapteyn, A., Smith, J. P., & van Soest, A. (2004). *International comparisons of work disability.* Discussion Paper IZA DP no. 1118. Institute for the Study of Labor.

(2003). Cross-cultural equivalence with structural equation modeling. In: Harkness, F. J. R. Van Der Vijver & P. Ph. Mohler (Eds), *Cross-cultural survey* . London: WileyEurope.

& Philippens, M. (2003). Data-based quality assessment in ESS — Round 1. In: *work package 7.* Available at www.europeansocialsurvey.org

& Gidengil, E. (1993). Things are not always what they seem: French-English and the problem of measurement equivalence. *Canadian Journal of Political* 26, 541–555.

Bollen, K. A., Entwisle, B., & Alderson, A. S. (1993). Macro-comparative research methods. *Annual Review of Sociology, 19*, 321–351.
Bolton, R. N., & Bronkhorst, T. M. (1996). Questionnaire pretesting: Computer-assisted coding of concurrent protocols. In: N. Schwarz & S. Sudman (Eds), *Answering questions: Methodology for determining cognitive and communicative processes in survey research.* San Francisco, CA: Jossey-Bass.
Braun, M., & Mohler, P. Ph. (2003). Background variables. In: J. A. Harkness, F. J. R. Van Der Vijver & P. Ph. Mohler (Eds), *Cross-cultural survey methods.* London: WileyEurope.
Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Research, 1*, 185–216.
Brislin, R. W. (1986). The wording and translation of research instruments. In: W. J. Lonner & J. W. Berry (Eds), *Field methods in cross-cultural research.* Newbury Park, CA: Sage.
Bullinger, M. (1995). German translation and psychometric testing of the SF-36 health survey: Preliminary results from the IQOLA Project. *Social Science Medicine, 41*, 359–1366.
Cantor, S. C., Byrd, T. L., Groff, J. Y., Reyes, Y., Tortolero-Luna, G., & Mullen, P. D. (2005). The language translation process in survey research: A cost analysis. *Hispanic Journal of Behavioral Sciences, 27*, 364–370.
Carlson, E. D. (2000). A case study in translation methodology using health-promotion lifestyle profile II. *Public Health Nursing, 17*, 61–70.
Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese personality assessment inventory. *Journal of Cross-Cultural Psychology, 27*, 181–199.
Cleland, J., & Scott, C. (Eds). (1987). *The world fertility survey: An assessment.* Oxford: Oxford University Press.
Cloutier, E. (1976). Les conception Americaine, Canadienne-Anglaise, et Canadienne Francaise l'idee d'eqalite. *Canadian Journal of Political Science, 9*, 581–604.
Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Beverly Hills, CA: Sage.
Couper, M., & De Leeuw, E. (2003). Nonresponse in cross-cultural and cross-national surveys. In: J. A. Harkness, F. J. R. Van Der Vijver & P. Ph. Mohler (Eds), *Cross-cultural survey methods.* London: WileyEurope.
DeMaio, T. J. (1984). Social desirability and survey measurement: A review. In: C. F. Turner & E. Martin (Eds), *Surveying subjective phenomena* (Vol. 2). New York: Russell Sage.
Durkheim, E. (1938). *The rules of sociological method.* Glencoe, IL: The Free Press.
Elias, P. (1997). *Occupational classification: Concepts, methods, reliability, validity, and national comparability.* Paris: Institute for Employment Research.
Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluations of attitude survey translations: investigations using item response theory. *International Journal of Psychology, 24*.
Eremenco, S. L., Cella, D., & Arnold, B. J. (2005). A comprehensive method for translation and cross-cultural validation of health status questionnaires. *Evaluation & Health Professions, 28*, 212–232.
European Social Survey. (2006). Translation strategy. Available at www.europeansocial
Fiske, A. P., Kitayama, S., Markus, H. R., & Nisbett, R. E. (1998). The cultural social psychology. In: D. T. Gilbert, S. T. Fiske & G. Lindzey (Eds), *The handbook psychology* (Vol. 2). Boston, MA: McGraw Hill.
Foddy, W. (1998). An empirical evaluation on in-depth probes used to pr questions. *Sociological Methods and Research, 27*, 103–133.
Fowler, F. J., Jr. (1993). *Survey research methods* (2nd ed). Newbury Park, CA:

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.
Fowler, F. J., Jr., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In: N. Schwarz & S. Sudman (Eds), *Answering questions: Methodology for determining cognitive and communicative processes in survey research.* San Francisco, CA: Jossey-Bass.
Gerber, E. R., & Wellens, T. R. (1997). Perspectives on pretesting: "Cognition" in the cognitive interview? *Biomedical Mass Spectrometry, 55*, 18–39.
Gerber, E. R., & Wellens, T. R. (1998). The conversational analogy, forms literacy, and pretesting in self-administered questionnaires. Paper presented to the International Sociological Association, Montreal, Canada.
Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*, 1115–1124.
Grunert, S. C., & Muller, T. E. (1996). Measuring values in international settings: Are respondents thinking 'real' life or "ideal" life. *Journal of International Consumer Marketing, 8*, 169–185.
Hahn, E. A., Bode, Du, & Cella (2006). Evaluating linguistic equivalence of patient-reported outcomes in a cancer clinical trial. *Clinical Trials, 3*, 280–290.
Hahn, E. A., Holzner, B., Kemmler, G., Sperner-Unterweger, B., Hudgens, S. A., & Cella, D. (2005). Cross-cultural evaluation of health status using item response theory: FACT-B comparisons between Austrian and U.S. patients with breast cancer. *Evaluation and the Health Professions, 28*, 233–259.
Hamilton, M., & Barton, D. (2000). The international adult literacy survey: What does it really measure? *International Review of Education, 46*, 377–389.
Harkness, J. A. (1999). In pursuit of quality: Issues for cross-national survey research. *International Journal of Social Research Methodology, 2*, 125–140.
Harkness, J. A. (2001). Questionnaire development, adaption, and assessment for the ESS. Paper presented to the International Conference on Quality in Official Statistics, Stockholm.
Harkness, J. A. (2006). Round 3 ESS translation guidelines. *ESS document*, April.
Harkness, J. A., Pennell, B. E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In: S. Presser, M. P. Couper, J. T. Lesser, E. Martin, J. Martin & J. M. Rothgeb, et al. (Eds), *Methods for testing and evaluating survey questionnaires.* New York: Wiley.
Harkness, J. A., Schoua-Glusberg, A. (1998). Questionnaires in translation. In: J. Harkness (Ed.), ZUMA-Nachrichten Spezial No. 3, Cross-Cultural Survey Equivalence, ZUMA, Mannheim.
Heath, A., Fisher, S., & Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science, 8*, 297–333.
Meyer-Zlotnik, J. H. P., & Wolf, C. (2003). *Advances in cross-national comparison: A European working book for demographic and socio-economic variables.* New York: Kluwer Academic.
Niland, J. G., Johnson, T. P., & Wolf, J. G. (1992). A fairly common ambiguity: Comparing rating and approval measures of public opinion. *Sociological Focus, 25*, 261–271.
Noelle, M., & Richter, R. (2001). *Theoretical and methodological concepts for future research and documentation on social reporting in cross-sectional surveys.* EuReporting Working Paper no. 18. Lazarsfeld-Gesellschaft, P. fuer Sozialforschung, Vienna.
Schuman, H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly*, 250–260.

Javeline, D. (1999). Response effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly, 63*, 1–28.

Johnson, T., Harkness, J., Mohler, P. van de Vijver, F., & Ozcan, Y. Z. (2000). 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, Germany.

Johnson, T., Kulesa, P., ISR LLC, Y. I., Cho, & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*, 1–14.

Johnson, T., & Van de Vijver, F. (2003). Social desirability in cross-cultural research. In: J. A. Harkness, F. J. R. Van Der Vijver & P. Ph. Mohler (Eds), *Cross-cultural survey methods*. London: WileyEurope.

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research, In: J. A. Harkness (Ed.), Nachrichten Spezial, Cross-Cultural Survey Equivalence, ZUMA, Mannheim.

Johnson, T. P., & Bowman, P. J. (2003). Cross-cultural sources of measurement error in substance use surveys. *Substance Use & Misuse, 38*, 1441–1483.

Johnson, T. P., O'Rourke, D., Sudman, S., Warnecke, R., Lacey, L., & Horm, J. (1997). Social cognition and responses to survey questions among culturally diverse populations. In: L. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo & N. Schwarz, et al. (Eds), *Survey measurement and process control*. New York: Wiley.

Jowell, R. (1998). How comparative is comparative research? *American Behavioral Scientist, 42*, 168–177.

Jusko, K. L., & Shively, W. P. (2005). Applying a two-step strategy to the analysis of cross-national public opinion data. *Political Analysis, 13*, 327–344.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*, 191–207.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. *International Statistical Review, 62*, 167–186.

Knoop, J. C. (1979). Assessing equivalence of indicators cross-national survey research: Some practical guidelines. *International Review of Sport Sociology, 14*, 137–156.

Krebs, D., & Schuessler, K. F. (1986). Zur Konstruktion von Einstellungsskalen in Internationalen Vergelich. *ZUMA-Arbeitsbericht* No. 86/01.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.

Kumata, H., & Schramm, W. (1956). A pilot study of cross-cultural meaning. *Public Opinion Quarterly, 20*, 229–238.

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology, 90*, 165–187.

Lessler, J. (1984). Measurement error in survey. In: C. F. Turner & E. Martin (Eds), *Surveying subjective phenomena*. New York: Russell Sage.

Levine, R., Gonzalez, R., Weidmer, B., & Gallagher, P. (2004). Cognitive testing of English and Spanish versions of health survey items. Paper presented to the American Association for Public Opinion Research, Phoenix, AZ.

Li, L., Wang, H., & Shen, Y. (2003). Chinese SF-36 health survey: Translation, cultural adaptation, validation, and normalization. *Journal of Epidemiology and Community Health, 57*, 259–263.

Lievesley, D. (2001). The challenge of improving the quality of internationally comparable data. In: *Proceedings of statistics Canada symposium 2001*, Statistics Canada, Ottawa.

Lin, Y., Chen, C., & Chiu, P. (2005). Cross cultural research and back-translation. *Sports Journal, 8*, 1–8.

Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions*. Beverly Hills, CA: Sage.

Lodge, M., & Tursky, B. (1982). The social-psychological scaling of political opinion. In: B. Wegener (Ed.), *Social attitudes and psychophysical measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lynn, P. (2003). Developing quality standards for cross-national survey research: Five approaches. *International Journal of Social Research Methodology, 6*, 323–336.

Lynn, P., Haeder, S. Gabler, & Laaksonen, S. (2004). *Methods for achieving equivalence of samples in cross-national surveys: The European Social Survey experience*. ISER Working Paper no. 2004-09. Institute for Social and Economic Research, University of Essex, Colchester, UK.

MacIntosh, R. (1998). Global attitude measurement: An assessment of the world values survey postmaterialism scale. *American Sociological Review, 63*, 452–464.

McGorry, S. Y. (2000). Measurement in cross-cultural environment: Survey translation issues. *Qualitative Market Research, 3*, 74–81.

Miller, J., Slumczynski, K. M., & Schoenberg, R. (1981). Assessing comparability of measurement in cross-national sociocultural settings. *Social Psychology Quarterly, 44*, 178–191.

Mohler, P. P., Smith, T. W., & Harkness, J. A. (1998). Respondent's ratings of expressions from response scales: A two-country, two-language investigation on equivalence and translation. In: J. A. Harkness (Ed.), Nachrichten Spezial No. 3, Cross-Cultural Survey Equivalence.

Newby, M., Amin, S., Diamond, I., & Naved, R. T. (1998). Survey experience among women in Bangladesh. *American Behavioral Scientist, 42*, 252–275.

Nicoletti, C., & Peracchi, F. (2005). Survey response and survey characteristics: Microlevel evidence from the European community panel. *Journal of the Royal Statistical Society, A*, (168), 763–781.

Nuyts, K., Waege, H., Loosvelts, G., & Bulliet, J. (1997). The application of cognitive interviewing techniques in the development and testing of measurement instruments for survey research. *Tijdschrift voor Sociologie, 18*, 477–500.

Olson, L., Osborn, L., Blumberg, S., & Brady, S. (2003). Collecting data in multiple languages: Development of a methodology. Paper presented to the American Association for Public Opinion Research, Nashville, TN.

Pan, Y. (2004). Cognitive interviews in languages other than English: Methodological and research issues. Paper presented to the American Association for Public Opinion Research, Phoenix, AZ.

Pasick, R. J., Sabogal, F., Bird, J., D'Onofrio, C., Jenkins, C. N. H., Lee, M., et al. (1996). Problems and progress in translation of health survey questions: The pathways experience. *Health Education Quarterly, 23*, 28–40.

Parneger, T. V., Leplege, A., & Etter, J. F. (1999). Cross-cultural adaptation of a psychometric instrument: Two methods compared. *Journal of Clinical Epidemiology, 52*, 1037–1046.

Potaka, L., & Cochrane, S. (2002). *Developing bilingual questionnaires: Experiences from New Zealand in the development of the 2001 Maori language survey*. Unpublished report, New Zealand Statistics.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology, 24*, 73–104.

Sato, A. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43*, 1–14.

Pruefer, J., & Rexroth, M. (1996). Verfahren zur Evaluation von Survey-Fragen: Ein Ueberblick. ZUMA-Arbeitsbericht No. 95/5.
Przeworski, A., & Teune, H. (1966). Equivalence in cross-national research. *Public Opinion Quarterly, 30*, 551–568.
Ryan, A. M., Chan, D., Ployhart, R. E., & Slade, L. A. (1999). Employee attitude surveys in a multinational organization: Considering language and culture in assessing measurement equivalence. *Personnel Psychology, 52*, 37–58.
Salomon, J., Tandon, A., & Murray, C. J. L. (2004). Comparability of self-rated health: Cross-sectional multi-country survey using anchoring vignettes. *BMJ*, on-line at www.bmj.com
Saris, W. (2003). Multi-trait — Multi-method studies. In: J. A. Harkness, F. J. R. Van Der Vijver & P. Ph. Mohler (Eds), *Cross-cultural survey methods*. London: WileyEurope.
Saris, W. E. (1998). The effects of measurement error in cross-cultural research. In: J. A. Harkness (Ed.), Nachrichten Spezial No. 3, Cross-Cultural Survey Equivalence.
Sasaki, M. (1995). Research design of cross-national attitude surveys. *Behaviormetrika, 22*, 99–114.
Schooler, C., Diakite, C., Vogel, J., Mounkoro, P., & Caplan, L. (1998). Conducting a complex sociological survey in rural Mali: Three points of view. *American Behavioral Scientist, 42*, 252–275.
Schuman, H., Steeh, C., Bobo, L., & Krysan, M. (1997). *Racial attitudes in America: Trends and interpretations* (Revised edition). Cambridge, MA: Harvard University Press.
Schwarz, N., Grayson, C., & Knaeuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research, 10*, 177–183.
Schwarz, N., & Hippler, H.-J. (1995). The numeric values of rating scales: A comparison of their impact in mail surveys and telephone interviews. *International Journal of Public Opinion Research, 7*, 72–74.
Skevington, S. M. (2002). Advancing cross-cultural research on quality of life: Observations drawn from the WHOQOL development. *Quality of Life Research, 11*, 135–144.
Smid, M., & Hess, D. (2003). Harmonising sampling frames and indicators in international market research: A German perspective. In: J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds), *Advances in cross-national comparison: A European working book for demographic and socio-economic variables*. New York: Kluwer Academic.
Smith, T. W. (1988). *The ups and downs of cross-national survey research*. GSS Cross-National Report no. 8. NORC, Chicago.
Smith, T. W. (1993). *Little things matter: A sampler of how differences in questionnaire format can affect survey responses*. GSS Methodology Report no. 78. NORC, Chicago.
Smith, T. W. (1994). An analysis of response patterns to the ten-point scalometer. In: *American Statistical Association 1993 proceedings of the section on survey research methods*. Alexandria, VA.
Smith, T. W. (1997). *Improving cross-national survey response by measuring the intensity of response categories*. GSS Cross-National Report no. 17. NORC, Chicago.
Smith, T. W. (2001). Developing nonresponse standards. In: R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds), *Survey nonresponse*. New York: Wiley.
Smith, T. W. (2003). Developing comparable questions in cross-national surveys. In: J. A. Harkness, F. J. R. Van Der Vijver & P. Ph. Mohler (Eds), *Cross-cultural survey methods*. London: WileyEurope.
Smith, T. W. (2004). Developing and evaluating cross-national survey instruments. In: S. Presser, M. P. Couper, J. T. Lesser, E. Martin, J. Martin & J. M. Rothgeb, eds., *Methods for testing and evaluating survey questionnaires*. New York: Wiley.

Smith, T. W. (2005). Total survey error. In: K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. New York: Academic Press.
Smith, T. W. (2006a). Advancing cross-national research in the social sciences: Collaboration and methodological innovation. Paper presented to the EuroScience Open Forum, Munich.
Smith, T. W. (2006b). International standards for market, opinion, and social research. *WAPORNEWS (2nd Quarter)*, (6).
Smith, T. W. (2006c). *Notes on the use of substitution in surveys*. Unpublished NORC report, August.
Smith, T. W. (2007a). *An evaluation of Spanish questions on the 2006 general social survey*. GSS Methodological Report no. 109, NORC, Chicago.
Smith, T. W. (2007b). Survey non-response procedures in cross-national perspective: The 2005 ISSP non-response surveys. *Survey Research Methods, 1*, 21–31.
Smith, T. W. (forthcoming). Codes of ethics and standards in survey research. In: W. Donsbach & M. Traugott (Eds), *Handbook of public opinion research*. London: Sage.
Smith, T. W., Mohler, P. P., Harkness, J., & Onodero, N. (2005). Methods for assessing and calibrating response scales across countries and languages. *Comparative Sociology, 4*, 365–415.
Stoop, I. A. L. (2005). *The hunt for the last respondent: Nonresponse in sample surveys*. The Hague: Social and Cultural Planning Office.
Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In: R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds), *Adapting educational and psychological test for cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum.
Tanzer, N. K., Gittler, G., & Ellis, B. B. (1995). Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychology Assessment, 11*, 170–183.
Thumboo, J., Kok-Yong, F., Machin, D., Chang, S.-P., Soh, C.-H., Leong, K.-H., et al. (2002). Does being bilingual in English and Chinese influence response to quality-of-life scales? *Medical Care, 40*, 105–112.
Tien, F. F. (1999). The application of cognitive interview on survey research: An example of contingent valuation method. IN: *Proceedings of the National Science Council*, Republic of China, No. 9, pp. 555–574.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
Uher, R., & Mohler, P. (2003). Documenting comparative surveys for secondary analysis. In: J. A. Harkness, F. J. R. Van Der Vijver & P. Ph. Mohler (Eds), *Cross-cultural survey methods*. London: WileyEurope.
Usunier, J. C. (1999). *Marketing across cultures* (3rd ed.). New York: Prentice-Hall.
Van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89–99.
Verma, V. (2002). Comparability in international survey statistics. Paper presented to the International Conference on Improving Surveys, Copenhagen.
Werner, O., & Campbell, D. (1970). Translating, working through interpreters, and the problem of decentering. In: R. Naroll & R. Cohen (Eds), *Handbook of cultural anthropology*. New York: American Museum of Natural History.
Willcox, C., Sigelman, L., & Cook, E. (1989). Some like it hot: Individual differences in responses to group feeling thermometers. *Public Opinion Quarterly, 53*(Summer), 246–257.
Willis, G. B., & Schechter, S. (1997). Evaluation of cognitive interviewing techniques: Do the results generalize to the field? *BMS, 55*(June), 40–66.