Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research

TIMOTHY P. JOHNSON

In cross-cultural (and cross-national) survey research, the equivalence of survey questions rivals the importance of their reliability and validity. This paper presents a review of the multiple dimensions of equivalence that must be addressed when conducting comparative survey research. Available methodologies for establishing one or more forms of equivalence are also identified and the strengths and limitations of each approach are examined. It is concluded that multiple methodologies must be implemented in order to insure the cross-cultural equivalence of survey measures.

1. Introduction

In perhaps no other subfield of social science research are issues of methodology and measurement as open to challenge and criticism as when they are applied in cross-cultural and cross-national settings. Indeed, the available protocols for conducting cross-cultural and cross-national survey research (to be subsequently referred to as cross-cultural survey research) would appear to be seriously underdeveloped in comparison to the methodologies available for the conduct of monocultural surveys. A major source of the criticism directed at cross-cultural survey research, in fact, has been the uncritical adaptation of the highly successful techniques developed for monocultural surveys. The simple application of this technology in cross-cultural settings usually and unfortunately makes gross assumptions regarding the equivalence of concepts and measurement. Although this problem is recognized by most practitioners, and many have made serious

attempts to address it, there is currently little consensus regarding how best to establish cross-cultural equivalence when conducting social surveys.

One possible explanation for this absence of methodological consensus, given the sheer quantity of cross-cultural surveys that have been conducted over many decades, is perhaps an even more fundamental lack of agreement regarding the notion of equivalence. As we shall shortly see, researchers concerned with cross-cultural inquiries have conceptualized and cataloged equivalence in numerous ways. It would seem obvious that differing views of what equivalence is would almost certainly lead to variability in the procedures proposed for investigating or establishing it. The purpose of this paper is to present an investigation of these closely-related problems. Specifically, it will review: (1) the concept of equivalence as it has been applied to cross-cultural survey research; and (2) available methodologies that have been proposed and/or previously implemented for the purpose of assessing or implementing one or more forms of cross-cultural equivalence when conducting social surveys.

2. Notions of "Equivalence"

Common sense definitions of the term "equivalent" include: "equal in force, amount, or value;" and "corresponding or virtually identical especially in effect or function" (Webster's Seventh New Collegiate Dictionary, 1965). Perhaps in no field of inquiry, though, has this seemingly elementary concept been assigned as many alternative meanings and disaggregated into as many components as in the field of cross-cultural research. Table 1 presents the results of an investigation of the types of "equivalence" that have been discussed or mentioned in the available literature on cross-cultural research. This review included work representing the disciplines of anthropology, business, communication, demography, economics, market research, political science, psychiatry, psychology, sociology, as well as other professions, and covered work reported over the past 35 years.

As Table 1 indicates, more than 50 specific terms have been used to discuss varieties of equivalence. Some of these have not been well defined. As might be expected, there is also considerable overlap, and many of these alternative labels probably represent "equivalent" concepts (see below). Two of the terms used in this table, *cross-cultural equivalence* (Hui and Triandis, 1985) and *cultural equivalence* (Devins et al., 1997) appear to have been used in a generic sense, referring collectively to all forms of equivalence. They will be used in a similar manner in this review. In addition, although it was not my intention in conducting this review to contribute to the plethora of equivalence labels inhabiting the literature, for purposes of parsimony, the remaining forms of equivalence listed in Table 1 can be subsumed under what can be defined as two fundamental domains of cross-cultural equivalence: interpretive and procedural. These two general domains will be examined in turn.

2.1 Table 1: Types of Equivalence Referenced in the Literature

1. ***Calibration Equivalence*** - Mullen (1995)
2. ***Complete Equivalence*** - Verba et al. (1978)
3. ***Conceptual Equivalence*** - Adams-Esquivel (1991); Elder (1976); Eyton and Neuwirth (1984); Flaherty et al. (1988); Green and White (1976); Hines (1993); Hui and Triandis (1985); Kohn and Slomczynski, 1990; Miller et al. (1981); Mitchell (1973); Narula (1990); Okazaki and Sue (1995); Sears (1961); Sechrest et al. (1972); Singh (1995); Straus (1969); Warwick and Osherson (1973)
4. ***Construct Equivalence*** - Singh (1995); Van de Vijver and Leung (forthcoming)
5. ***Construct Operationalization Equivalence*** - Hui and Triandis (1983)
6. ***Content Equivalence*** - Flaherty et al. (1988)
7. ***Contextual Equivalence*** - Elder (1973)
8. ***Credible Equivalence*** - Teune (1990)
9. ***Criterion Equivalence*** - Flaherty et al. (1988)
10. ***Cross-cultural Equivalence*** - Devins et al. (1997); Hui and Triandis (1985); Hui

et al. (1983)

11. ***Cultural Equivalence*** - Devins et al. (1997)
12. ***Definitional Equivalence*** - Eyton and Neuwirth (1984)
13. ***Direct Equivalence*** - Frey (1970)
14. ***Exact Equivalence*** - Verba et al. (1978)
15. ***Experiential Equivalence*** - Sechrest et al. (1972)
16. ***Factor Equivalence*** - Dressler et al. (1991)
17. ***Factorial Equivalence*** - Singh (1995)
18. ***Formal Equivalence*** - Frey (1970); Marsh (1967); Miller et al. (1985); Mohler et al. (1996)
19. ***Functional Equivalence*** - Alwin et al (1994); Allerbeck (1977); Berry (1969); Braun and Scott (1996); Czudnowski (1976); Frey (1970); Frijda and Jahoda (1966); Green and White (1976); Hui and Triandis (1983; 1985); Marsh (1967);Mitchell (1973); NieBen (1982); Pareek and Rao (1980); Peschar (1982); Scheuch, (1993); Sekaran (1983); Singh (1995); Teune (1990); Van de Vijver and Poortinga (1982); Verba (1969); Verba et al. (1978)
20. ***Grammatical-Syntactical Equivalence*** - Sechrest et al. (1972)
21. ***Indicator Equivalence*** - Kuechler (1987)
22. ***Idiomatic Equivalence*** - Sechrest et al. (1972)
23. ***Instrument Equivalence*** - Frey (1970); Green and White (1976); Singh (1995)
24. ***Item Equivalence*** - Borg (1996); Hui and Triandis (1983; 1985); Mohler et al. (1996)
25. ***Lexical Equivalence*** - Blumer and Warwick (1993); Deutscher (1973); Elder (1973); Warwick and Osherson (1973)
26. ***Linguistic Equivalence*** - Berry et al. (1992); Ellis et al. (1989); Hines (1993); Hulin (1987); Iyengar (1993); Kohn and Slomczynski, 1990; Okazaki and Sue (1995); Prince and Mombour (1967); Sechrest et al. (1972); Warwick and Osherson (1973)

27. ***Literal Equivalence*** - Frijda and Jahoda (1966); Mohler et al (1996)
28. ***Meaning Equivalence*** - Prince and Mombour (1967)
29. ***Measurement Equivalence*** - de Vera (1985); Drasgow and Kanfer (1985); Dressler et al. (1991); Ellis et al. (1989); Green and White (1976); Hui et al. (1983); Iyengar (1993); Leung and Drasgow (1986); Mullen (1995); Poortinga (1989); Singh (1995); Straus (1969)
30. ***Measurement Unit Equivalence*** - Van de Vijver and Leung (1996)
31. ***Metaphorical Equivalence*** - Dunnigan et al. (1993)
32. ***Metric Equivalence*** - Hui and Triandis (1983); Leung and Bond (1989); Mullen (1995); Okazaki and Sue (1995); Straus (1969); Van de Vijver and Leung (1996); Van de Vijver and Poortinga (1982)
33. ***Motivational Equivalence*** - Triandis (1972)
34. ***Operational Equivalence*** - Mohler et al (1996); Narula (1990); Prince and Mombour (1967)
35. ***Psychological Equivalence*** - Eckensberger (1973)
36. ***Psychometric Equivalence*** - Devin et al. (1997); Ellis et al. (1989); Hulin (1987); Van de Vijver and Poortinga (1982)
37. ***Relational Equivalence*** - Ellis et al. (1989)
38. ***Relative Equivalence*** - Frey (1970)
39. ***Response Equivalence*** - Anderson (1967); Frey (1970); Sekaran (1983)
40. ***Scalar Equivalence*** - Hui and Triandis (1983; 1985); Mullen (1995); Van de Vijver and Leung (1996); Van de Vijver and Poortinga (1982)
41. ***Semantic Equivalence*** - Flaherty et al (1988); Kleinman (1987)
42. ***Situational Equivalence*** - Anderson (1967)
43. ***Stimulus Equivalence*** - Anderson (1967); Verba et al. (1978)
44. ***Structural Equivalence*** - Van de Vijver and Leung (1996); Watkins(1989)
45. ***Substantive Equivalence*** - Czudnowski (1976)
46. ***Syntactic Equivalence*** - Kohn and Slomczynski (1990)

47. ***Technical Equivalence*** - Flaherty et al (1988)
48. ***Text Equivalence*** - Alwin et al (1994)
49. ***Theoretical Equivalence*** - Teune (1977; 1990)
50. ***Translation Equivalence*** - Anderson (1967); Berry et al. (1992); Candell and Hulin (1987); Hui and Triandis (1983); Hulin (1987); Mullen (1995)
51. ***Verbal Equivalence*** - Adams-Esquivel (1991)
52. ***Vocabulary Equivalence*** - Sechrest et al. (1972)

2.2 Interpretive Equivalence

Several types of equivalence that have been discussed in the literature are primarily concerned with similarities in how abstract, or latent, concepts are interpreted across cultures. As such, these types are very similar in their emphasis on equivalence of meaning, and will consequently be classified as subtypes of "interpretive" equivalence. One of the more commonly cited forms is *conceptual equivalence*, which Hui and Triandis (1985) would apply to constructs that can be meaningfully discussed within each of the cultures of interest. They identify conceptual equivalence as a necessary condition for making cross-cultural comparisons. Similarly, Okazaki and Sue (1995) associate conceptual equivalence with the degree to which a particular concept has identical meaning within two or more cultural groups.

An emphasis on concordance of meaning also appears to be the central requirement for *functional equivalence*. In discussing this form, Van de Vijver and Poortinga (1982) state that "concepts with functional equivalence are universal in a qualitative, although not necessarily a quantitative sense." Pareek and Rao (1980) also emphasize the commonality of meaning across cultures when discussing functional equivalence, suggesting that it "exists when the behavior in question has developed in response to a problem shared by two or more social/cultural groups, even though the behavior in one society may be superficially quite different from the behavior in another society." Additionally, Singh (1995) argues that functional equivalence exists to the degree that the

concept serves similar functions within each society being investigated. *Definitional equivalence*, as discussed by Eyton and Neuwirth (1984), would appear to have a similar meaning.

Other forms of equivalence that have been discussed in the literature also appear to be primarily concerned with meaning. One of these is *semantic equivalence*, a concept which Flaherty et al. (1988) would apply to survey items that exhibit identical meaning across two or more cultures after translation. Similarly, Prince and Mombour (1967) define questionnaires that have successfully retained their original meaning after translation as having *linguistic equivalence.* Iyengar (1993) uses the same label to describe questionnaires that have validity across two or more languages. *Translation equivalence* (Hui and Triandis, 1983), *meaning equivalence* (Prince and Mombour, 1967), and *contextual equivalence* (Elder, 1973) would also appear to be concerned with similarity of construct interpretation across groups. Similarly, Sechrest et al. (1972) discuss *idiomatic equivalence*, which refers to the equivalence or inequivalence of idiomatic expressions used in survey items across cultural groups. Finally, three other terms that have been put forth by researchers, *experiential equivalence* (Sechrest et al., 1972), *theoretical equivalence* (Teune, 1977) and *substantive equivalence* (Czudnowski, 1976), are concerned with the cross-group similarity of the social processes being investigated.

2.3 Procedural Equivalence

A second form of equivalence that has been discussed at varying levels of detail in the literature is concerned with the measures and procedures used to make cross-cultural comparisons. For purposes of this review, these concepts will be defined as subtypes of "procedural" equivalence. One of these includes forms which focus on cross-cultural consistency of measurement. Among these are *exact equivalence* (Verba et al., 1978), *lexical equivalence (*Warwick and Osherson, 1973), *literal equivalence* (Frijda and

Jahoda, 1966), *verbal equivalence* (Adams-Esquivel, 1991), *vocabulary equivalence* (Sechrest et al., 1972), and perhaps also *indicator equivalence* (Kuechler, 1987) *stimulus equivalence* (Anderson, 1967) and *text equivalence* (Alwin et al., 1994), each of which suggests or implies a strict similarity of question wording across language groups. Related forms of equivalence include *formal equivalence* (Frey, 1970), *instrument equivalence* (Singh, 1995), *item equivalence* (Hui and Triandis, 1985), *measurement equivalence* (Leung and Drasgow, 1986), *psychometric equivalence* (Hulin, 1987), *syntactic equivalence* (Kohn and Slomczynski, 1990), and *grammatical-syntactical equivalence* (Sechrest et al., 1972), each of which emphasize the applicability of mechanically identical procedures across groups. Experienced researchers recognize both the pitfalls of uncritically assuming these forms of equivalence and the difficulties of formally demonstrating their presence. These concepts often represent what Berry (1969) has referred to as an "imposed etic" process, in that survey instruments initially designed for one culture are subsequently adapted in a strict technical sense for use with other cultural groups.

Another set of procedural equivalence concepts are concerned with varying levels of psychometric comparability among cross-cultural samples. *Metric equivalence*, for example, is thought to exist when survey questions exhibit similar statistical properties when measured across varying cultural groups (Hui and Triandis, 1983; Okazaki and Sue, 1995; Straus, 1969; Van de Vijver and Leung, 1996). Even more precisely, *measurement unit equivalence* exists when a measurement scale is identical across groups, but there is no common origin (Van de Vijver and Leung, 1996). When measures also have a common origin across groups, they are considered to have *scalar equivalence* (Hui and Triandis, 1983; 1985; Van de Vijver and Poortinga, 1982; Van de Vijver and Leung, 1996) or *calibration equivalence* (Mullen, 1995). *Structural equivalence* assesses the degree to which survey data collected across cultures produce equal data structures, such as what might be observed using factor analysis and similar procedures (Van de Vijver and Leung, 1996). *Factor equivalece* is also concerned with similarity of data structures,

but only to the degree that equal numbers of factors are identified across cultures via factor analysis. *Factorial equivalence* is concerned with the degree to which factor loadings are similar across cultural groups (Singh, 1995). Finally, *measurement equivalence*, as defined by Singh (1995; although see competing definitions provided by Leung and Drasgow (1986) and Straus, 1969), represents instances in which both factor loadings and error variances are identical across groups. A strict burden of equivalence indeed!

Frey (1970) discusses procedural equivalence from the perspective of the cross-cultural equating of measures. Specifically, he discusses *direct equivalence* as existing when measures can be directly compared across cultural groups without reference to culture-specific criteria. In contrast, *relative equivalence* exists when measures collected across two or more cultures must be standardized in reference to some other norm or criteria before they can be compared. For example, annual income can be reasonably compared across nations, but only after being standardized to one metric.

Another cluster of concepts share a concern with the cross-cultural validation of survey items and/or survey scales. Hui and Triandis (1983), for example, discuss *construct operationalization equivalence* as being a form of construct validity. A measure can be identified as having this type of equivalence to the degree that it exhibits a consistent theoretically-derived pattern of relationships with other variables across the cultural groups being examined. *Construct equivalence* (Singh, 1995) and *relational equivalence* (Ellis et al., 1989) would appear to have much the same meaning. *Criterion equivalence*, in contrast, is concerned with the degree to which a variable is consistently associated with other measures of the same construct across cultural groups (Flaherty et al., 1988). Flaherty et al. (1988) also discuss *content equivalence*, which they identify as being the extent to which the items in a measurement scale adequately represent the theoretical domain of interest within each culture being examined. Eckensberger (1973) assigns a very similar meaning to the term *psychological equivalence*. One additional form is *response equivalence*, which Frey (1970) defines as the degree to which responses

obtained from bilingual persons are similar when expressed in two or more different languages.

Both *situational* (Anderson, 1967) and *technical* (Flaherty et al., 1988) *equivalence* are concerned with the conditions under which surveys are administered. Of primary concern here is that the method of data collection used within each culture produces a similar stimulus. *Motivational equivalence* (Triandis, 1972) reflects an interest in assessing the degree to which respondents from varying cultures have similar motivations for their responses.

Another form of procedural equivalence has been referred to as *operational equivalence*. Although its use by Prince and Mombour (1967) is somewhat vague, Mohler et al. (1996) refer to measures as having operational equivalence if "one can be substituted for the other with no detectable change in statistical analyses."

Finally, without distinguishing between interpretational and procedureal forms of equivalence, Verba et al. (1978) refer to *complete equivalence* as a hypothetical achievement that will never be attainable in practice. In contrast, Teune's (1990) discussion of *credible equivalence* implies that some minimum level of either interpretational or procedural similarity may need to be demonstrated in practice before cross-cultural comparisons can be made.

How are these various types of equivalence established within the context of cross-cultural survey research? Just as there are multiple forms of equivalence with which researchers must be concerned, there are numerous methodological approaches that may be useful for addressing them. It is this issue to which our attention is next directed.

3. Available Methods for Establishing Equivalence

In reviewing available research methodologies for assessing cross-cultural equivalence in survey measurement, it may be useful to utilize the "etic-emic" conceptual model (Berry,

1969; Triandis, 1972) from anthropology and psychology. According to this framework, concepts, ideas and behaviors represented by survey questions can be classified as universal or "etic" to the degree that they are universal, or understood in a consistent manner across cultural and national boundaries (i.e., to the extent that they have interpretive equivalence). In contrast, some ideas and concepts are considered "emic" if they have meaning only to one or a few cultural groups, that is, if they are culture-specific or nation-specific.

Interpretive equivalence can never be established for emic phenomena because they do not have shared meaning across cultures. Some forms of procedural equivalence, ironically, can be obtained for emic phenomena. Survey instruments, for example, may impose identical wording on survey questions that are to be used across cultural groups, even if the concepts represented by those questions are emic to a single group. This, however, would be most appropriately referred to as a pseudoetic application of an emic construct. As mentioned earlier, Berry (1969) would refer to such an application as an "imposed etic" practice. This terminology will be useful throughout the remainder of this review.

The techniques which have been applied to problems of cross-cultural equivalence in survey research have been organized around four specific phases of survey research projects: question development, questionnaire pretesting, data collection, and data analysis (see Table 2). It should be noted that the discussion of each technique is intended to serve as a brief overview and not as a comprehensive presentation. References are provided for readers interested in obtaining additional information regarding any of these approaches.

Table 2: Available Methods for Addressing Equivalence in Cross-Cultural Survey Research

A. ***Question Development Phase***

(1). Expert consultation/collaboration

(2). Ethnographic and other qualitative approaches

(3). "Good" question wording practices

(4). "Good" translation practices

(5). Facet analysis

B. ***Questionnaire Pretesting Phase***

(6). Cognitive interviews/structured probes

(7). Measuring response category intensity

(8). Comparative behavior coding

(9). Compare alternative data collection modes

C. ***Data Collection Phase***

(10). Use multiple indicators

(11). Use both emic and etic questions

(12). Respondent/interviewer matching

D. ***Data Analysis Phase***

(13). Item analysis

(14). Item response theory

(15). Generalizability theory

(16). Confirmatory factor analysis

(17). Multidimensional scaling

(18). Applying statistical controls

(19). Identity-equivalence method

3.1 Question Development Phase

Perhaps the most intuitive method for improving the interpretive equivalence of survey questionnaires is the active participation of experts representative of each culture to be studied. This participation may take a number of forms. Two of the primary ones have been *expert consultation* and *expert collaboration*. Examples of expert consultation include: (a) Straus' (1969) proposal to employ cultural experts as judges for evaluating the appropriateness of specific survey items within their culture; and (b) Henderson et al.'s (1992) recommendation that members of each culture being examined be consulted in order to assure that topics of relevance to them are considered. Berry et al., (1992), Elder (1976) and Okazaki and Sue (1995) have each suggested a similar approach. Flaherty et al. (1988) have made more detailed recommendations for expert consultation, suggesting that such teams should include both content specialists and social scientists from each culture. Such teams would be asked to review the appropriateness of instrument content and data collection methods, and to identify other culture-specific considerations. A team or committee approach to questionnaire translation has also been recommended by several researchers (Adams-Esquivel, 1991; Brislin, 1986; Jones and Kay, 1992; Werner and Campbell, 1970). Although clearly very helpful, consultation is not the same as collaboration and may sometimes carry with it some of the less desirable connotations of "hired-hand" research, such as lack of commitment and status inconsistencies.

Others have emphasized more formal integration of cultural representatives as full research collaborators. Frey (1970), for example, has written that "the basic procedure is to assemble a research group possessing deep familiarity with the nations to be studied and with existing research techniques. This group must agree on the objectives of the research and reach a mutual understanding of its major concepts and hypotheses." More recently, Van de Vijver and Hambleton (1996) have stated that "successful avoidance of ethnocentric tendencies in instruments may require a multicultural, multilingual team

with an expertise in the construct under study." Brislin (1986), Johnson et al. (1996a), Kuechler (1987) and Triandis (1972) have made similar recommendations In the United States, the active collaboration of representatives from all participating cultural groups is now often a requirement for the receipt of research funding from federal agencies. The advantages of this approach for assessing and contributing to interpretive equivalence are clear and there appear to be few disadvantages. However, most of the recommendations cited above tend to emphasize collaboration only during the early hypothesis development and questionnaire design phases of research efforts. At the risk of stating the obvious, it is also important to recognize that collaboration should continue throughout all stages of the research process.

Ethnographic and other qualitative approaches have also been recommended as methods for developing interpretively equivalent survey measures. Marin and Marin (1991), for example, suggest cultural immersion, contact with informants, and familiarity with the available literature as appropriate means of improving cultural awareness prior to study design and question development. Word (1992) has also indicated that, prior to constructing survey instruments, ethnographic research may be useful for achieving a more in-depth understanding of the cognitive processes used by persons in different cultures. While these procedures offer obvious advantages, many researchers unfortunately find them less attractive because they are often time-consuming (Ferketich, Phillips and Verran, 1993). For those without the resources to conduct their own ethnographic inquiries, useful information may nonetheless be obtained from the Human Relations Area Files (HRAF), a large data base that maintains information regarding hundreds of unique social and cultural groups (Barry, 1980; Marsh, 1967).

There are also other less intensive qualitative strategies that may be employed during the development of survey questionnaires. One such approach is the antecedent-consequent method described by Triandis (1977). The method is both simple and powerful. Respondents representing the cultures of interest are asked to contribute phrases to a

series of incomplete sentences in order to complete them. By doing so, they can provide researchers with important insights into cross-cultural similarities and differences in perceptions of both the causes and consequences of various phenomena. Another approach is to ask respondents to perform card sorting tasks. These exercises can provide comparative information regarding how respondents organize and manipulate domains of content information. Johnson et al. (forthcoming), for example, have successfully employed this technique to investigate the social identities of multiracial individuals. Focus groups, of course, are a well-known qualitative technique that can provide additional insights when formulating survey questions for use in cross-cultural surveys (Harari and Beaty, 1990). Other qualitative approaches are discussed by Hines (1993).

Adherence to *"good" question wording practices* is another method that focuses primarily on procedural equivalence. Although there is no consensus on what those best practices might be, Brislin (1973; 1986) has over several decades refined a set of general principles that have received considerable attention. In brief, these include the following (Brislin, 1986):

(1). Use short, simple sentences of less than sixteen words;
(2). Employ the active rather than the passive voice;
(3). Repeat nouns instead of using pronouns;
(4). Avoid metaphors and colloquialisms;
(5). Avoid the subjunctive;
(6). Add sentences to provide context for key ideas;
(7). Avoid adverbs and prepositions telling "where" or "when;"
(8). Avoid possessive forms where possible;
(9). Use specific rather than general terms;
(10). Avoid words indicating vagueness regarding some event or thing;
(11). Use wording that will be familiar to translators; and
(12). Avoid sentences with two different verbs if the verbs suggest two different actions.

Bernard (1988) also provides a basic set of recommendations for the development of survey questions that are to be used cross-culturally.

Suggestions for "good" wording practices that will contribute to successful question translation have also been offered by several other researchers. Scheuch (1993), for example, posits that more abstract concepts have a greater likelihood of producing differences in meaning across languages and should therefore by avoided when possible. Prince and Mombour (1967) warn that "if there is a discrepancy in the frequency of usage of a word in two cultures, the words do not have meaning equivalence for survey purposes" and should also be avoided. In addition, it has been suggested by Warwick and Osherson (1973) that "one of the most effective aids to linguistic equivalence is a research problem that is salient to the cultures involved." The more relevant a concept is to everyday existence within a culture, they posit, the fewer the difficulties of language and translation that will be experienced. McKay et al. (1996) suggest the avoidance of slang terms. They also suggest avoiding modifiers and providing examples designed to increase comprehension, as these may also contribute to cross-cultural differences in interpretation.

Cultural differences in response styles are also a challenge to interpretive equivalence. For example, the well known "courtesy bias" found in many societies (Jones, 1963) suggests that questions that might invite obviously socially desirable responses should be avoided wherever possible. To further combat this problem, Mitchell (1973) has recommended that "moral" words be avoided when preparing survey questions, as they are also likely to encourage socially desirable responses. Inkeles and Smith (1974) suggest that "agree-disagree" response formats be avoided for the same reasons.

Smith (1988) provides several suggestions for improving the equivalence of the response scales used in cross-cultural studies. One of these is to consider the use of numerical scales, which he argues can "reduce problems by providing a universally understood set

of categories that have precise and similar meanings," and avoid the use of vague quantifiers, which are more likely to exhibit cross-cultural differences in interpretation. He acknowledges that this approach is also less than perfect in that numeric scales are often more complex than the simple Likert-type scales they are designed to replace, and that different cultures may vary in the ways they manipulate numeric information. Another approach suggested by Smith (1988) is the use of simple dichotomous response options, which may be less susceptible to misunderstanding than traditional ordinal response scales. Smith (1997) also provides useful recommendations regarding the use of various response options across cultures. For example, he indicates that symmetrical, bipolar scales with a clear middle point will likely be most successful in cross-cultural studies.

Collectively, these recommendations for "good" question wording practices can be expected in many instances to contribute to the interpretive equivalence of survey questions. These approaches, however, do not necessarily rule out equivalence threats associated with cross-cultural differences in the fundamental understanding of the concepts, ideas and/or behaviors being assessed. The emphasis of this approach to similarity of question wording, even "good" question wording, will always insure some degree of procedural equivalence at the risk of failing to achieve interpretive equivalence. Survey researchers will need to recognize that there are likely to be many etic concepts that can nonetheless not be assessed using identical survey questions across any random pair of cultures. In recognition of this, some have advocated the use of open-ended questions as a method of verifying equivalence of meaning across cultures (Verba et al., 1978).

Over the past several decades, effort has also been invested in the development of *"good" translation practices* for survey questionnaires. It has been clear for some time that a simple, unidirectional translation of a survey instrument from a source language into one or more target languages is an unacceptable procedure. A commonly referenced improvement is the back-translation model (Brislin, 1970; 1976; 1986). Although there

are countless variations (see for example: Anderson, 1967; Frey, 1970; Marin and Marin, 1991), the basic procedure calls for a bilingual person to translate a source questionnaire into a target language. A second bilingual person is then asked to translate this version back into the source language without knowledge of the original instrument. The initial and revised versions of the source language version are then compared, discrepancies are identified, and appropriate revisions are made.

Questionnaire translation, however, may be more art than science, and serious disagreements continue to be raised regarding the efficacy of these traditional procedures to which several generations of students have been introduced. Deutscher (1973) has warned that back-translation "can instill a false sense of security by demonstrating a spurious lexical equivalence," at the expense of interpretive equivalence. Reliance on back-translation may be particularly dangerous for researchers unfamiliar with one or more of the target languages, as these procedures are unlikely to provide critical information regarding the issues underlying translation discrepancies. In this regard, back-translation may be appropriately referred to as a "black box" technique (Harkness, 1996). Other concerns, discussed by Brislin, Lonner and Thorndike (1973) include the fact that, due to their varied backgrounds, translators may not always have an adequate awareness of the methodological requirements of cross-cultural translation, or experience with the subject material they are asked to translate. However, Sperber, DeVellis and Boehlecke (1994) have suggested that highly skilled translators may be successful in developing precise translations of poorly-worded survey questions.

Werner and Campbell (1970) have addressed some of these concerns with their proposal for "decentering" questionnaires. They identify two forms of questionnaire translation: symmetrical and asymmetrical. The basic back-translation process described above is an example of asymmetrical (or unicentered) translation because it emphasizes loyalty to a source language questionnaire that remains unchanged and serves as the standard for the development of target language instruments. Symmetrical (or decentered) translation, in

contrast, may involve multiple iterations of translation and back-translation, with each language version being continually refined to bring them into closer concordance of meaning. This "decentering" approach should be more successful in achieving interpretive equivalence compared to simple back-translation alone.

Another potential approach to addressing the problem of interpretive equivalence in translation is a variation of the back-translation procedure described by Anderson (1967). In essence, he recommends employing groups of bilinguals to work independently to develop a number of alternative versions of both the source and target language instruments. Although costly, this approach would produce a pool of alternate versions of each questionnaire item within which the effects of language, translation, and translator personal idiosyncracies would be random. Use of randomly selected question versions from such a pool and/or the use of different versions with randomly selected subsamples of survey respondents, he suggests, may be one method of producing cross-cultural equivalence.

Sperber, Devellis and Boehlecke (1994) have recently contributed a new step into the translation process in which they quantitatively evaluate source and back-translated questionnaire versions by asking substantive experts (in their example, medical students and faculty) to rank the degree to which the two alternative versions in the source language are comparable. Some practical guidelines for translating psychological tests and instruments have also been recently presented by Van de Vijver and Hambleton (1996).

Another recent innovation in translation research is the development and testing of cognitive thinkaloud protocol translation methodologies by Harkness (1996). The purpose of this approach is to supplement other translation procedures with information regarding how translator's interpret their role, how they approach and perform the task of translation, and the types of information they consider when translating survey questionnaires. Harkness (1996) reports an experiment in which traditional back-

translation procedures were compared with a thinkaloud translation protocol. The procedure was found to contribute a considerable amount of useful information above and beyond that obtained from back-translation alone. This approach should be viewed as an important complement to back-translation, in that it can provide important insights into the reasons for disagreements among translation versions that might otherwise be unavailable to monolingual researchers.

Facet analysis (Canter, 1983) is a related technique that has been recently proposed as a method for improving the development of equivalent survey questions in different languages (Borg, 1996). Consistent with the concept of interpretive equivalence, facet analysis enables one to emphasize shared meaning rather than shared stimulus. This methodology may be useful in identifying the dimensions, or facets, of survey questions. By doing so, questions might be "mapped" into equivalent counterparts in another language without reliance on fallible literal translations. Borg (1996) lists several additional advantages of this technique, including the ability to catalog question types, and to model the conceptual structure of survey questions. He also identifies one important limitation of this approach: the fact that the mapping of survey questions can become very complex, technical and abstract. Translators not expert in a particular substantive area may find such mapping sentences of little help. As mentioned earlier, the lack of substantive knowledge on the part of the translator is a general problem when translating survey instruments. Borgs's paper (this volume) provides an empirical example of how facet analysis might be usefully applied to a questionnaire translation problem.

3.2 Questionnaire Pretesting Phase

Several special techniques for pretesting monocultural survey instruments have also been applied to problems of cross-cultural equivalence. One set of these are *structured probes and/or cognitive interviews*. Schuman's (1966) introduction of the random probing

technique in a cross-cultural setting provides an early example of how follow-up questions can be used to identify respondent difficulties with question interpretation. In his example, responses to these open-ended probes were coded according to the degree to which a subject's response was able to correctly predict their substantive answer to the survey question. More recently, Johnson et al. (1996a; 1997) and Krause and Jay (1994) have employed thinkaloud interviews to examine cross-cultural differences in the cognitive processing of survey questions. Although these techniques are often able to provide important qualitative information that can be used to assess the interpretive equivalence of survey items, there is also the danger that they may interfere with or otherwise influence respondent answers to substantive survey questions. While this risk may be small relative to the potential advantages of cognitive interviewing, it should be recognized, particularly when working with cultural groups that may be unfamiliar with this general methodology.

Another pretesting methodology that has only recently been applied in a cross-cultural setting will be labeled here as *measuring response category intensity*. Unlike most of the other techniques reviewed, which focus on the interpretive equivalence of survey questions, this approach focuses on the interpretation of the response scales used to measure respondent attitudes and opinions. The essential procedure involves asking samples of respondents from multiple cultural groups to assign numeric values to the responses of various classification schemes. Mohler et al. (1996) and Smith (1997) have reported a cross-national experiment recently conducted as part of the ISSP (International Social Survey Programme) that compared the strength of meanings assigned by German and U.S. respondents to the various elements of several commonly employed survey response scales. For example, they evaluated 28 potential response options that reflect various degrees of agreement and disagreement. Smith (1997) concludes that this approach is more advantageous than other potential methods, including simple ranking and magnitude estimation, for measuring the strength of response categories.

Nonetheless, it should be noted that this approach relies on the untested assumption that numeric scales are interpreted in an equivalent manner across cultures.

The *behavior coding* of respondent difficulties in the interpretation of survey items has also been applied to cross-cultural research. Johnson et al. (1996b) employed this technique to examine composite variability in difficulties with interpreting health survey questions across four cultural groups in the U.S.: African Americans, Mexican Americans, Puerto Ricans, and non-Hispanic Whites. More than 300 interviews were tape-recorded and subsequently evaluated to identify respondent behaviors and/or statements that could be reasonably classified as problems relevant to question interpretation (e.g., requests for clarification, inadequate answers). Inkeles and Smith (1974), and Kohn and Slomczynski (1990) have also used behavior coding of pretest data to examine question comprehension problems across cultural groups. Comparative behavioral coding appears to have promise as a method for collecting somewhat more objective evidence of differential interpretation problems across cultures. This procedure, however, rests on the often-questionable assumption of cross-cultural similarities in response styles, such as satisficing (Krosnick, 1991) and courtesy bias (Jones, 1963), which may influence respondent expressions and indications of interpretation difficulty.

A final approach to evaluating cross-cultural equivalence during questionnaire pretesting is to examine respondent answers across *alternative data collection modes*.

This approach is recommended by Flaherty et al. (1988) in order to insure technical equivalence across groups. Although it may often be tempting and convenient to do so, of course, it cannot be assumed that all cultures will react to the same survey methods in an identical manner. Aquilino and LoSciuto (1990), for example, have provided evidence that African American, but not White, respondents may be significantly less likely to report drug use during telephone, compared to in-person, interviews. Unfortunately, although findings such as these have important implications for the collection of cross-

cultural survey data, the mode of data collection is often fixed and questions of cultural differences in mode effects are never considered, let alone addressed.

3.3 Data Collection Phase

Many researchers recommend using ***multiple indicators*** to measure each topic examined in cross-cultural surveys (Braun and Scott, 1996; Mitchell, 1973; Okazaki and Sue, 1995; Przeworski and Teune, 1970; Smith, 1988). Although this recommendation is also relevant to monocultural surveys (Elder, 1976), as it can demonstrably improve measurement quality, it is likely to take on added importance in cross-cultural surveys. This is because post-survey data analyses (see next section) may identify some questions that do not perform in an identical manner (for example, do not cluster in a similar fashion) across cultures. One can therefore avoid "placing-all-of-the-eggs-in-one-basket" by developing multiple survey indicators for each construct to be measured. Smith (1988) suggests using at least three indicators of each construct; items that employ different response scales as well as different questions. These recommendations are very reasonable and should be considered even by those researchers who either: (1) do not have the resources to implement any of the other strategies discussed up to this point; or (2) are "certain" that their own research will be graced with interpretational and procedural equivalence without the need to resort to any of these additional methodologies.

Another approach goes beyond the simple collection of multiple indicators by ***including both etic and emic questions*** in the survey instrument. That is, this procedure asks a set of questions that are thought to have universal relevance across the cultures being surveyed, as well as additional sets believed to be relevant only to some cultures or to have unique meanings across all cultures. This alternative follows the recommendations of both Przeworski and Teune (1970) and Triandis (1972), who have presented

methodologies (to be discussed below) for jointly analyzing both types of questions. It is of further interest because it appears to address both interpretive and procedural equivalence by acknowledging that conceptually identical phenomena may be successfully measured across cultures using different instruments. While this is a powerful approach, it poses significant challenges to researchers. As Frijda and Jahoda (1966) observe, developing survey materials that are appropriate for a given culture makes the often questionnable assumption that the researcher has a detailed and intimate understanding of the culture(s) being studied. Some of the collaborative suggestions discussed earlier may help address this important concern. In addition, as Warwick and Osherson (1973) have observed, because this approach recommends that the emic questions be asked of respondents within each culture, respondents may sometimes be asked to answer survey questions that appear irrelevant or even foolish to them. In order to avoid this latter possibility, investigators may sometimes be inclined to exclude important emic questions from the survey instrument, even at the risk of restricting the relevant question content for one or more cultures.

Another data collection procedure that is commonly employed in hopes of approximating procedural equivalence is *respondent-interviewer matching* on one or more demographic characteristics, although primarily race/ethnicity or gender is taken (Couper, 1991; Schaeffer, 1980), or the use of indigenous interviewers (Bloom and Padilla, 1979). These practices are usually implemented with the expectation that respondents will feel more at ease, and be more forthcoming with their answers, when the perceived social distance between themselves and their interviewer is low. Brislin (1986), for instance, has argued that matching will contribute to the minimization of various types of response bias that may result from the uncertainties of cross-cultural communication. Language problems should also be minimized under these conditions. Hanna and Hanna (1966) have stated that in some societies failure to match respondents with similar interviewers will produce data in which we can have no “confidence.” However, there is not universal agreement on the applicability of matching procedures. Ferketich, Phillips and Verran (1993) have

observed that in communities where the need for privacy may be strong, outside or otherwise dissimilar interviewers may be preferred. Others have argued that a highly trained staff of interviewers who are given random interview assignments is the most effective approach to minimizing response bias (Collins, 1980; Freeman and Butler, 1976).

3.4 Data Analysis Phase

The most basic form of data analysis for assessing one or more forms of cross-cultural equivalence is to employ *item analysis* techniques. At a minimum, researchers should examine frequency distributions for obvious indications of variability across groups, such as differing or high proportions of "don't know" responses, which may indicate lack of interpretive equivalence (Frijda and Jahoda, 1966; Smith, 1988). Likewise, an indicator that lacks variability in one culture but not another is in all likelihood representing an emic concept. Frey (1970) suggests that these types of simple psychometric comparisons may identify the "tip-of-the-iceberg," providing warning of a more serious lack of equivalence hidden below the surface. More elaborate forms of item analysis rely on assessments of cross-group differentials in item functioning using analysis of variance (ANOVA) and other bivariate statistical techniques (Van de Vijver and Leung, 199**??**; Devins et al., 1997).

Other preliminary analysis procedures may examine cross-cultural differences in response styles, such as acquiescence, social desirability, and extreme response style, in an effort to assess the degree to which these variables may be influencing responses from each culture. Another approach is to determine if multiple indicators of each construct correlate with one another in a similar manner across cultural groups. Iyengar (1993) suggests that increased similarity in correlation patterns across groups may be an indicator of interpretive equivalence across groups. Comparisons of scale reliabilities across cultural groups is also used as a preliminary method of investigating procedural

equivalence (Devins et al., 1997). Kuechler (1987) takes a somewhat different approach to item analysis, suggesting that a thorough set of within-group analyses should be completed prior to the conduct of cross-cultural comparisons.

Item response theory methodology is a more sophisticated approach to identifying survey questions that do and do not behave in a similar manner across cultures (Leung and Drasgow, 1986). This technique is commonly used by psychologists to identify test items that do not reflect the underlying latent construct purportedly being measured. Several authors have provided useful examples of the application of item response models to assessments of the translation equivalence (Candell and Hulin, 1987; Ellis et al., 1989; Hulin, 1987) and cross-cultural relevance (Hui et al., 1983) of survey scale items. It does so by comparing cross-group item characteristic curves, which represent the conditional probabilities of responding in a given manner to individual questions for various levels of a latent variable represented by a measurement scale. Similar item characteristic curves across cultural groups are interpreted as evidence of similar behavior, and hence equivalence. Several limitations of this approach have been noted, including the very strong assumption that the underlying latent trait represented by the survey items is unidimensional, an assumption that may seldom be realistic (Hulin et al., 1982). In addition, these models require fairly large numbers of items in order to function properly, also often an unrealistic assumption for many survey data sets, and the requirement that all observed variables be measured on a dichotomous scale (Drasgow and Kanfer, 1985). This methodology, however, does have the ability to incorporate both etic and emic questions into cross-cultural measures (de Vera, 1985; Hulin, 1987), and should thus be considered an option whenever practical.

Another analytic method that is used to evaluate the equivalence of translated instruments is based on *generalizability theory* (Van de Vijver and Poortinga, 1982). Using an analysis of variance framework, this procedure can partial out variability in survey responses due to the effects of language, individuals, other variables, and all interactions

(Hulin, 1987). Katerberg et al. (1977) provide an application of generalizability theory to an evaluation of the equivalence of English and Spanish versions of two job attitude measures. A unique advantage of this method is its potential to view cross-cultural equivalence as a relative, rather than an absolute, concept (Van de Vijver and Poortinga, 1982). A potential limitation of generalizability theory models are their reliance on the responses of bilingual respondents, who are asked to complete the survey instrument in each language. This necessary reliance on bilinguals is an obvious concern because they may not be representative of the monolingual populations that many researchers are more interested in generalizing to. This technique also assumes that the bilingual respondents will answer in a similar manner in either language, a questionable assumption at best. Marin et al. (1983), for instance, found that Spanish-English bilinguals use more complex cognitive structures when completing the questionnaire in their native language.

Several techniques have also been used to compare the structural relationships among sets of survey items across two or more cultural groups. One of these is *confirmatory factor analysis*. This procedure was introduced by Joreskog (1971), who described it as a theory-driven tool that could be used to compare simultaneously the factor structure of a set of survey questions across multiple population groups and make assessments of their equivalence through comparisons of large sample chi-square statistics. There are numerous excellent examples of the application of confirmatory factor analysis to equivalence problems in cross-cultural research (Devins et al., 1997; Drasgow and Kanfer, 1985; Kohn et al., 1997; Miller et al., 1981; Watkins, 1989). Singh (1995) describes several increasingly precise levels of procedural equivalence that can be obtained using this technique. Unlike the item response models discussed earlier, confirmatory factor analysis is useful in examining the relatively small numbers of items that might be available to represent a given construct in many survey questionnaires. Another advantage of confirmatory factor analysis is its ability to take full advantage of the information available in ordinal and interval rating scales, unlike item response theory models which require dichotomous data (Drasgow and Kanfer, 1985). Kuechler (1987),

however, correctly observes that confirmatory factor analysis requires a large number of assumptions that information collected using survey methodologies are often unable to meet.

Multidimensional scaling has additionally been employed to compare the structure of survey measures cross-culturally (Allerbeck, 1977). This technique examines the relative proximities among sets of survey measures to identify their underlying structure. In practice, multidimensional scaling often produces findings similar to factor analyses, although the latter technique permits more rigorous comparisons of alternative models (Van de Vijver and Leung, 199**??**). Schwartz and Sagiv (1995) and Braun and Scott (1996) have utilized multidimensional scaling to conduct cross-cultural comparisons of the dimensionality of survey instruments. Hayashi et al. (1992) report cross-cultural comparisons using a similar technique which they refer to as *minimum dimension analysis*.

Other analytic approaches have also been used to establish procedural equivalence between samples when investigating cultural effects in survey research. One basic approach has been to examine the effects of culture after first *applying statistical controls* for other sources of variation that might be confounded with culture, such as socioeconomic status (cf., Johnson et al., 1997). Another strategy has been suggested by Leung (1989) and Van de Vijver and Leung (forthcoming), who have observed that the concept of culture is far too broad and complex to serve as an acceptable explanatory variable. They suggest that the analyses of survey data collected across cultures may be improved if an approach is adopted that replaces the commonly used global indicators of culture, such as race, ethnicity, and country of origin, with more specific measures that represent the qualities or features of various cultures that are believed to account for the cross-group differences of interest. This strategy is known as the "unpackaging" of culture (Whiting, 1976). A related procedure has been demonstrated by Johnson et al. (1996a), who provide empirical examples of how variability in survey question

interpretation may be able to account for cultural differences in self-reported physical and mental health. Poortinga (1989) has referred to this approach as "interpreting equivalence."

One final approach to establishing cross-cultural equivalence when analyzing survey data is what Przeworski and Teune (1970) have referred to as the *identity-equivalence method*. Briefly, this method would include survey instrument items that are thought to be etic across each of the cultures of interest, as well as questions believed to be emic to one or some of the cultures being examined (see above). A subsequent set of statistical analyses using correlation matrices, factor analysis or some other technique, would be used to verify empirically which measures were representing the same construct cross-culturally. Survey questions not identified as etic may nonetheless be valid emic indicators of the construct being examined if they correlate with the etic items within a given culture. The measure of an etic construct may thus be developed using a common set of emic indicators and group-specific sets of emic items. A important feature of this approach is its attempt to reconcile interpretive and procedural equivalence. It should be noted that this procedure is similar to the concept of "etic + emic" analysis outlined by Harry Triandis and colleagues (Davidson et al., 1976; Triandis, 1972; Triandis and Marin, 1983). Kohn and Slomczynski (1990) provide an excellent example of the application of the identity-equivalence method as part of their comparative analyses of the relationship between social structure and personality in Poland and the U.S. Examples of other studies that have employed this technique include Funkhouser (1993), Miller et al. (1985), Przeworski and Teune (1966), and Verba et al. (1978). A disadvantage of this approach is its seeming inability to be used in conjunction with pooled analyses of cross-cultural data sets.

4. Discussion

In addition to the traditional reliability and validity requirements for monocultural survey instruments, researchers conducting cross-cultural survey research have the added concern of equivalence. Indeed, cross-cultural research demands a commitment to the establishment of equivalence that is at least equal to the attention routinely reserved for the problems of reliability and validity. As this review suggests, cross-cultural equivalence has been conceptualized in a multitude of ways, and social scientists have in turn devised a variety of methods for use in hopes of achieving it. Although equivalence has multiple dimensions, there seems to be a natural distinction between interpretive and procedural equivalence. While interpretive equivalence is primarily concerned with the subjective cross-cultural comparability of meaning, procedural equivalence, broadly speaking, refers to the objective development of comparable survey measures across cultural groups. Depending on the research questions of interest, the various dimensions of equivalence represented by these two general labels may take on different levels of importance.

It should also be noted that not all forms of equivalence are necessarily created equal. Whereas most would agree that interpretational equivalence is an absolute requirement, certain forms of procedural equivalence may not always be necessary, or even desirable. Specific forms of procedural equivalence that emphasize pure replication of survey questions across cultures may, for example, be inappropriate in many situations where differing norms or frames of reference may require unique survey measures of the same construct. Nonetheless, many otherwise conscientious researchers prefer working with identically-worded survey questions in cross-cultural studies, even when evidence of poor interpretational equivalence is readily available, because such procedural equivalence facilitates data analysis. Indeed, the challenges that an emphasis on interpretational equivalence can pose for data analysis is likely the main reason why so many cross-cultural studies prefer to emphasize forms of procedural equivalence instead. The general

underdevelopment of cross-cultural survey research methodology mentioned earlier in this paper can probably be attributed to this expediency more than anything else.

Ironically, despite this state of affairs, numerous methods for establishing or assessing one or more forms of cross-cultural equivalence are currently available. The best advice to researchers is probably to employ as many of these techniques as possible and within reason, given that various methodologies may be more appropriate to one specific form of equivalence or another. Several other researchers, including Hui and Triandis (1985), and Van de Vijver and Poortinga (1997), have made similar recommendations. Certainly, efforts to establish cross-cultural equivalence should be made during each phase of survey implementation. Various forms of interpretive equivalence, for example, can be and are more easily assessed during question development and questionnaire pretesting phases, while issues of procedural equivalence tend to predominate during the data collection and analysis stages. One gross indicator of the success researchers have had in establishing cross-cultural equivalence may simply be the number of alternative methods they employed throughout the course of their study to achieve this goal.

Finally, efforts to improve the available tools for developing cross-cultural equivalence should be recognized now as one of the more pressing needs of the survey research community. As the cultural composition of many countries continues to diversify, an ever increasing proportion of all researchers will need to confront issues of equivalence in the conduct of their work. The international research community would be the beneficiary if all graduate programs and survey research centers emphasized the importance of cross-cultural equivalence and encouraged ongoing theoretical and methodological assessments of this fundamental problem.

References

Adams-Esquivel, H. (1991) Conceptual adaptation vs. Back-translation of multilingual instruments: How to increase the accuracy and actionability of multlingual surveys. Paper resented at the annual meeting of the American Association for Public Opinion Research, Pheonix, AZ.

Allerbeck, K.R. (1977) Analysis and inference in cross-national survey research. Pp. 373-402 in A. Szalai and R. Petrella (Eds.) Cross-National Comparative Survey Research: Theory and Practice. Oxford: Pergamon.

Alwin, D.F., Braun, M., Harkness, J. and Scott, J. (1994) Measurement in multi-national surveys. Pp. 26-39 in I. Borg and P. Mohler (Eds.) Trends and Perspectives in Empirical Social Research. Berlin: Walkter de Gruyter.

Anderson, R.B. (1967) On the comparability of meaningful stimuli in cross-cultural research. Sociometry 30: 124-136.

Aquilino, W. and LoSciuto, L. (1990) Effects of interview mode on the validity of drug use surveys." Public Opinion Quarterly 54: 362-395.

Barry, H. (1980) Description and uses of the Human Relations Area Files. Pp. 445-478 in H.C. Triandis and J.W. Berry (Eds.) Handbook of Cross-Cultural Psychology: Methodology, Volume 2. Boston: Allyn and Bacon.

Berry, J.W. (1969) On cross-cultural comparability. International Journal of Psychology 4: 119-128.

Berry, J.W., Poortinga, Y.H., Segall, M.H. and Dasen, P.R. (1992) Cross-Cultural Psychology: Research and Applications. New York: Cambridge University Press.

Bernard, H.R. (1988) Research Methods in Cultural Anthropology. Newbury Park, CA: Sage.

Bloom, D. And Padilla, A.M. (1979) A peer interviewer model in conducting surveys among Mexican-American youth. Journal of Community Psychology 7: 129-136.

Blumer, M. and Warwick, D.P. (1993) Social Research in Developing Countries: Surveys and Censuses in the Third World. London: John Wiley & Sons.

Borg, I. (1996) Using facet theory to control item content in cross-cultural surveys. Paper presented at the International Sociological Association Conference on Social Science Methodology, University of Essex, Colchester, England.

Borg, I. And Shye, S. (1995) Facet Theory: Form and Content. Thousand Oaks, CA: Sage.

Braun, M. and Scott, J. (1996) Data-based procedures for the detection of problems in functional equivalence - or: Is "having a job" the same as "working"? Paper presented at the International Sociological Association Conference on Social Science Methodology, University of Essex, Colchester, England.

Brislin, R.W. (1970) Back-translation for cross-cultural research. Journal of Cross-Cultural Psychology 1: 195-216.

Brislin, R.W. (1976) Translation: Applications and Research. New York: John Wiley.

Brislin, R.W. (1986) The wording and translation of research instruments. Pp. 137-164 in Lonner, W.J. and Berry, J.W. (Eds.) Field Methods in Cross-Cultural Research. Beverly Hills, CA: Sage.

Brislin, R.W., Lonner, W.J. and Thorndike, R.M. (1973) Cross-Cultural Research Methods. New York: John Wiley & Sons.

Candell, G.L. and Hulin, C.L. (1987) Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. Journal of Cross-Cultural Psychology 17: 417-440.

Collins, M. (1980) Interviewer variability: A review of the problem. Journal of he Market Research Society 22: 77-95.

Couper, M.P. (1991) Modeling survey participation at the interviewer level. Proceedings of the Section on Survey Research Methods, American Statistical Association, 98-107.

Czudnowski, M.M. (1976) Comparing Political Behavior. Beverly Hills, CA: Sage.

Davidson, A.R., Jaccard, J.J., Triandis, H.C., Morales, M.L. and Diaz-Guerrero, R. (1976) Cross-cultural model testing: Toward a solution of the etic-emic dilemma. International Journal of Psychology 11: 1-13.

de Vera, M.V. (1985) Establishing cultural relevance and measurement equivalence using emic and etic items. Unpublished dissertation. Urbana, IL: University of Illinois.

Deutscher, I. (1973) Asking questions cross-culturally: Some problems of linguistic comparability. Pp. 163-186 in D.P. Warwick and S. Osherson (Eds.) Comparative Research Methods. Englewood Cliffs, NJ: Prentice-Hall.

Devins, G.M., Beiser, M., Dion, R., Pelletier, L.G. and Edwards, R.G. (1997) Cross-cultural measurements of psychological well-being: The psychometric equivalence of Cantonese, Vietnamese, and Laotian translations of the affect balance scale. American Journal of Public Health 87: 794-799.

Drasgow, F. and Kanfer, R. (1985) Equivalence of psychological measurement in heterogeneous populations. Journal of Applied Psychology 70: 662-680.

Dressler, W.M., Viteri, F.E., Chavez, A., Grell, G.A.C. and Dos Santos, J.E. (1991) Comparative research in social epidemiology: Measurement issues. Ethnicity and Disease 1: 379-393.

Dunnigan T., McNall, M. and Mortimer, J.T. (1993) The problem of metaphorical nonequivalence in cross-cultural survey research. Journal of Cross-Cultural Psychology 24: 344-365.

Eckensberger, L.H. (1973) Methodological issues of cross-cultural research in developmental psychology. Pp. 43-64 in J.R. Nesselroade and H.W. Reese (Eds.) Life-Span Developmental Psychology: Methodological Issues. New York: Academic Press.

Elder, J.W. (1973) Problems of cross-cultural methodology: Instrumentation and interviewing in India. Pp. 119-144 in M. Armer and A.D. Grimshaw (Eds.) Comparative Social Research: Methodological Problems and Strategies. New York: John Wiley & Sons.

Elder, J.W. (1976) Comparative cross-national methodology. Pp. 209-230 in A. Inkeles, J. Coleman and N. Smelser (Eds.), Annual Review of Sociology, Volume 2. Palo Alto, CA: Annual Reviews, Inc.

Ellis, B.B., Minsel, B. and Becker, P. (1989) Evaluation of attitude survey translations: An investigation using item response theory. International Journal of Psychology 24: 665-684.

Eyton, J. And Neuwirth, G. (1984) Cross-cultural validity: Ethnocentrism in health studies with special reference to the Vietnamese. Social Science and Medicine 5: 447-453.

Ferketich, S., Phillips, L. and Verran, J. (1993) Development and administration of a survey instrument for cross-cultural research. Research in Nursing & Health 16: 227-230.

Flaherty, J.A., Gaviria, M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J.A. and Birz, S. (1988) Developing instruments for cross-cultural psychiatric research. Journal of Nervous and Mental Disease 176: 257-263.

Freeman, J. and Butler, E.W. (1976) Some sources of interviewer variance in surveys. Public Opinion Quarterly 40: 79-92.

Frey, F.W. (1970) Cross-cultural survey research in political science. Pp. 173-294 in Holt, R.T. and Turner, J.E. (Eds.) The Methodology of Comparative Research. New York: Free Press.

Frijda, N. and Jahoda, G. (1966) On the scope and methods of cross-cultural research. International Journal of Psychology 1: 109-127.

Funkhouser, G.R. (1993) A self-anchoring instrument and analytical procedure for reducing cultural bias in cross-cultural research. The Journal of Social Psychology 133: 661-673.

Green, R.T. and White, P.D. (1976) Methodological considerations in cross-national consumer research. Journal of International Business Studies 7: 81-87.

Hanna, W.J. and Hanna, J.L. (1966) The problem of ethnicity and factionalism in African survey research. Public Opinion Quarterly 30: 290-294.

Harari, O. and Beaty, D. (1990) On the folly of relying solely on a questionnaire methodology in cross-cultural research. Journal of Managerial Issues 11: 267-281.

Harkness, J. (1996) Cognitive approaches to survey translation. Paper presented at the International Sociological Association Conference on Social Science Methodology, University of Essex, Colchester, England.

Hayashi, C., Suzuki, T. and Sasaki, M. (1992) Data Analysis for Comparative Social Research: International Perspectives. Amsterdam: North-Holland.

Henderson, D.J., Sampselle, C., Mayes, F. and Oakley, D. (1992) Toward culturally sensitive research in a multicultural society. Health Care for Women International 13: 339-350.

Hines, A.M. (1993) Linking qualitative and quantitative methods in cross-cultural survey research: Techniques from cognitive science. American Journal of Community Psychology 21: 729-746.

Hui, C.H., Drasgow, F. and Chang, B.H. (1983) An analysis of the modernity scale: An item response theory approach. Journal of Cross-Cultural Psychology 14: 259-278.

Hui, C.H. and Triandis, H.C. (1983) Multistrategy approach to cross-cultural research: The case of locus of control. Journal of Cross-Cultural Psychology 14: 65-83.

Hui, C.H. and Triandis, H.C. (1985) Measurement in cross-cultural psychology: A review and comparison of strategies. Journal of Cross-Cultural Psychology 156: 131-152.

Hulin, C.L. (1987) A psychometric theory of evaluations of item and scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology 18: 115-142.

Hulin, C.L., Drasgow, R. and Comocar, J. (1982) Applications of item response theory to analysis of attitude scale translations. Journal of Applied Psychology 67: 818-825.

Inkeles, A. and Smith, D.H. (1974) Becoming Modern: Individual Change in Six Developing Countries. Cambridge, MA: Harvard University Press.

Iyengar, S. (1993) Assessing linguistic equivalence in multilingual surveys. Pp. 173-182 in M. Blumer and D.P. Warwick (Eds.) Social Research in Developing Countries: Surveys and Censuses in the Third World. London: John Wiley & Sons.

Johnson, T.P., Jobe, J., O'Rourke, D., Sudman, S., Warnecke, R., Chavez, N., Chapa-Resendez, G. and Golden, P. (forthcoming) Dimensions of identification among multiracial and multiethnic respondents in survey interviews. Evaluation Review.

Johnson, T., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R., Lacey, L. and Horm, J. (1996a) Cultural variations in the interpretation of health survey questions. Pp. 57-62 in Warnecke, R. (Ed.) Health Survey Research Methods Conference Proceedings. DHHS Publication No. (PHS) 96-1013. Hyattsville, MD: National Center for Health Statistics.

Johnson, T.P., O'Rourke, D., Sudman, S., Warnecke, R. and Chavez, N. (1996b) Assessing question comprehension across cultures: Evidence from the United States. Paper presented at the International Sociological Association Conference on Social Science Methodology, University of Essex, Colchester, England.

Johnson, T., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R., Lacey, L. and Horm, J. (1997) Social cognition and responses to survey questions among culturally diverse populations. Pp. 87-113 in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. and Trewim, D. (Eds.) Survey Measurement and Process Quality. New York: John Wiley & Sons.

Jones, E.L. (1963) The courtesy bias in South-East Asian surveys. International Social Science Journal 15: 70-76.

Jones, E.G. and Kay, M. (1992) Instrumentation in cross-cultural research. Nursing Research 41: 186-188.

Joreskog, K.G. (1971) Simultaneous factor analysis in several populations. Psychometrika 36: 409-426.

Katerberg, R., Smith, F.J. and Hoy, S. (1977) Language, time, and person effects on attitude scale translations. Journal of Applied Psychology 62: 385-391.

Kleinman, A. (1987) Anthropology and psychiatry: The role of culture in cross-cultural research on illness. British Journal of Psychiatry 151: 447-454.

Kohn, M.L. and Slomczynski, K.M. (1990) Social Structure and Self-Direction: A Comparative Analysis of the United States and Poland. Cambridge, MA: Basil Blackwell.

Kohn, M.L., Slomczynski, K.M., Janicka, K., Khmelko, V., Mach, B.W., Paniotto, V., Zaborowski, W., Gutierrez, R. and Heyman, C. (1997) Social structure and

personality under conditions of radical social change: A comparative analysis of Poland and Ukraine. American Sociological Review 62: 614-638.

Krause, N.M. and Jay, G.M. (1994) What do global self-rated health items measure? Medical Care 32: 930-942.

Krosnick, J.A. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology 5: 213-236,

Kuechler, M. (1987) The utility of surveys for cross-national research. Social Science Research 16: 229-244.

Leung, K. (1989) Cross-cultural differences: Individual-level vs. Culture-level analysis. International Journal of Psychology 24: 703-719.

Leung, K. and Bond, M.H. (1989) On the empirical identification of dimensions for cross-cultural comparisons. Journal of Cross-Cultural Psychology 20: 133-151.

Leung, K. And Drasgow, F. (1986) Relation between self-esteem and delinquent behavior in three ethnic groups: An application of item response theory. Journal of Cross-Cultural Psychology 17: 151-167.

Marin, G. and Marin, B. (1991) Research with Hispanic Populations. Newbury Park, CA: Sage.

Marin, G., Triandis, H.C., Betancourt, H. And Kashima, Y. (1983) Ethnic affirmation versus social desirability: Explaining discrepancies in bilinguals' responses to a questionnaire. Journal of Cross-Cultural Psychology 14: 173-186.

Marsh, R.M. (1967) Comparative Sociology: A Codification of Cross-Societal Analysis. New York: Harcourt, Brace & World.

McKay, R.B., Breslow, M.J., Sangster, R.L., Gabbard, S.M., Reynolds, R.W., Nakamoto, J.M. and Tarnai, J. (1996) Translating survey questionnaires: Lessons learned. New Directions for Evaluation 70: 93-104.

Miller, J., Slomczynski, K.M. and Kohn, M.L. (1985) Continuity of learning generalization: The effect of job on men's intellective process in the United States and Poland. AJS 91: 593-615.

Miller, J., Slomczynski, K.M. and Schoenberg, R.J. (1981) Assessing comparability of measurement in cross-national research: Authoritarian-conservatism in different sociocultural settings. Social Psychology Quarterly 44: 178-191.

Mitchell, R.E. (1973) Survey materials collected in the developing countries: Sampling, measurement, and interviewing obstacles to intra- and inter-national comparisons. Pp. 204-226 in D.P. Warwick and S. Osherson (Eds.) Comparative Research Methods. Englewood Cliffs, NJ: Prentice-Hall.

Mohler, P., Harkness, J., Smith, T.W. and Davis, J.A. (1996) Calibrating response scales across two languages and cultures. Paper presented at the International Sociological Association Conference on Social Science Methodology, University of Essex, Colchester, England.

Mullen, M.R. (1995) Diagnosing measurement equivalence in cross-national research. Journal of International Business Studies 36: 573-596.

Narula, U. (1990) Practical constraints in social field research in India. Pp. 123-149 in U. Narula and W.B. Pearce (Eds.) Culturs, Politics, and Research Programs: An International Assessment of Practical Problems in Field Research. Hillsdale, NJ: Lawrence Erlbaum Associates.

NieBen, M. (1982) Qualitative aspects in cross-national comparative research and the problem of functional equivalence. Pp. 83-104 in M. Niessen and J. Peschar (Eds.) International Comparative Research. Oxford: Pergamon Press.

Okazaki, S. and Sue, S. (1995) Methodological issues in assessment research with ethnic minorities. Psychological Assessment 7: 367-375.

Pareek, U. and Rao, T.V. (1980) Cross-Cultural surveys and interviewing. Pp. 127-179 in H.C. Triandis and J.W. Berry (Eds.) Handbook of Cross-Cultural Psychology, Volume 2. Boston: Allyn and Bacon.

Peschar, J. (1982) Quantitative aspects in cross-national comparative research: Problems and issues. Pp. 57-81 in M. Niessen and J. Peschar (Eds.) International Comparative Research. Oxford: Pergamon Press.

Poortinga, Y.H. (1989) Equivalence of cross-cultural data: An overview of basic issues. International Journal of Psychology 24: 737-756.

Prince, R. and Mombour, W. (1967) A technique for improving linguistic equivalence in cross-cultural surveys. Journal of Social Psychology 13: 229-237.

Przeworski, A. and Teune, H. (1966) Equivalence in cross-national research. Public Opinion Quarterly 30: 33-43.

Przeworski, A. and Teune, H. (1970) The Logic of Comparative Social Inquiry. New York: John Wiley & Sons.

Schachter, S. (1954) Interpretive and methodological problems of replicated research. Journal of Social Issues 10: 52-60.

Schaeffer, N.C. (1980) Evaluating race-of-interviewer effects in a national survey. Sociological Methods & Research 8: 400-419.

Scheuch, E.K. (1993) The cross-cultural use of sample surveys: Problems of comparability. Historical Social Research 18: 104-138.

Schuman, H. (1966) The random probe: A technique for evaluating the validity of closed questions. American Sociological Review 31: 218-222.

Schwartz, S.H. and Sagiv, L. (1995) Identifying culture-specifics in the content and structure of values. Journal of Cross-Cultural Psychology 26: 92-116.

Sears, R.R. (1961) Transcultural variables and conceptual equivalence. Pp. 445-455 in B. Kaplan (Ed.) Studying Personality Cross-Culturally. Evanston, IL: Row, Peterson and Co.

Sechrest, L., Fay, T.L. and Hafeez Zaidi, S.M. (1972) Problems of translation in cross-cultural research. Journal of Cross-Cultural Psychology 3: 41-56.

Sekaran, U. (1983) Methodological and theoretical issues and advancements in cross-cultural research. Journal of International Business Studies 14: 61-73.

Singh, J. (1995) Measurement issues in cross-national research. Journal of International Business Studies 26: 597-619.

Smith, T.W. (1988) The ups and downs of cross-national survey research. GSS Cross-National Report No. 8. National Opinion Research Center, University of Chicago.

Smith, T.W. (1997) Improving cross-national survey research by measuring the intensity of response categories. GSS Cross-National Report No. 17. National Opinion Research Center, University of Chicago.

Sperber, A.D., DeVellis, R.F. and Boehlecke, B. (1994) Cross-cultural translation: Methodology and Validation. Journal of Cross-Cultural Psychology 25: 501-524.

Straus, M.A. (1969) Phenomenal identity and conceptual equivalence of measurement in cross-national comparative research. Journal of Marriage and the Family 81: 233-239.

Teune, H. (1977) Analysis and interpretation in cross-national survey research. Pp. 95-128 in A. Szalai and R. Petrella (Eds.) Cross-National Comparative Survey Research: Theory and Practice. Oxford: Pergamon.

Teune, H. (1990) Comparing countries: Lessons learned. Pp. 38-62 in Comparative Methodology: Theory and Practice in International Social Research. London: Sage.

Triandis, H.C. (1972) The Analysis of Subjective Culture. New York: John Wiley & Sons.

Triandis, H.C. (1977) Interpersonal Behavior. Monterey, CA: Brooks/Cole.

Triandis, H.C. and Marin, G. (1983) Etic plus emic versus pseudoetic: A test of a basic assumption of contemporary cross-cultural psychology. Journal of Cross-Cultural Psychology 14: 489-500.

Van de Vijver, F. and Hambelton, R.K. (1996) Translating tests: Some practical guidelines. European Psychologist 1: 89-99.

Van de Vijver, F. and Leung, K. (1996) Methods and data analysis of comparative research. Pp. 257-300 in J.W. Berry, Y.H. Poortinga and J. Pandey (Eds.) Handbook of Cross-Cultural Psychology, Second Edition, Volume 1. Chicago: Allyn & Bacon.

Van de Vijver, F. and Leung, K. (Forthcoming) Methods and Data Analysis for Cross-Cultural Research. Thousand Oaks, CA: Sage.

Van de Vijver, F.J.R. and Poortinga, Y.H. (1982) Cross-cultural generalization and universality. Journal of Cross-Cultural Psychology 13: 387-408.

Van de Vijver, F.R. and Poortinga, Y.H. (1997) Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, forthcoming.

Verba, S. (1969) The uses of survey research in the study of comparative politics: Issues and strategies. Pp. 56-106 in S. Rokkan, S. Verba, J. Viet, and E. Almasy (Eds.) Comparative Survey Analysis. Paris: Mouton.

Verba, S., Nie, N.H. and Kim, J.O. (1978) Participation and Political Equality: A Seven-Nation Comparison. London: Cambridge University Press.

Warwick, D.P. and Osherson, S. (1973) Comparative Research Methods. Englewood Cliffs, NJ: Prentice-Hall.

Watkins, D. (1989) The role of confirmatory factor analysis in cross-cultural research. International Journal of Psychology 24: 685--701.

Webster's Seventh New Collegiate Dictionary (1965) Springfield, MA: G. & C. Merriam Co.

Werner, L. And Campbell, D.T. (1970) Translating, working through interpreters and the problem of decentering. Pp 398-420 in R. Naroll and R. Cohen (Eds.) American Handbook of Methods in Cultural Anthropology. Garden City, NY: Natural History Press.

Whiting, B.B. (1976) The problem of the packaged variable. Pp. 303-309 in K. Riegel and J. Meacham (Eds.) The Developing Individual in a Changing World, Volume 1. The Hague: Mouton.

Word, C.O. (1992) Cross-cultural methods for survey research in Black urban areas. Pp. 28-42 in Burlew, A.K.H., Banks, W.C., McAdoo, H.P. and Azibo, D.A. (Eds.) African American Psychology: Theory, Research, and Practice. Newbury Park, CA: Sage.

Towards a Theory of Bias and Equivalence

FONS J. R. VAN DE VIJVER

Bias refers to the presence of nuisance factors in cross-cultural research. Three types of bias are distinguished, depending on whether the nuisance factor is located at the level of the construct (construct bias), the measurement instrument as a whole (method bias) or the items (item bias or differential item functioning). Equivalence refers to the measurement level characteristics that apply to cross-cultural score comparisons; three types of equivalence are defined: construct (identity of constructs across cultures), measurement unit (identity of measurement unit), and scalar equivalence (identity of measurement unit and scale origin). Bias often jeopardizes equivalence. Implications of the occurrence of bias on equivalence are described. Examples of how equivalence can be enhanced in multilingual studies are given.

1. Introduction

Cross-cultural research is a generic name here for all comparative studies that involve either different nation states or different cultural groups within a single country. This kind of research is coming of age. A recent tally of *PsycLit*, an electronic medium publishing summaries of a large number of psychology journals and books, showed that during the last ten years there is a continuous increase of the number of publications dealing with cross-cultural differences (Van de Vijver & Lonner, 1995). Surveys in social sciences will probably reveal the same picture. The increased interest may be related to societal developments. Due to large migration streams, Western countries have become multicultural. For example, in the largest cities in the Netherlands about half of the pupils entering primary school are not native Dutch. In the same vein, it is predicted that by 2020, cities like San Francisco and Los Angeles will have more Hispanic than White Anglo residents. The increased interest may also be fueled by the internationalization of economic life. There are more companies than ever before that operate on an international

market. The booming market of intercultural communication training provides a telling example of this interest.

Although it is reassuring to see the tremendous interest in cross-cultural studies, it is regrettable that there is no generally accepted way of dealing with issues that are specific to cross-cultural research. One can come across empirical studies in which Western instruments have been applied without considering the cultural appropriateness of the measure. There are too many studies in which a test is administered in two cultural groups and in which the only question addressed refers to the difference in average score of the two cultural groups. A comparison of average scores should be preceded by an analysis of the suitability of the instrument. Unless a good theoretical framework is available which can rule out various bias sources, the observation of a significant difference is often open to multiple interpretations such as differential stimulus familiarity (in the case of mental tests) and differential social desirability (on personality and attitude questionnaires). Unfortunately, we do not have well-established and widely adopted practices in cross-cultural search to deal with issues like instrument feasibility and multiple interpretations.

In order to establish such practices we will need to have a theoretical framework that attempts to incorporate aspects that are specific to cross-cultural research. In the present author's view, bias and equivalence are concepts that form the core of such a framework. It will be argued that bias and equivalence are concepts that can guide our plans and actions at all stages of a project, in much the same way as the concepts of validity and reliability underlie many decisions taken in intracultural research. Bias can be viewed as the generic name for all validity-related issues that are specific for cross-cultural research.

In the next section bias and equivalence are defined. The third section links these theoretical concepts to such well-known problems in cross-cultural research as sample incomparability. The fourth section applies this framework to problems encountered in multilingual studies. Conclusions will be drawn in the final section.

2. Bias and Equivalence Defined

The concepts of bias and equivalence have their own history in cross-cultural psychology. *Bias* is related to validity. An instrument is biased if its scores do not have the same psychological meaning across the cultural groups involved; more precisely, an instrument is biased if statements about (similarities and differences of) its scores do not apply in the psychological domain of the scores. For example, individual differences in intelligence test scores may reflect differences in intelligence in a single cultural group, whereas intergroup differences may be largely due to differences in education and test experience. Equivalence has historically become associated with the measurement level at which cross-cultural comparisons can be made. Suppose that in the example of the intelligence test individual differences are measured at ratio level in each cultural group. *Equivalence* refers to the question whether there is any difference in measurement level of within- and between-group comparisons. If the measure is biased against some cultural group, individual differences within a cultural population and across cultural populations are not measured at the same scale.

Three characteristics can be derived from these definitions. First, bias refers to unintended sources of variation that constitute alternative explanations of intergroup differences. If bias is present, cross-cultural score differences are not engendered by the target construct (e.g., intelligence or political affiliation) but by some other characteristic (e.g., social desirability or education). Second, bias and equivalence are not intrinsic to an instrument but characteristics of a specific cross-cultural comparison. Both instrument and sample characteristics will influence the likelihood of occurrence of bias. A questionnaire that can be used to measure political affiliation in, say, France and Germany may be biased in a comparison of France and China. Bias will often increase with the cultural distance to be bridged by the instrument and is also more likely when an instrument shows more cultural saturation. In particular in mental testing much effort has been invested in the development of instruments that can be applied across a wide variety of cultures. Labels used in the past for these tests, such as "culture-free" and "culture-fair"

(e.g., Cattell, 1940; Cattell & Cattell, 1963), sound presumptuous to us; still, the underlying idea that stimulus features can unintentionally and systematically distort observed cross-cultural differences has never been challenged. Finally, bias is a source of systematic variation that is -- at least in principle -- replicable across parallel instruments administered to the same samples.

2.1 Three Types of Bias

Table 1. Types of bias and their description

Type of bias	Description
Construct bias	• dissimilarity of constructs
Method bias	
• Sample bias	• incomparability of samples
• Instrument	• stimulus features that induce cross-cultural differences such as stimulus familiarity
• Administration	• procedural aspects such as communication problems
Item bias	• anomalies at item level such as poor translations

Following Van de Vijver and Leung (1997) three types of bias will be distinguished (cf. Table 1). The first is *construct bias*. It is characterized by dissimilarity of construct across cultures. An example comes from Ho's (1996) work on filial piety in China. The concept refers to the behaviors associated with being a good son or daughter. In Western countries the core of the concept is made up of immaterial aspects such as love and respect; the Chinese concept is broader. In China it is more commonly expected that children play an active role in taking care of their parents once these are unable to support themselves. A

Western-based measure of filial piety will insufficiently cover the Chinese concept while a Chinese questionnaire will be overinclusive according to Western standards; in Embretson's (1983) words, the test will show a poor construct representation. If one is interested in a cross-cultural comparison of constructs that show or are susceptible to construct bias such as filial piety, there is a need to clearly define the behaviors included in the measure.

Method bias is a generic name for all sources of bias emanating from methodological-procedural aspects of a study. The name was coined because in empirical papers most sources of bias meant here are described in the method section. This type of bias can be further subdivided in three subtypes. The first is *sample bias*, subsuming all differences in scores that are related to specific aspects of a sample. Comparability of samples can be a cumbersome issue in cross-cultural comparisons. Two types of sampling schemes are often employed in cross-cultural studies. The first is based on random sampling and aims at securing the results from a single sample to a cultural population at large. The second applies a matched sampling procedure and attempts to control or at least to measure the influence of a potentially confounding variable such as age or education on a target variable. For instance, if one is interested in religious beliefs in different countries, the educational level of the interviewees may be relevant to consider. Sample bias is particularly important to take into account in an examination of culturally highly divergent groups. A random sampling scheme may amount to a comparison of dissimilar groups in terms of background characteristics that are related to instrument scores (e.g., education). On the other hand, a matching procedure may yield atypical samples (e.g., matching Aboriginals and Australians from European descent on education may yield atypical groups in either or both populations). A common way to reduce such sampling problems is the measurement of potentially confounding variables at individual level. In many cases it may be possible to apply statistical procedures to examine the influence of confounding variables such as an analysis of covariance or hierarchical regression procedures (Poortinga & Van de Vijver, 1987).

Instrument bias is the second type of method bias. It is induced by instrument characteristics to which individuals from different cultural groups react in a consistently dissimilar way. Examples are stimulus familiarity (which can influence mental test scores) and differential social desirability or response styles (in personality and attitude measurement). *Administration bias* is triggered by communication problems (e.g., poor mastery of the testing language by one of the parties), interviewer characteristics (e.g., sex and cultural group), or other procedural aspects of the data collection.

Item bias (also known as *Differential Item Functioning*) is the third type of bias. It refers to anomalies of an instrument at item level. Examples are poor translations. Hambleton (1994) gives an example from a Swedish-English comparison of educational achievement: "Where is a bird with webbed feet most likely to live? (a) in the mountains; (b) in the woods; (c) in the sea; (d) in the desert." In the Swedish translation "webbed feet" became "swimming feet," thereby giving a clear cue about the correct answer. Item bias has received much more attention in the literature than construct and method bias. For example, there is a widely accepted, statistically-oriented definition of item bias (e.g., Holland & Wainer, 1993). An item is said to be biased if persons from different cultural groups with the same score on the underlying trait have the same expected score on the item. In other words, persons who are equally dominant (or whatever is measured) and who come from different groups should have the same averages on the item. Equal standing on the underlying trait is usually derived from the total test score.

Numerous techniques have been developed to identify item bias. The most popular technique to date is the Mantel-Haenszel procedure which detects bias in dichotomously scored items (Camilli & Shepard, 1994; Holland & Wainer, 1993). The technique for interval-level scores described here closely follows the rationale of the Mantel-Haenszel procedure. Suppose that a test of dominance consisting of 10 five-point Likert-type items is administered to 400 persons in two countries. An item bias procedure starts with the computation of total test scores (i.e., the sum scores on the 10 items). These range from

10 (10 x 1) to 100 (10 x 10). The extreme scores of 10 and 100 are not taken into account, because by definition persons with these scores have identical response profiles for all

items. The remaining scores are split up into score levels; the number of score levels will be determined by the total sample size; a group size of at least 50 persons in each score group is recommended. An analysis of variance is carried out, with culture and score level group as independent variables and item score as dependent variable. An item is said to be uniformly biased (Mellenbergh, 1982) if the main effect of culture is significant. This implies that for each observed total score level the item is consistently easier or more endorsed in one culture than in another. An item is said to show nonuniform bias if the interaction of score level and culture is significant. In such a case the cross-cultural score differences vary with the observed total test score. In empirical applications, uniform bias is much more common than nonuniform bias.

2.2 Four Types of Equivalence

There is a hierarchical order in the types of equivalence presented here (cf. Table 2). The first refers to the incomparability of constructs across cultures and is labeled *construct inequivalence*; it amounts to "comparing apples and oranges." The other three types show some form of equivalence. The weakest type of equivalence is *construct equivalence*, also known as *functional equivalence* and *structural equivalence*. It occurs when the same

Table 2. Types of equivalence and their description

Type of equivalence	Description
Construct inequivalence	dissimilarity of constructs
Construct equivalence	same construct is measured in each cultural group
Measurement unit equivalence	same scale (measurement unit) with different origins in each cultural group
Scalar equivalence	same scale with same origin in each cultural group

construct has been measured across cultural groups (not necessarily using the same instrument). Construct equivalence is sometimes studied in a comparison of nomological networks across cultures, addressing the question of the construct validity of the measure in each cultural group. Factor analysis is a more frequently employed procedure. In most instances, an exploratory factor analysis is carried out separately in each culture, followed by a target rotation procedure (e.g., Mc Donald, 1985) and the computation of factorial agreement. The target rotation is needed in order to deal with the freedom in rotating factor analytic solutions. So, first the solutions obtained in two cultural groups should be rotated to each other before the agreement can be computed (Van de Vijver and Leung, 1997b, provides an SPSS procedure to carry out the target rotations and compute the agreement index). As an example, Piedmont and Chae (1997) describe the development of a Korean version of a measure of the Big Five personality factors (e.g., McCrae & Costa, 1985), originally developed for the US. In the literature one also finds applications of structural equation modeling to examine construct equivalence. In most cases a confirmatory factor analysis is fitted to the data and the cross-sample stability of the parameters is scrutinized. Taylor and Boeyens (1991), for example, applied confirmatory factor analysis, among other techniques, to study the adequacy of the South African Personality Questionnaire among Blacks and Whites in South Africa.

The third type is *measurement unit equivalence*. We assume here, as below, that the measure is of interval or ratio level in all the cultural populations studied. A measure shows this type of equivalence if the measurement unit is identical across groups while the origins differ. As an example, suppose that temperature is measured using Celsius and Kelvin scales. The measurement units are identical but there is a constant difference (an offset) of 273 degrees of the measures. This type of equivalence will arise if the same instrument has been administered across cultures and method bias (e.g., stimulus familiarity) influences the measure. Individual differences may be measured at ratio level in each group while there is no comparison possible across cultures. Unlike the temperature example, we hardly ever know the offset in measures in the social and behavioral sciences.

In the case of *scalar equivalence* or *full score comparability*, the same interval or ratio level applies to measures in the cultures compared. This is the type of equivalence assumed when averages are compared across cultures, such as in *t* tests and analyses of variance.

3. The Influence of Bias on Equivalence

Bias can be seen as a threat to the validity of cross-cultural studies in that it can lead to inequivalence. The relationship between bias and equivalence is schematically presented in Table 3.

Table 3. Is the level of equivalence affected by bias? (after Van de Vijver & Leung, 1997b)

Type of bias		Level of equivalence		
		Construct	Measurement unit[a]	Scalar[a,b]
Construct bias		yes	yes	yes
Method bias:	uniform	no	no	yes
	nonuniform	no	yes	yes
Item bias:	uniform	no	no	yes
	nonuniform	no	yes	yes

[a]The same measurement unit is assumed in each cultural group;
[b]The same origin is assumed in each cultural group.

There are a few rules underlying the table:

- higher types of equivalence are less robust against bias, for example, scalar equivalence is more susceptible to bias than measurement unit equivalence.
- in terms of actions required for recovery, construct bias is more consequential than are method and item bias;
- nonuniform bias is more consequential than uniform bias because nonuniform bias affects both the origin and the measurement unit of a measure while uniform bias influences merely the origin of the scale.

Scalar equivalence is the strictest type of equivalence, allowing for statements of the type "Culture A has a higher score on propensity F than Culture B." In order to make such strong statements, the absence of any bias is assumed. On the other hand, if one is only interested in the construct equivalence, neither item bias nor method bias will be a threat.

In many empirical applications a choice has to be made whether measurement unit equivalence or scalar equivalence applies. The heated debates about racial differences in intelligence focus on this issue. In the terminology of the present chapter, the debate is about the presence or absence of method bias. In many instances, method bias will lead to an offset in the scales: method bias will induce differences in average scores of cultural groups. Cross-cultural differences in stimulus familiarity, social desirability, and response styles tend to affect many items of an instrument; hence, they will often exert a more or less uniform influence on most or all items of an instrument. From a statistical perspective such an influence may well show up as a significant difference in average scores (e.g., in a *t* test or analysis of variance). Yet, such a cross-cultural difference can be mistakenly interpreted as a real difference on a target construct such as intelligence, while an interpretation in terms of some other characteristic (e.g., educational quality) is more appropriate.

3.1 Example: Multilingual Studies

Multilingual studies are an important area of application of the bias and equivalence issues described above. In most multilingual projects a target instrument is already available that has shown desirable characteristics (reliability and validity) in a particular linguistic group; this instrument is translated for use with other lingustic groups. Studies in which an instrument is simultaneously developed in different languages are less common. Therefore, the present discussion will mainly focus on successive development.

Whereas in the past there has been a tendency to see the linguistic aspects of a translation as the focal area of attention in multilingual studies, there is now a growing awareness that more is involved in the translation of an instrument than rendering text from a source into a target language. In the behavioral sciences, there is rarely much interest in the specific contents of questions and items. Instead, instruments are almost always a means to an end and the operationalizations as expressed in questions and items provide access to underlying constructs, such as political involvement, alienation, and egalitarian commitment. Multilingual studies are often based on the tacit assumption that a careful translation of the instrument will lead to a full transfer of all measurement characteristics such as construct validity and reliability. In the terminology of the present chapter, such a full transfer amounts to an assumption of bias-free measurement and the attainment of the highest level of equivalence possible. The transfer of characteristics from a source-language version to a target language should be empirically scrutinized, since the transfer of the characteristics of the original instrument can be anywhere between absent and complete. In order to maintain the highest level of equivalence possible, the translation and subsequent application of an instrument should be as free of bias as possible. In this, linguistic aspects are important, but not the only ones to be considered. Multilingual studies should focus on validity issues (cf. Bracken & Barona, 1991; Hambleton, 1994; Vallerand, 1989; Van de Vijver & Hambleton, 1996).

In retrospect, it is probably fair to say that the theoretical framework of multilingual studies has become broader in recent times. Recommendations about how to

carry out multilingual studies tended to describe procedures for arriving at accurate translations and provide rules for (in)appropriate item writing, such as the avoidance of the passive and long sentences or the care needed in using referential words such as "his," "her," "this," and "that" because languages differ in their systems of reference. The more recent treatment of multilingual studies from a validity perspective is an acknowledgment of the potential threat of bias and the need to minimize bias in all stages of such a study. A group of researchers recruited from several international psychological associations, headed by Ronald Hambleton (University of Amherst, Massachusetts), recently formulated a set of *guidelines* on how to carry out multilingual studies. Instead of discussing the guidelines (see Hambleton, 1994; Van de Vijver & Hambleton, 1996), I shall briefly present the first two principles which adequately capture the general atmosphere of all guidelines:

> Principle 1. Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.
>
> Principle 2. The amount of overlap in the constructs in the populations of interest should be assessed.

It is characteristic for this approach that central principles of multilingual studies do not relate to linguistic issues but to the reduction of bias and the enhancement of construct validity of the measures.

A multilingual study that is carried out from a validity perspective does not primarily address the question of the translation of an instrument but deals with the question of how to measure the particular construct of the source instrument in the target group, using the characteristics of the latter instrument as much as possible. Such an approach is less direct and more involved than preparing a translation of an instrument; yet it will increase the likelihood that a variety of questions are addressed directly which are answered implicitly, though probably incorrectly, in direct translations, such as:

- Do the items cover the construct in the target group adequately?
- Does the instrument have a format and scoring that is appropriate in the target group?
- Are all items relevant and adequately phrased for the target group?

The broad perspective adopted by validity studies has various implications. A literal translation, quite often seen as the only available option in multilingual studies, is one of the possibilities from a validity perspective. In general, translation studies can apply three strategies depending on the type of bias to be expected. First, when construct bias can be expected to threaten a literal translation of the original measure, the *assembly* of an entirely new instrument may be needed to obtain a good representation of the construct in the new cultural context. A good example can be found in the work by Cheung et al. (1996). These authors argued that Western personality measures do not address all relevant dimensions of the Chinese personality. They developed the Chinese Personality Assessment Inventory. In order to examine construct bias of common Western measures, a pilot study was carried out addressing important characteristics of personality as seen by Chinese subjects. The pilot study pointed to the need to include constructs such as "face" and "harmony." The final version of the inventory has both universal and culture-specific aspects of personality. Their study illustrates various features of an assembly approach towards test development: adequate representation of a local construct instead of cross-cultural comparability (and scalar equivalence) is the aim of the project, thereby maximizing the suitability of the instrument for the local context though precluding the opportunity to compare scores across cultures. Furthermore, assembly studies tend to require huge amounts of resources (time and money).

Adaptations constitute the second type of multilingual study. Some (or even most) stimuli are considered appropriate but as a whole the instrument is not taken to yield an appropriate measure of the target construct. Adaptations amount to the literal translation of some stimuli and, depending on the specific features of the instrument, to adding, changing, or removing other stimuli. Adaptation will be the preferred choice when there

is an incomplete overlap in the behaviors or attitudes associated with a construct. A good example of the adaptation option is the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970). This instrument had been translated into more than 40 languages. Most versions are not literal translations of the English-language original, but are adapted in such a way that the underlying constructs, state and trait anxiety, are measured adequately in each language (e.g., Laux, Glanzmann, Schaffner, & Spielberger, 1981). Another example is the Minnesota Multiphasic Personality Inventory (Dahlstrom, Welsh and Dahlstrom, 1972), which has been adapted to various cultural contexts. The constructs of the tests are broad and various items have a limited applicability outside the US, where the inventory was developed. A Mexican adaptation has been described by Lucio, Reyes-Lagunes, and Scott (1994) and a Chinese adaptation by Cheung (1989).

The statistical analyses of adapted instruments often amount to an examination of the construct validity of the new instrument. For example, Cheung (1989), who adapted the MMPI to China, provides evidence for the validity of the scale by examining its ability to discriminate between normals and patients and by computing profiles for different diagnostic groups. She reported patterns similar to those found in the US.

Due to developments in statistical methods, the opportunities for analysis have been expanded in the last decades. The first important development is item response theory (e.g., Hambleton, & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Molenaar & Fischer, 1995). Scores of subjects can be compared across instruments that are based on partially dissimilar item sets. As a hypothetical example, suppose that a German inventory of 15 items to measure political interest is translated for use in an entirely different political system. Furthermore, let us assume that five items have to be replaced by new items, leaving a common set of 10 items. If the assumptions of item response theory are met, a comparison of scores and even a statistical comparison of means of cultural groups in a *t* test can be obtained. The most relevant assumption will be that the 15 items measure a single latent trait in both groups and that the 10 common items measure the same latent trait in both groups. Statistical tests of the assumptions are

available (Hambleton, & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Molenaar & Fischer, 1995).

The second relevant development has taken place in the area of factor analyses, both exploratory (e.g., Kiers, 1990; Kiers & Ten Berge, 1989) and confirmatory (e.g., Bollen, 1989; Bollen & Long, 1993; Byrne, 1989, 1994). Target rotations are a common way to explore the similarity of factors obtained in an exploratory factor analysis across cultural populations (cf. van de Vijver & Leung, 1997a,b). Because factor analytic solutions can be arbitrarily rotated, solutions obtained in different populations have first to be rotated towards each other (i.e., their agreement has to be maximized) before their correspondence can be assessed. The computation of an agreement of adapted instruments amounts to a factor analysis of all items in each cultural group, thereby allowing that both common and culture-specific items define factors, and a target rotation of the common items. (The culture-specific items are defined as missing values). Van de Vijver and Leung (1997**??**) provide an SPSS procedure to carry out target rotations and compute agreement indices (including the frequently reported Tucker's phi).

The way in which so-called multisample analyses in confirmatory factor analysis deal with test adaptations is somewhat similar. Both common and culture-specific items are utilized to get an adequate factorial representation in each cultural population. The use of multisample procedures in confirmatory factor analysis allows for a fine-grained (i.e., item-level) test of similarities of loadings of common variables across cultural populations. As an example, De Groot, Koot, and Verhulst (1994) examined the cross-cultural stability of the Child Behavior Checklist, a measure of child pathology, in the US and the Netherlands. Most syndromes (factors) were similar across these countries.

Both item response theory and structural equation modeling have enlarged the tools of the cross-cultural researcher in an interesting way; however, the limitations of the techniques can be easily overlooked. Suppose that in our example there were five common and ten culture-specific items. With such a small core of common items, the common and ten culture-specific items. With such a small core of common items, the

culture-specific aspects may describe salient aspects of the construct not covered by the other items; the common core may underrepresent the construct. The poorer the representation will be, the more likely it will become that construct bias endangers the comparability of scores.

By far the most popular option in multilingual studies is *application*. It amounts to the literal translation of the original stimulus material. Translation-backtranslations are often employed to arrive at appropriate translations of stimulus material. In most cases the translator will be hired for his or her linguistic expertise. Such an approach may be inappropriate if method or construct bias jeopardize the equivalence of scores. A so-called committee approach in which persons from different areas of expertise participate is better equipped to deal with the complexities of method and in particular construct bias (cf. Hambleton, 1994).

The literature contains many examples of the application option. Smith, Tisak, Bauman, and Green (1991) studied the equivalence of a translated circadian rhythm questionnaire in English and Japanese. Several discrepancies between the original and translated scales were found. Ellis, Becker, and Kimmel (1993) studied the equivalence of an English-language version of the Trier Personality Inventory and the original German version. Among the 120 items tested, 11 items were found to be biased.

The reason for the popularity of literal translations can be easily appreciated. Compared to the assembly of new instruments or the adaptations of existing ones, applications are cheap and retain all opportunities for scalar equivalence. As can be expected, these advantages are not without costs: applications require the absence of bias. Reading the cross-cultural literature, one cannot escape from the impression that the assumption of the absence of bias is often readily made and that claims about absence of bias are only infrequently substantiated. In the social and behavioral sciences, we are often inclined to work from the assumption that our measures are unobtrusive (Webb, Campbell, & Schwartz, 1966), despite the impressive evidence to the contrary. Thus, in a recently completed meta-analysis of cross-cultural differences in cognitive test

performance, the present author found that commercially available Western tests such as Raven's Colored, Standard, and Advanced Progessive Matrices and the Wechlser intelligence scales for children and for adults yielded consistently larger cross-cultural differences than did locally developed non-Western tests (Van de Vijver, 1997).

Validity Enhancement in Multilingual Studies

Many multilingual studies are designed with the aim to compare scores or score patterns across languages. Such an aim amounts to the attainment of the highest level of measurement equivalence possible. Various measures can be taken to enhance the validity of multilingual studies (cf. Van de Vijver & Tanzer, 1997). Obviously, a listing of the measures cannot be exhaustive and some selection criterion is needed. The present overview provides a small overview of frequently proposed measures. The types of bias that were distinguished previously (construct, method, and item bias) constitute the framework in which the measures will be presented (see Table 4).

There are a few ways in which construct bias can be adequately addressed. In the first, decentering (Werner & Campbell, 1970), an instrument is simultaneously developed in all target languages. Ideally, a team with an expertise in both psychology and linguistics is set up for each language. These teams exchange information about the construct and its associated behaviors or attitudes. Culture-specific aspects, such as problematic wording or the use of particular answer rubrics, are likely to be detected and can be removed. An instrument developed this way will not have the implicit or explicit references to the cultural background of the test developer that are characteristic for many measures in the social and behavioral sciences. An interesting variation to this technique is the so-called 'convergence approach,' in which researchers and cultures are crossed. As an example, an Indian and a German political scientist want to study political interest. Both write an inventory for their own cultural group. The instrument is translated in the other language. Both instruments are then administered in both countries. A comparison

Table 4. Strategies for Identifying and Dealing with Bias in Cross-Cultural Assessment (from Van de Vijver & Tanzer, in press)

Type of bias	Strategies
Construct bias	• decentering (i.e., simultaneously developing the same instrument in several cultures) • convergence approach (i.e., independent within-culture development of instruments and subsequent cross-cultural administration of all instruments)
Construct bias and/or method bias	• use of informants with expertise in local culture and language • use samples of bilingual subjects • use of local surveys (e.g., content analyses of free-response questions) • nonstandard instrument administration (e.g., "thinking aloud") • cross-cultural comparison of nomological networks (e.g., convergent/discriminant validity studies, monotrait--multimethod studies, connotation of key phrases)
Method bias	• extensive training of administrators (e.g., increasing cultural sensitivity) • detailed manual/protocol for administration, scoring, and interpretation • detailed instructions (e.g., with sufficient number of examples and/or exercises) • use of subject and context variables (e.g., educational background) • use of collateral information (e.g., test-taking behavior or test attitudes) • assessment of response styles • use of test-retest, training and/or intervention studies • detailed manual/protocol for administration, scoring, and interpretation • use of test-retest, training and/or intervention studies
Item bias	• judgmental methods of item bias detection (e.g., linguistic and psychological analysis) • psychometric methods of item bias detection (e.g., differential item functioning analysis) • error or distracter analysis • documentation of "spare items" in the test manual which are be equally good measures of the construct as actually used test items

of the results may provide insight into the universal and culture-specific aspects of the instrument.

Another set of measures addresses construct and/or method bias. Examples are the use of bilingual subjects and of local surveys. If there is doubt about the applicability of an instrument, nonstandard administrations (e.g., think aloud protocols) can be an aid in the identification of problematic aspects. Another way of addressing construct and/or method bias is the cross-cultural comparison of nomological networks. Such a comparison attempts to answer the question whether an instrument shows a convergent and discriminant validity that may be expected in each culture. Structural equation modeling provides a data-analytic tool to compare nomological networks across cultures.

The measures that can be taken to reduce method bias are numerous. The general procedure behind most measures is the reduction or measurement of relevant confounding variables. Examples aimed at the reduction of nuisance factors are the extensive training of test administrators/interviewers and the preparation of a detailed protocol for administering, scoring, and interpreting an instrument. When the cultural distances to be bridged by an instrument are large, procedures to reduce the influence of confounding factors may be insufficient. For example, when groups of literate and illiterates are compared, lengthy instructions and well-defined administration guidelines cannot make up for the immense differences in relevant background variables. In such cases, an alternative to reduction may be measurement of the most relevant background variables. The influence of these variables can be assessed in an analysis of covariance or hierarchical regression analysis.

An interesting way to examine method bias is the repeated administration of the same instrument in various cultural groups and the examination of score changes, usually score increments, upon retesting. If subjects with similar scores on the pretest show differential gain patterns, strong evidence for method bias has been obtained. Gain patterns on cognitive tests that are larger in non-Western groups than in Western groups

have been reported (e.g., Kendall, Verster, & Von Mollendorf, 1988). Nkaya, Huteau, and Bonnet (1994) administered Raven's Standard Matrices three times to sixth graders in France and Congo. Under power conditions (i.e., when no time limit was applied) a moderate improvement from the first to the second and no progress from the second to the third administration were observed in both groups. Under timed conditions both groups progressed rapidly from the first to the second; however, only the Congolese pupils progressed from the second to the third session. Such findings retrospectively cast doubt on the score equivalence of the first administration.

Disturbances at item level are commonly detected by either of two procedures. The first is the use of judgmental procedures. A few years ago a committee of Dutch psychologists carried out a content analysis of commonly employed psychological tests; the adequacy of these instruments for individuals whose native tongue is other than Dutch was judged. The committee concluded that ethnocentrism is rampant (Hofstee, 1990). The second procedure to detect item-level disturbances is the use of item bias techniques (which have been described before).

Despite their relevance and widespread use, particularly in the area of educational testing, these techniques are not without their problems. Apart from statistical-technical problems mentioned earlier (such as the need for huge samples), there is a problem of interpretation: expert judgments and item bias procedures are more or less consistently found to be unrelated. Sources of item anomalies as identified by experts such as implicit ethnocentrism are often not flagged as biased by statistical procedures. A recent example is a study by Van Leest (1997) investigating the suitability of two personality questionnaires frequently employed among native Dutch for the selection of migrants in the Netherlands. Experts from minority groups (from the target groups of the study) were asked to judge the instruments. Entirely in line with the Hofstee committee, they found many items inadequate for use among migrants. Statistical procedures also identified many biased items; yet, there was no relationship between the conclusions of the

judgmental and statistical procedures. Furthermore, empirical research has shows that item bias is poorly understood. Item bias is often not at all stable across instruments and samples. Thus, Scheuneman (1987) studied bias in items for American Blacks and Whites on the Graduate Record Examination General Test. Various hypotheses about the influence of formal characteristics on item bias were tested (such as a negative phrasing of item stem, clarity of content, and ordinal position of the correct alternative). Some systematic relationships were found; however, Scheunemann concluded "what emerges most clearly from the study is how little we know about the mechanisms that produce differential performance between black and white examinees" (p. 117). Or in Linn's (1993; 359) words: "The majority of items with large DIF values seem to defy explanation of the kind that can lead to more general principles of sound test development practice".

4. Conclusion

Bias and equivalence are integral elements of each and every cross-cultural study. Bias refers to the absence or presence of nuisance factors while equivalence refers to the implications of bias on the cross-cultural score comparisons to be made. In order to safeguard the highest possible level of equivalence, bias should be scrutinized in each and every stage of an empirical project. Hopefully, a serious concern for bias and equivalence will become a routine consideration in cross-cultural studies, in much the same way as validity and reliability have become standard concepts that have deeply influenced our thinking about to design, administer, score, and interpret test scores. In an era in which cross-cultural encounters are becoming more frequent and cross-cultural research is gaining momentum, it is important to design agreed-upon procedures to carry out such research.

References

Bollen, K.J. (1989): *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K.J. & Long, J.S. (eds.) (1993): *Testing Structural Equation Models*. Newbury Park, CA: Sage.

Bracken, B.A. & Barona, A. (1991): State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International 12*: 119-132.

Byrne, B.M. (1989): *A Primer of LISREL: Basic Applications and Programming for Confirmatory Factor Analytic Models*. New York: Springer.

Byrne, B.M. (1994): *Structural Equation Modelling with EQS and EQS/Windows: Basic Concepts, Applications, and Programming*. Thousand Oaks, CA: Sage.

Camilli, G. & Shepard, L.N. (1994): *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage.

Cattell, R.B. (1940): A culture-free intelligence test, I. *Journal of Educational Psychology 31*: 176-199.

Cattell, R.B. & Cattell, A.K.S. (1963): *Culture Fair Intelligence Test*. Champaign, IL: Institute for Personality and Ability Testing.

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

Dahlstrom, W.G., Welsh, G.S. & Dahlstrom, L.E. (1972): *An MMPI Handbook*. Minneapolis: University of Minnesota Press.

De Groot, A. Koot, H.M. & Verhulst, F.C. (1994): Cross-cultural generalizability of the Child Behavior Checklist cross-informant syndromes. *Psychological Assessment 6*: 225-230.

Ellis, B.B., Becker, P. & Kimmel, H.D. (1993): An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology 24*: 133-148.

Embretson, S.E. (1983): Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin 93*: 179-197.

Hambleton, R.K. (1994): Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment (Bulletin of the International Test Commission) 10*: 229-244.

Hambleton, R.K. & Swaminathan, H. (1985): *Item Response Theory: Principles and Applications*. Dordrecht: Kluwer.

Hambleton, R.K. Swaminathan, H. & Rogers, H.J. (1991): *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Ho, D.Y.F. (1996): Filial piety and its psychological consequences. In: Bond, M.H. (ed.), *Handbook of Chinese Psychology* (pp. 155-165). Hong Kong: Oxford University Press.

Hofstee, W.K.B. (1990). Toepasbaarheid van psychologische tests bij allochtonen. *De Psycholoog 25*: 291-294.

Holland, P.W. & Wainer, H. (eds.) (1993). *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Kendall, I.M., Verster, M.A. & Von Mollendorf, J.W. (1988): Test performance of blacks in South Africa. In: Irvine, S.H. & Berry J.W. (eds.), *Human abilities in cultural context* (pp. 299-339). Cambridge: Cambridge University Press.

Kiers, H.A.L. (1990): *SCA: A Program for Simultaneous Components Analysis*. Groningen: IEC ProGamma.

Kiers, H.A.L. & Ten Berge, J.M.F. (1989): Alternating Least Squares Algorithms for Simultaneous Components Analysis with equal component weight matrices for all populations. *Psychometrika 54*: 467-473.

Laux, L., Glanzmann, P., Schaffner, P. & Spielberger, C.D. (1981): *Das State-Trait Angstinventar. Theoretische Grundlagen und Handanweisung* [The German Adaptation of the State-Trait Anxiety Inventory. Theoretical Background and Manual]. Weinheim, Germany: Beltz Test.

Linn, R. L. (1993): The use of differential item functioning statistics: A discussion of current practice and future implications. In: Holland, P.W. & Wainer, H. (eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Erlbaum.

Lucio, E., Reyes-Lagunes, I. & Scott, R.L. (1994): MMPI-2 for Mexico: Translation and adaptation. *Journal of Personality Assessment 63*: 105-116.

McCrae, R.R. & Costa, P.T. (1985): Updating Norman's "adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology 49,* 710-721.

McDonald, R.P. (1985): *Factor Analysis and Related Methods*. Hillsdale, NJ: Erlbaum.

Mellenbergh, G.J. (1982): Contingency table models for assessing item bias. *Journal of Educational Statistics 7*: 105-118.

Molenaar, I.W. & Fischer, G.H. (eds.) (1995): *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer.

Nkaya, H.N., Huteau, M. & Bonnet, J. (1994): Retest effect on cognitive performance on the Raven-38 Matrices in France and in the Congo. *Perceptual and Motor Skills 78*: 503-510.

Piedmont, R.L. & Chae, J-H. (1997): Cross-cultural generalizability of the five-factor model of personality: Development and validation of the NEO PI-R for Koreans. *Journal of Cross-Cultural Psychology 28*: 131-155.

Poortinga, Y.H. & Van de Vijver, F.J.R. (1987): Explaining cross-cultural differences: Bias analysis and beyond. *Journal of Cross-Cultural Psychology 18*: 259-282.

Scheuneman, J.D. (1987): An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement 24*: 97-118.

Smith, C.S., Tisak, J., Bauman, T. & Green, E. (1991): Psychometric equivalence of a translated circadian rhythm questionnaire: Implications for between- and within-population assessments. *Journal of Applied Psychology 76*: 628-636.

Spielberger, C.D., Gorsuch, R.L. & Lushene, R.E. (1970): *Manual for the State-Trait Anxiety Inventory ("Self-Evaluation Questionnaire")*. Palo Alto, CA: Consulting Psychologists Press.

Taylor, T.R. & Boeyens, J.C. (1991): The comparability of the scores of Blacks and Whites on the South African Personality Questionnaire: An exploratory study. *South-African Journal of Psychology 21*: 1-11.

Vallerand, R.J. (1989): Vers une methodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue francaise. *Canadian Psychology 30*: 662-680.

Van de Vijver, F.J.R. (1997): Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology 28*, 670-709.

Van de Vijver, F.J.R. & Hambleton, R.K. (1996): Translating tests: Some practical guidelines. *European Psychologist 1*: 89-99.

Van de Vijver, F.J.R. & Leung, K. (1997a): Methods and data analysis of comparative research. In: Berry, J.W., Poortinga, Y.H. & Pandey, J. (eds.), *Handbook of Cross-Cultural Psychology* (2nd ed., vol. 1, pp. 257-300). Boston: Allyn & Bacon.

Van de Vijver, F.J.R. & Leung, K. (1997b): *Methods and Data Analysis for Cross-Cultural Research*. Newbury Park, CA: Sage.

Van de Vijver, F.J.R. & Lonner, W. (1995): A bibliometric analysis of the Journal of Cross-Cultural Psychology. *Journal of Cross-Cultural Psychology 26*: 591-602.

Van de Vijver, F.J.R. & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*. (in press)

Van Leest, P.F. (1997): Bias and equivalence research in the Netherlands. *European Review of Applied Psychology*. (in press)

Webb, E.J., Campbell, D.T. & Schwartz, R.D. (1966): *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.

Werner, O. & Campbell, D.T. (1970): Translating, working through interpreters, and the problem of decentering. In: Naroll, R. & Cohen, R. (eds.), *A Handbook of Cultural Anthropology* (pp. 398-419). New York: American Museum of Natural History.

The Effects of Measurement Error in Cross Cultural Research

WILLEM E. SARIS

In survey research many decisions are made in order to design an instrument for data collection. These choices have to do with the formulation of the question, the response categories, the instruction, the sample, the mode of data collection, etc. Each of these choices can lead to different errors (Sudman and Bradburn, 1974; Belson, 1981; Schuman and Presser, 1981; Dijkstra and Van der Zouwen, 1982; Andrews, 1984; Molenaar, 1986; Billiet et al., 1986; Groves, 1989; Alwin and Krosnick, 1991, and Scherpenzeel and Saris, 1997) and consequently to incomparability of results with respect to estimates of correlations and effect parameters across studies and also across countries. It is common knowledge that cross-cultural comparison can only be made if the measurement procedures are completely the same. In this study, we want to argue that this requirement is not enough. We will show that the results can also differ if the same procedures have been used because of differences in measurement errors in the different countries. We therefore propose a procedure to correct for measurement error, in order to make comparisons across countries with respect to correlations and regression coefficients. To correct for measurement error, we have chosen an approach that can be used by every researcher involved in social science research. This in particular is why we advocate this approach, even though, from a methodological point of view, more suitable approaches are available. We avoid using these methods because one purpose of this project is that we want to demonstrate a procedure for the correction of measurement error which can be used in any study, once prior methodological research is done. We begin with a discussion of the problems connected with measurement error in comparative survey research and then we describe the solution we have chosen for these problems. All examples given are based on the satisfaction studies done in the context of a methodological, comparative research project involving 13 language areas.

1. The effect of measurement error

The problem of measurement error in research is quite well known. These errors can bias the correlations between the variables in a study, and as a consequence, bias the estimates of parameters in models (see, for example, Bollen, 1989, chapter 5). In comparative research an extra complication is that the choices of the different instruments might make the results incomparable across countries. Let us give a simple example. In a cross cultural study (Saris et al., 1996), the same respondents were asked repeatedly to indicate their satisfaction with life in general (GLS), and their satisfaction with housing (SH), with their financial situation (SF) and with social contacts (SC). Each time the questions were presented, a different response scale was used. In the Dutch study used as an example here, the questions were presented first with a line-drawing scale and repeated with a 10-point scale in a first interview; in a second interview four weeks later, the questions were presented with a 100-point scale and a 5-point scale (for a more detailed description of the study design, see Scherpenzeel (1996)). It is therefore possible to compare the correlations between these four variables measured, using different scales for the same respondents. In Table 1 the correlation for the 1,599 respondents are presented. The

coefficients of the 5-point scale and 10-point scale measures,[1] presented in Table 1 are polychoric correlation coefficients resulting from calculations with PRELIS 1 (Jöreskog and Sörbom, 1988). In the same way, data were collected in Hungary from a sample of 300 people. Here, however, three instead of four different procedures were used (Münnich, 1996). The correlations estimated in the same way for this study are presented in Table 2.

When it is realised that in each of these tables with correlation matrices, the relationships between the same variables for the same respondents are given, then it is surprising that such large differences in correlations are found between the matrices. One might think that this is related to the different points in time of some of the measures; but even when the time is held constant for the Dutch correlations, comparing the 10-point scale correlations with the line production correlations, and comparing the 5-point scale correlations with the 100-point scale correlations, the differences are still considerable. The correlations of SC with SH and SF in the 100-point matrix, for example, are twice as high as they are in the 5-point matrix, even though they were collected at the same point in time. The Hungarian correlation matrices vary just as much, but these data were all collected in one interview with the same people.

[1] Because some of the measures are categorical in nature, polychoric correlation coefficients were calculated with PRELIS 1 (Jöreskog and Sörbom, 1988) to avoid effects of categorisation of, in principle, continuous variables. The advantage of this type of coefficient is that it provides an estimate of the correlation between the variables correcting for the categorical nature of the observed variables. A categorical measure is defined as a measure with less than 15 categories used. The 100-point measures were treated as continuous when at least 15 numbers were used by the respondents. The graphical line-drawing scale was always continuous.

Table 1. Correlations between four satisfaction variables measured with four different methods obtained from the same respondents at two different points in time in the Netherlands.

					TIME 1			
	GLS	SH	SF	SC	GLS	SH	SF	SC
line production					10-point scale (polychoric corr)			
GLS	1.00				1.00			
SH	.356	1.00			.458	1.00		
SF	.370	.364	1.00		.456	.434	1.00	
SC	.454	.253	.303	1.00	.491	.325	.333	1.00
					TIME 2			
	GLS	SH	SF	SC	GLS	SH	SF	SC
100-point scale					5-point scale (polychoric corr)			
GLS	1.00				1.00			
SH	.570	1.00			.381	1.00		
SF	.544	.529	1.00		.445	.349	1.00	
SC	.644	.515	.518	1.00	.462	.232	.270	1.00

Table 2. The same data collected in Hungary.

	GLS	SH	SF	SC	GLS	SH	SF	SC
10-point scale (polychoric corr)					5-point scale (polychoric corr)			
GLS	1.00				1.00			
SH	.490	1.00			.341	1.00		
SF	.637	.468	1.00		.664	.380	1.00	
SC	.519	.254	.308	1.00	.296	.182	.247	1.00
	GLS	SH	SF	SC				
100-point scale								
GLS	1.00							
SH	.450	1.00						
SF	.614	.460	1.00					
SC	.401	.320	.310	1.00				

These differences between the different methods are clear illustrations of the problem of measurement error in survey research and we do not know what the correct estimates of the correlations between the satisfaction variables are. Since the correlations should be the same, because they represent the correlations between the same variables for the same people, the only explanation for the differences is that the methods produce different error structures, and that these errors have large effects on the correlations and consequently on all the estimates which are derived from these correlations. In this study, these questions were asked several times with different methods, allowing

us to see that such differences exist. In studies where only one method is used, this cannot be seen, but the obtained correlations can be just as incorrect, because they, too, are affected by the typical errors of the specific method used.

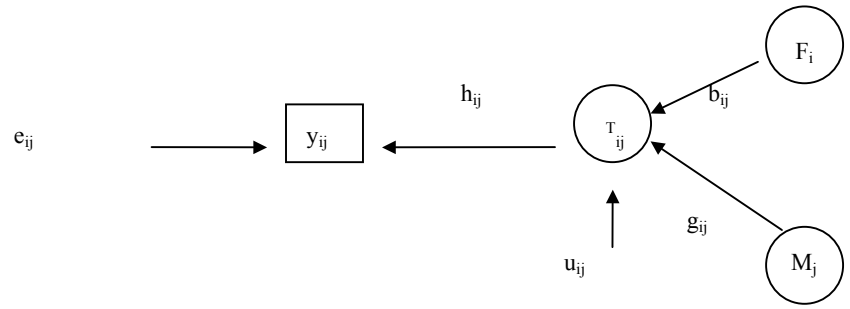
In addition, it is clear that comparisons of correlations across different countries is also very difficult, since even correlations obtained with the same measurement procedures lead to different conclusions. For example, comparing the correlation between GLS and SH for a 10-point scale would lead to the conclusion that the correlation in Hungary is higher. However, looking at the same correlation for the 100-point scale, the Dutch correlation is higher. Many similar examples can be given. These results suggest that even using the same procedure, the conclusions depend on the method which is used. It illustrates that the commonly accepted wisdom that one can only make cross-cultural comparisons if the methods are exactly the same is not, in fact, correct. The equality of the methods is neither a necessary nor sufficient condition for cross-cultural comparability. The reason for this will be clarified in the next section.

2. Explanation of the differences in correlations

Several studies have been published about measurement error and method effects (e.g., Sudman and Bradburn, 1974; Belson, 1981; Schuman and Presser, 1981; Dijkstra and Van der Zouwen, 1982; Andrews, 1984; Molenaar, 1986; Billiet et al., 1986; and Alwin and Krosnick, 1991). The approach suggested by Andrews for estimating the size of the effects of the errors and the procedure to correct for them is discussed in this paper. We have chosen this approach because it is the most explicit and general one of the different procedures introduced by these researchers. It provides all researchers, after a specialised methodological study, with information to make different measurement instruments comparable within a study and across studies. To be able to describe this approach, we first have to formulate the problem of measurement error in a more formal way. For this we use the formulation given in a publication of Saris and Andrews (1991) and Saris and Münnich (1995). In these studies, the authors suggest the path model presented in Figure 1 as a summary of their idea.

Figure 1.

A model for the response on a question incorporating method effects, unique components, and random error.



In a more formal way this idea can be formulated as follows: The responses y_{ij} on item i using method j, can be decomposed into a stable component T_{ij}, which is called the "true score" in classical test theory (Heise and Bohrnstedt, 1970; Lord and Novick, 1968) and a random error component e_{ij}. If the response variable and the variable representing the stable part are standardised, we get equation (1):

$$y_{ij} = h_{ij}\, T_{ij} + e_{ij} \qquad (1)$$

where h_{ij} represents the strength of the relationship between the stable component, or true score, and the response. The true score can further be decomposed into a component representing the score on the variable of interest, F_i, a component due to the method used, M_j, and a unique component due to the combination of method and trait, u_{ij}. After standardisation, this leads to the formulation of equation (2):

$$T_{ij} = b_{ij}F_i + g_{ij}M_j + u_{ij} \qquad (2)$$

where b_{ij} represents the strength of the relationship between the latent variable of interest and the true score and g_{ij} indicates the method effect on the true score. All variables are standardised, except for the disturbance variables. Furthermore, we assume, as is normally done, that the correlations between the disturbance variables and the explanatory variables in each equation and across equations is zero, and we assume that the method and trait factors are uncorrelated.

If all variables except the disturbance terms are standardised, the coefficients h_{ij}, b_{ij} and g_{ij} indicate the strength of the relationships between the variables in the model, and these coefficients have been given a special interpretation:

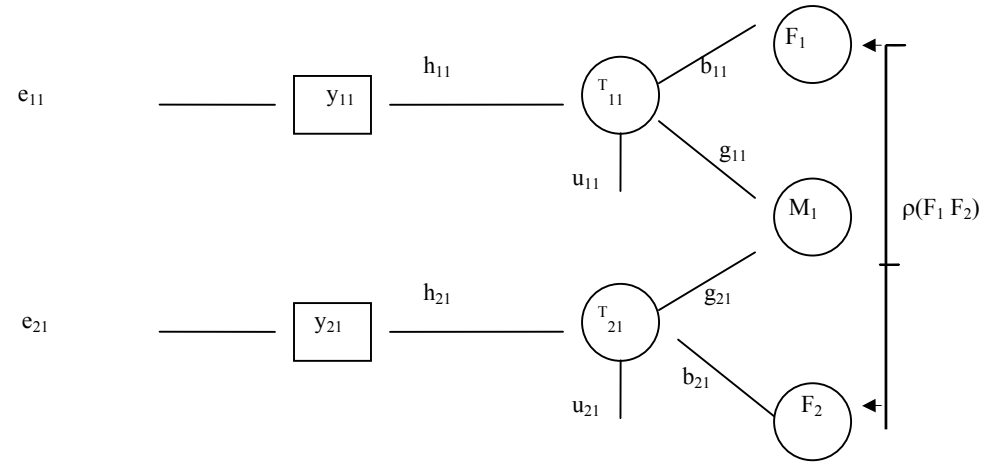
- h_{ij} is called the "reliability coefficient". The square of this coefficient is an estimate of the test-retest reliability in the sense of classical test theory (Heise and Bohrnstedt, 1970; Lord and Novick, 1968).
- b_{ij} is called the "true score validity coefficient" because the square of this coefficient is the explained variance in the true score due to the variable of interest.
- g_{ij} is called the "method effect" because the square of this coefficient is the explained variance in the true score due to the method used.
- The variance of u_{ij} plus g_{ij}^2 is sometimes called the "invalidity", because it is the variance explained in the true score which is not due to the variable of interest (Heise and Bohrnstedt, 1970).

It can be seen that with this information, the total measurement error in the responses (y_{ij}) can be decomposed into a random component ($var(e_{ij})$) and a systematic component ($var(u_{ij}) + g_{ij}^2$).
With this notation and simple path analysis, we can demonstrate all possible effects of measurement error on the correlations and effect parameters.

Figure 2 illustrates the effect of measurement error on the correlations. The only difference between Figure 1 and Figure 2 is that now two variables are studied at the same time and that we assume that these two variables are correlated. This correlation is denoted by $\rho(F_1\ F_2)$. It is assumed that the same method is used for each variable but that the method factor is uncorrelated with the trait factors. The disturbance variables are assumed to be uncorrelated with each other and the factors. All other assumptions made for the model in Figure 1 also hold, and the parameters have the same meaning as described before.

Figure 2.

A model for two correlated variables incorporating method effects, unique components, and random error.



Path analysis suggests that the correlation between the observed variables, denoted by $r(y_{11}y_{21})$ is equal to the correlation produced by F_1 and F_2 and the spurious relationship due to the method-specific variation in the observed variables. This result is specified in (3):

$$r(y_{11}y_{21}) = h_{11}*b_{11}*\rho(F_1\ F_2)*\ b_{21}*h_{21} + h_{11}*g_{11}*g_{21}*h_{21} \qquad (3)$$

Since the validity coefficients and the reliability coefficients are maximally 1, it follows from (4) that $r(y_{11}y_{21}) = \rho(F_1\ F_2)$ only if the reliability and validity are maximal and the method effect is zero. A situation like this is extremely unlikely. Therefore, the two correlations will in general be different. Since the effects of reliability, validity, and method differ from method to method, this might be the explanation for the differences in correlations found between the different methods in Table 1 and 2. The reader can easily check for him/her self that any correlation between the factors of interest can produce very different correlations, depending on the size of the validity and reliability coefficients and method effects. This variation makes it impossible to compare correlations obtained in different studies.

3. An empirical illustration

The International Research group for Methodological and Comparative Survey research (IRMCS) has done a number of projects to estimate these quality indicators for survey instruments in general. A description of the approach can be found in Saris and Münnich (1995) and Scherpenzeel and Saris (1997). An application of the approach on life satisfaction research can be found in Saris et al. (1996). For this project, in each language area, a study was carried out to obtain estimates of reliability, validity, and method effects for that country. After that, a meta-analysis was made in order to study the effects of the different characteristics of the instruments used on the validity and reliability of the instruments (for details, see Scherpenzeel, 1995). In Table 3, the results of the Scherpenzeel (1995) study are summarised.

In the first row of this table, the overall mean validity and reliability coefficients for satisfaction measures can be found. In the other rows of the table, the adjustments for this expected value are specified for different data collection situations. In each row, the adjustment for a different specific study characteristic is mentioned (for a fuller description of the table, see Scherpenzeel (1995; 64-68). It can be seen that a large variety of characteristics has been taken into account, such as the specific trait studied, the scale, the method of data collection, the position in the questionnaire, and some factors which have to do with the design of the study, such as whether an instrument is used alone or in combination with others, what the position of the instrument was in the sequence of methods used and the country in which the data collection took place.

Table 3. Meta-analysis of life satisfaction data across countries.

	N measures	Validity Coefficient Mean = .940 Multivariate Deviations	Reliability Coefficient Mean = .911 Multivariate Deviations
SATISFACTION DOMAIN			
Life in general	54	-.006	-.038
House	54	.005	.029
Finances	54	.003	.020
Social contacts	54	-.001	-.011
RESPONSE SCALE			
100 p. number scale	64	-.021	-.027
10 p. number scale	72	.011	.051
5/4 p. category cale	72	-.022	-.026
graphical line scale	8	.058	-.007
DATA COLLECTION			
Face-to-face interview	96	.011	.012
Telephone interview	52	.002	-.051
Mail questionnaire	40	-.014	-.011
Tele-interview	28	-.022	.067
POSITION			
1 - 5 48		.011	.026
6 - 45	68	.017	-.001
50+ 100		-.017	-.012
TIME BETWEEN REPETITIONS			
alone in interview	32	.010	-.071
first/last 5-20 minutes	64	.017	.063
first/last 30- 60 minutes	80	-.021	-.023
middle, 5-20 minutes	16	.043	.028
middle, 30-60 minutes	24	-.017	-.016
ORDER OF PRESENTATION			
first measurement	60	-.015	-.025
repetition	156	.006	.010
COUNTRY			
Slovenia	12	.020	-.013
Germany	16	.007	.028
Catalonia (Spain)	12	-.039	-.022
Italy 12		.013	.043
Flanders (Belg)+ Netherlands	64	-.028	-.039
Wallonia (Belgium)	12	-.026	-.028
Brussels (Belgium)	12	.006	.000
Sweden	12	.023	.099
Hungary	12	.050	.046
Norway	16	-.018	.031
Russians (Russia)	12	.043	.004
Tatarians (Russia)	12	.033	.003
Other nationalities in Russia	12	.039	.000

The last point is of special interest to us. This study suggests, for example, that on average the validity will be .94, but depending on the chosen instruments, this quality indicator will be higher or lower.

Table 4. Prediction of the validity and reliability of a measure in the Dutch study, on the basis of the instrument characteristics.

	Validity coefficient	Reliability coefficient
	Mean = .940	Mean = .911
Adjustments for:		
Domain: GLS	-.006	-.038
10-point scale	+.011	+.051
Data collection by mail	-.014	-.011
Position 6-45	+.017	-.001
Design time: alone	+.010	-.071
Design order: first	-.015	-.025
The Netherlands	-.028	-.039
Sum	.915	.777

Table 5. Prediction of the validity and reliability of a measure in a Hungarian study, on the basis of the instrument characteristics.

	Validity coefficient	Reliability coefficient
	Mean = .940	Mean = .911
Adjustments for:		
Domain: GLS	-.006	-.038
10-point scale	+.011	+.051
Data collection by mail	-.014	-.011
Position 6-45	+.017	-.001
Design time: alone	+.010	-.071
Design order: first	-.015	-.025
Hungary	+.050	+.046
Sum	.993	.862

On the other hand, even if the instruments are identical in two countries, the validity can be different due to country-specific differences. For example, the validity in Slovenia will on average be .02 higher than the mean, while in Catalonia the validity on average will be .039 lower. Similar effects can be found for other countries and for reliability. This suggests that the quality of the data differs from country to country, even if they use the same data collection procedure. We illustrate this important point below. Any researcher who has one measure of a satisfaction variable can determine the quality of this measure on the basis of the results presented in Table 3. For example, if we say GLS was measured by mail using a 10-point scale at the beginning of the interview in the Netherlands and in Hungary, we can estimate the validity and reliability coefficients with the information from Table 3, as shown in Table 4 and 5. By adding up all the adjustments to the mean value, we obtain an estimate of the validity and reliability coefficient for this variable. For the Dutch study the result is presented in Table 4, for Hungary, in Table 5.

The tables indicate that even if the same instruments are used in both countries for measurement of satisfaction, large differences in results are found for the Netherlands and Hungary. In the same way, these two coefficients can be estimated for the other traits and other methods. For the Dutch and Hungarian study, the results of these calculations for all satisfaction traits using the same method (10-point scale) are presented in Table 6.

Table 6. Quality estimates of the indicators in the MTMM study, predicted on the basis of the meta-analysis for the Netherlands and Hungary.

		Validity		Reliability		Method Effect	
		NL	H	NL	H	NL	H
10-point scale							
	GLS	.92	.99	.78	.86	.39	.14
	SH	.93	1.0	.85	.93	.37	.00
	SF	.93	1.0	.84	.92	.37	.00
	SC	.92	.99	.81	.89	.39	.14

In Table 6, the method effects are also included. This effect can easily be calculated from the information on the validity coefficient, because the method variance should be 1- b^2_{ij} if the unique

variance is zero[2]. So the estimate of the method effect parameter is the square root of the method variance, or:

$$g_{ij} = \sqrt{(1 - b^2_{ij})} \qquad (4)$$

If the measurement procedure indicated above is used, in both countries the reliability, validity and method effects for both variables will be different, as demonstrated above. Using equation 3, it can be shown that in that case the correlation will also be different, although the correlation was the same between the variables of interest. For instance, if we take a correlation of .8 for the variables GLS and SH, for the Netherlands we would get:

r (R1,R2) = .78*.92*(.8)*.93*.85 + .78*.39*.37*.85 = .45 + .096 = .55

In the same way, for Hungary we would get:

r (R1, R2) = .86*.99*(.8)*1.0*.93 + .86*.14*.00*.93 = .63 + .00 = .63

First of all we see that the resulting correlation is much lower in both cases than the correlation between the variables of interest due to the relatively low reliability. In addition, we see a difference of .08 between the resulting correlation in the two countries, even though the correlations between the theoretical variables in both countries were identical. This difference has no substantive meaning, it is only due to the difference in quality of the measurement procedures in the two countries. It seems that the Hungarian public, somehow less bothered by questionnaires, gives better answers to the same questions than Dutch respondents do.

This result indicates that comparisons between correlations from different countries cannot be made without correction for measurement error. How these corrections can be made is the subject of the next section.

[2] This assumption is necessary for identification of the model. This assumption is realistic if in the experiment exactly the same question is used combined with each method. For details we refer to Saris (1990).

4. Correction for measurement error

Now we will concentrate on the correction for the effect of the specific method on the obtained correlation. In other words, we are interested in the correlation between the latent factors, and not in the correlations between the observed variables. To derive these correlations, we have to express the correlation between the factors in the observed correlations and the different validity's, reliability's and method effects. This expression follows immediately from equation (3):

$$\rho(F_1 F_2) = [r(y_{11}y_{21}) - (h_{11}*g_{11}*g_{21}*h_{21})] / (h_{11}*b_{11}*b_{21}*h_{21}) \quad (5)$$

This result suggests that the correlation between the factors can be estimated simply from the observed correlation if estimates for the validity and reliability coefficients and the method effects are known. Table 3 above provides the information from which the reliability, validity and method effects for different measurement instruments can be derived. These results can be used, as before, to estimate the correlation between the variables of interest corrected for measurement error. This could be done by hand, but it is also possible to use programs like LISREL (Jöreskog and Sörbom, 1989) to estimate the corrected correlations, using the model specified in Figure 2, or a larger model for all traits for which data have been collected. Appendix A provides the LISREL input for such an analysis.

Below we give some examples using equation 5 or Figure 2. First of all, the example of the last section can be reversed. For the instruments presented in Table 6, the validity, reliability and method effects were calculated. If in the Netherlands a correlation of .55 is obtained with these instruments, and in Hungary a correlation of .63, then equation 5 can be used to show that in both countries the correlation between the two variables, corrected for measure-ment error, is identical and equal to .8.

On the other hand, if, under these conditions in both countries, a correlation between GLS and SH of .63 is found, then, using equation 5 and the results of Table 3, it can be shown that the correlation between these variables, corrected for measurements error is .95 in the Netherlands, and .80 in Hungary.

This example shows that equal correlations obtained with identical instruments can be due to quite different correlations between the variables of interest. This means that by using this correction for measurement, one can control for differences in error structures between countries and make the results comparable.

5. Conclusion

In all textbooks about structural equation models, a multiple indicators approach is recommended for the estimation of, and correction for, measurement error. Although this approach is statistically correct, many practical and substantive problems are associated with it.

First of all, it is rather expensive to measure each theoretical variable in at least two different ways. It means that one doubles the interview time, which usually is quite costly.

Second, it is difficult to ask the same question twice in one interview. Although possible, it is not easy to organise, and one risks irritating respon-dents who notice the repetition. As a substitute, researchers often vary the formulation of the repeated question. However, Heise (1969) and Saris (1982) have argued that variation in question wording might change the meaning of the variable one measures. There are, moreover, many studies which demonstrate this point, even for the mean and variance of the variables (see Schuman and Presser (1981); Belson (1981)). Consequently, it is not clear what a multiple indicator model in such a situation represents. The latent variable will be a common factor of two or more indicators, but because these indicators are substantively different, it is unclear what this common factor stands for.

On the other hand, correction for measurement error seems to be a necessity, as we have tried to indicate. We have shown that the commonly accepted idea that results can only be compared across countries if the same method has been used is, in fact, incorrect. Even if the same method is used, one can get different results due to differences in the error structure in the different countries. Therefore, correction for measurement error is necessary. Corrected correlation coefficients are more comparable, not only across different studies but also across different countries. Also, the correction for measurement error provides a better estimate of the explained variance in each equation. This is important for the evaluation of the quality of different explanatory models.

We hope to have indicated in this chapter that the proposed procedure allows correction for measurement error even if only one indicator is used for each theoretical variable. When large methodological studies as described in Scherpenzeel and Saris (1997) are involved, and tables like Table 3 here are constructed for more topics than life satisfaction (see, for example, Andrews, 1984; Rodgers et al., 1992; Költringer, 1993; Scherpenzeel, 1995), the procedure described here can be used for any correlation matrix and any structural equation model. This is what makes it an attractive approach for national and cross-national studies.

The discussion in this paper has been limited to the effect of measurement error on the correlation between variables in cross-cultural research. There are, of course, more reasons for incomparability, such as coverage differences and fieldwork differences, mode effects, etc. The discussion has focused on problems with respect to the correlations; one can also study the effect on distributions of variables. A more general approach, covering a wider range of issues, can be found in Saris and Kaase (1997). Here we have concentrated on the misleading assumption that equality of measurement procedures is sufficient to guarantee comparability in cross-cultural research. We have shown that the situation is much more complex. Without correction for measurement error in each separate study, comparability is not guaranteed. We have also shown that many methodological studies are available to realise these corrections for measurement error.

References

Alwin, D. F. & Krosnick, J. A. (1991): The reliability of survey attitude measurement: the influence of question and respondent attributes. *Sociological Methods and Research 20*: 139-181.

Andrews, F. M. (1984): Construct validity and error components of survey measures: a structural modelling approach. *Public Opinion Quarterly 48*: 409-422.

Belson, W. (1981): *The design and understanding of survey questions*. London: Gower.

Billiet, J., Loosveldt, G. & Waterplas, L. (1986): *Het survey-interview onderzocht: effecten van het ontwerp en gebruik van vragenlijsten op de kwaliteit van de antwoorden*. Leuven: Sociologisch Onderzoeksinstituut KU Leuven.

Bollen, K. A. (1989): *Structural equations with latent variables.* New York: Wiley.

Dijkstra, W. & Zouwen, J., van der (1982): *Response Behaviour in the Survey-Interview*. London: Academic Press.

Groves, R. M. (1989): *Survey Errors and Survey Costs*. New York: Wiley and Sons.

Heise, D. R. (1969): Separating reliability and stability in test-retest correlation. *American Sociological Review 34:* 93-101.

Heise, D. R. & Bohrnstedt, G. W. (1970): Validity, invalidity, and reliabiity. In: Borgatta, E. F. & Bohrnstedt, G. W. (eds.). *Sociological Methodology*. San Francisco: Jossey-Bass.

Jöreskog, K. G. & Sörbom, D. (1988): *Prelis: a program for multivariate data screening and data summarization* (second edition). Mooresville: Scientific Software.

Jöreskog, K. G. & Sörbom, D. (1989): *Lisrel VII: Users reference guide*. Mooresville: Scientific Software.

Költringer, R. (1993): *Messqualität in der sozialwissenschaftlichen Umfrageforschung.* Wien: Endbericht Project P8690-SOZ des Fonds zur Förderung der wissenschaftlichen Forschung (FWF).

Lord, F. & Novick, M. R. (1968): *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Molenaar, N. J. (1986): *Formuleringseffecten in Survey-interviews*. Amsterdam: VU-uitgeverij.

Rodgers, W. L., Andrews, F. M. & Herzog, A. R. (1992): Quality of Survey Measures: a structural modeling approach. *Journal of Official Statistics 8:* 251-275.

Saris, W. E. (1982): Different questions, different variables. In: C. Fornell (eds*.). A second generation of multivariate analysis: Vol. 2. Measurement and Evaluation*. New York: Praeger.

Saris, W. E. (1990): The choice of a model for evaluation of measurement instruments. In: Saris, W.E. & Meurs, A., van (eds.). *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies*. Amsterdam: North Holland.

Saris, W. E. & Andrews, F. M. (1991): Evaluation of measurement instruments using a structural modelling approach. In: Biemer, P. P., Groves, R.M., Lyberg, L.E., Mathiowetz, N. and Sudman, S. (eds.). *Measurement Errors in Surveys*. New York: Wiley and Sons.

Saris, W. E. & Münnich A. (1995): *The Multitrait-Multimethod approach to evaluate measurement instruments.* Budapest: Eötvös University Press.

Saris, W. E., Veenhoven, R., Scherpenzeel, A. & Bunting, B. (eds.) (1996): *Life Satisfaction in West and Eastern Europe*. Budapest: Eötvös University Press.

Saris, W. E. & Kaase M. (eds.) (1997). *Eurobarometer: measurement instrument for opinions in Europe* ZUMA-Nachrichten Spezial No. 2. Mannheim: ZUMA.

Scherpenzeel, A. C. (1995): *A Question of Quality: Evaluating survey questions by MTMM studies.* Ph.D thesis. Amsterdam: University of Amsterdam.

Scherpenzeel, A. (1996): Life Satisfaction in the Netherlands. In: Saris, W. E., Veenhoven, R, Scherpenzeel, A. & Bunting, B. (eds.): *Life Satisfaction in West and Eastern Europe.* Budapest: Eötvös University Press. Chapter 3.

Scherpenzeel, A. & Saris, W. E. (1997): The validity and reliability of survey questions: A Meta analysis of MTMM studies, *Sociological Methods and Research 25*, 341-383.

Schuman, H. & Presser, S. (1981): *Questions and answers in attitude surveys: experiments on question form, wording and context*. New York: Academic Press.

Sudman, S. & Bradburn, N. L. (1974): *Response Effects in Surveys.* Chicago: Aldin.

Appendix A.

LISREL input to estimate corrected correlations between four satisfaction variables.

```
        Satisfaction Netherlands, 5p scales, correction on basis of meta-analysis
da ni=4 no=1599 ma=pm
la
*
'sat5p1'  'sat5p2'  'sat5p3'  'sat5p4'
pm file=sat5p.pm
model ny=4 nk=5 ne=4 te=fi ga=fi ps=ze ph=sy,fr
le
*
'truesco1' 'truesco2' 'truesco3' 'truesco4'
lk
*
'general' 'house' 'financial' 'contacts' '5p'
value .86 ly 1 1
value .93 ly 2 2
value .92 ly 3 3
value .89 ly 4 4
value .89 ga 1 1
value .90 ga 2 2 ga 3 3
value .89 ga 4 4
value .46 ga 1 5
value .44 ga 2 5 ga 3 5
value .45 ga 4 5
value .26 te 1 1
value .14 te 2 2
value .16 te 3 3
value .21 te 4 4
fi ph 4 5 ph 3 5 ph 2 5 ph 1 5
fi ph 1 1 ph 2 2 ph 3 3 ph 4 4 ph 5 5
value 1 ph 1 1 ph 2 2 ph 3 3 ph 4 4 ph 5 5
start .5 all
output ns ss
```

Questionnaires in Translation

JANET A. HARKNESS AND ALICIA SCHOUA-GLUSBERG

Translation of questionnaires is the most frequently chosen route to implementing 'equivalent' instruments in cross-national and cross-lingual survey research. The article presents the framework of current survey translation practice: the various procedures proposed for translation and for assessment of translation products and the respective advantages or disadvantages of each. In doing so, pointers are made to research gaps in questionnaire adaptation and evaluation for cross-cultural work and to the need for interdisciplinary transfer from cognitive survey research, translation studies and statistical analysis in order to establish a thorough-going methodology of questionnaire adaptation, assessment and documentation.

1. Why and when questionnaires are translated

The most common reason for translating questionnaires is to be able to field an instrument not available in the language required for fielding. Thus the best-known cross-national survey projects operating on a regular basis (including EUROBAROMETER, ISSP, WVS, LATINOBAROMETER)[1] translate from source questionnaires into the other languages required. Within countries with more than one official language, questionnaires for the different linguistic populations are usually produced from one questionnaire. In America, Spanish-speaking populations are frequently interviewed using Spanish questionnaires translated from the English source questionnaires.

[1] EUROBAROMETER: the official regular survey carried out for the European Commission; ISSP: the International Social Survey Programme, an annual survey of topics of interest to the social sciences, 30 member countries in 1998; WVS: World Values Survey; LATINOBAROMETER: the South American counterpart to the EUROBAROMETER.

The need to translate a questionnaire is sometimes apparent from the outset if one or more targeted populations is known to need a different language from the one in which the questionnaire is\will be designed. Alternatively, the need for a translation may only become apparent at a later stage. In some American studies, for example, the 'luck of the (sample) draw', i.e., where the sample falls, decides whether a translation is made. A multi-stage probability sample which selects rural counties at the first stage may, for example, end up including counties with a high density of Spanish speakers who require a Spanish questionnaire. Translations are also made in some surveys on an *ad hoc* basis during fielding; interviewers orally translate their questionnaire in order to field with respondents who require another language (see section 4.7).

Instruments are also translated when researchers wish to field items originally conducted in another language. Two further reasons are a) when questionnaires are translated so as to consider their items or coverage in developing new questionnaires and b) when translations (or glosses) of items are made, usually into English, for the electronic question banks and databases now appearing. The issues related to these last two contexts are not discussed here.

2. Materials Used to Produce Translations

2.1 Source Language Questionnaires (SLQs)

A common point of departure for translation is what we call a *source language questionnaire* (SLQ) in finalised form. In a *finalised* questionnaire, every component has basically been decided and fixed. In European multi-national and in international projects, the SLQ is often in English and is finalised before translation begins. One notable exception is the EUROBAROMETER, for which French and English source questionnaires are provided. Occasionally, translation begins when the SLQ is still at the drafting stage. The aim here may be to use *advance translating* (section 4.5) to refine the draft towards a final version.

In some studies, there may not *be* a questionnaire to translate. Instead, topics, dimensions, and perhaps numbers of items may be set out in one language; the questionnaire is then developed in another language on the basis of these. Although elements of 'translation' of concepts are involved in this situation (cf. Gutknecht and Rölle, 1996:297f.), it is best thought of as *foreign language implementation of design specifications*. In this situation, a questionnaire in the language of the specifications may never appear, or only appear at a later stage to allow designers to discuss the implementation.

2.1.1 Development of SLQs

SLQs for multi-lingual implementation are developed under different conditions, which in turn may impact on the products. They include the following:

- mono-culturally by people all sharing the same general cultural-linguistic background;
- by people from one country with different first languages or habitually using different languages (Switzerland, Canada);
- by people from one country with different standard varieties of one language (in the UK, the Welsh, Scots, N. Irish and English);
- by people from several countries speaking varieties of one language (Spanish-speaking South American countries, or GB, USA, and New Zealand);
- by people from different nations and cultures speaking different languages using one language as a lingua franca (e.g., the ISSP, multi-lingual, multi-cultural);
- (potentially related to the above) by a group developing an SLQ in a language which is the first language of very few or no-one in the group.

The processes and dynamics of multi-cultural development raise numerous issues, not least since questionnaire development involves detailed consideration of formulations. These cannot be gone into here.

2.1.2 Types of Finalised SLQs

Finalised SLQs take various forms:

- a questionnaire exactly in the format fielded in a country using the source language. It may indeed be some country's questionnaire;
- a questionnaire *text* in the source language which is not set out as a questionnaire and also not pre-tested in the source language;

  The wording for this text is thus fixed but format and lay-out, for example, may not be. In view of the communicative nature of all the elements in a questionnaire (Harkness 1995, 1996; Schwarz, 1996), this may not be an optimal source document from which to produce a translated questionnaire expected to be 'equivalent'.
- an SLQ which includes background variable items to be used everywhere or, as a further variation, an outline of required background variable information (cf. Harkness, Mohler and McCabe (1997) on background variables and cross-cultural comparability);
- an *annotated* questionnaire;

  Recent ISSP modules have annotated questionnaires.[2] The questionnaire now distributed in English for a module is the 'prototype' questionnaire for the Programme. It cannot be used *exactly as it stands* in any country. Countries fielding in English would, for example, remove certain notes in brackets and insert their country-specific elements in the same way as countries required to translate. The ISSP annotation includes:

  – notes on elements to be adapted in a country-specific manner, e.g., school-leaving ages and culture-specific institutions such as *Parliament;*
  – general (non-country-specific) glosses of elements thought or known to be problematic in translation (e.g., British English terms such as *civil servants*, *social security,* (cf. Role of Government, 1996);
  – notes on dimensions in items expected to need free translation (cf. Work Orientations, 1997);
  – reminders to implementers to observe certain goals of the design (cf. Religion, 1998);
  – indications where an ISSP member has permission either *not* to field or to adapt a question (cf. Religion, 1998 for Japan);

– indications which, if any, questions are optional;
– specifications of special background demographics beyond the compulsory ISSP set.

- a questionnaire which incorporates items which are to be translated and fielded in each language and/or culture (etic items) and items which are to be individually developed as cultural equivalents for all or some of the languages and/or cultures (emic items);

Despite the appeal of the increased and more detailed coverage of local information through emic items, concerns about how to establish and code 'equivalence' mean that these options have been little used in multi-national survey work (but see, for example, Przeworski and Teune, 1970; Flaherty et al., 1988; Triandis, 1994; and Johnson, this volume). Hui and Triandis (1985:143-144) outline the 'combined etic-emic approach' in which etic constructs are identified and then measured in emic ways. They, too, note that the lack of item equivalence and scalar equivalence make "direct comparison of cultures impossible".

- a 'new' SLQ which includes items already used in other studies.

Items already used in accredited studies will generally be preferred over newly developed items, partly because actual use is seen as the best possible 'pre-test' and because replicating them offers some opportunity for comparison of findings. The conditions under which such 'old' items were developed and the existence of translations which have already been used affect the questionnaire currently being translated. Although the new questionnaire may have been developed multi-culturally, 'borrowed' items may not have been. Tension then arises between tinkering with item wording to 'improve' them and using 'tried and tested' items. Existing translations, even if considered sub-optimal, may be adopted for similar reasons.

2.1.3 Draft Source Questionnaires

Different types and stages of questionnaire text are variously referred to as *draft questionnaires*. A source language draft questionnaire used in a decentering approach

[2] Source questionnaires are available from the ISSP secretariat and ISSP archive (addresses at the Programme's web site: www.issp.org)

(Werner and Campbell, 1970) is the text from which both source and target language versions are produced more or less simultaneously (section 4.1). In contrast, draft translations in committee translation approaches are the texts from which the final translated version emerges (section 4.3). ISSP drafting involves a series of draft questionnaires in English, produced by a multi-cultural drafting group in working towards a final source language questionnaire. Each draft incorporates feedback in English from (potentially) all member countries of the Programme.

3. Facts of Translation

Translation of instruments is not the only means available to gather information on dimensions and constructs across cultures, but it is generally seen as the only means to ensure item equivalence and scalar equivalence (cf. Hui and Triandis, 1985; Flaherty et al., 1988; Van de Vijver, this volume). Acquadro et al. (1996:575) identify two major arguments for using the same (translated) questionnaire in different countries: a) "a common international interpretation and analysis of the results is only possible if the data come from the same instrument" and b) all new data acquired about an instrument contribute to the validation and reputation of the instrument (especially relevant in the context of much-used instruments). Translation is at all events the most frequently adopted approach and certainly the approach the majority of researchers see as the most viable option (cf. Guillemin et al., 1993; Van de Vijver, this volume).

Translation beyond the field of instrument translation takes many forms, with different outputs for different purposes. The goals of a particular translation – whether this be to convey the factual information, the sound effects, or the communicative intention of a source text – determine the product of the translation process (cf. Kiraly, 1995:55). In survey research, questionnaire translations are generally, if vaguely, required to ask the same questions and offer the same response options as a source text. Rightly or wrongly, they are expected to do so by means of a close rendering of the source questionnaire (section 4.4).

3.1 Equivalence and Adequacy/Appropriateness

Languages are not isomorphic and so translation cannot be expected to operate on a one-to-one basis across languages. This means that what goes in (the source language text) cannot be completely matched by what comes out (the target language text). Indeed, a mechanistic notion of input and output is itself misleading. Moreover, translation is not solely concerned with translating 'meaning' (on 'meaning', see section 3.2). As mentioned, some translations are aimed at conveying sound effects or emotional effects, while others focus on conveying factual information or (distinct from this) communicative intention.

Translation necessarily involves difference as well as similarity. Absolute absence of difference would amount to replication of the source text in the source language, absolute absence of similarity would force us to query the status of one text as a translation of the other. In terms of 'equivalence' between texts, difference is sometimes discussed in terms of 'loss' and 'gain' (Newmark, 1988:7f.). S*emantic* loss and gain occurs as soon as nuances of meaning associated with a lexeme sense (see Lyons (1977:197f.) on sense) in the source language are not covered by the lexeme(s) in the target language, while other nuances, anchored in the target language and culture, are 'gained'. An example of *grammatical* loss or gain could be that the sex of individual people referred to is indicated in one language (as frequently is the case in German) but not in another language. Harkness (1996a) discusses issues this raises for translating survey items, while Acquadro et al. (1996:582) recommend gendered versions of questionnaires.

Social science research has its own rich array of kinds of 'equivalence' (Johnson, this volume). Different but equally varied kinds of equivalence are referred to in translatology writings (e.g., Snell-Hornby, 1988:13-22; Kiraly, 1995:54f.). These include expressions which formally match some used in survey research but have different senses (e.g., 'functional equivalence'). Gutt (1991:10-17) argues against applying the term equivalence to translations at all and demonstrates that equivalence (however understood) cannot be automatically equated with quality (see, too, Reiss and Vermeer, 1984; Hönig and Kussmaul, 1984). We avoid using the term 'translatory equivalence' here. In

considering translation quality, we prefer to think of assessment in terms of appropriateness or adequacy for a given task. The appropriateness or adequacy of a given translation is then defined in terms of the degree to which it successfully fulfils stipulated goals for the translation, within the constraints of what is possible. Admittedly, a major problem here for survey research is that concrete translation goals are rarely articulated (sections 3.4 and 3.5).

3.2 Meaning is defined in and by use

The second constraint on our expectations about translation and equivalence has to do with survey standpoints on the meaning of well-written items. Handbooks outline how to avoid writing poorly formulated and ambiguous items. The implication is that the meaning of well-written items will be clear or unambiguous. This, in turn, implies that there is such a thing as 'the' meaning of an item.

However, in many fields of research this is by no means the currently accepted view of meaning as related to words and larger units of language in use. Over the past three decades, in research fields concerned with meaning – such as linguistics, literary theory, social psychology, language philosophy and translation studies – the meaning of words and the larger units they constitute has come to be seen as determined in and by use. Meaning in a given context is thus seen as determined in and by that context in its widest sense and as co-constructed between users. By this is meant that the co-text (the surrounding or accompanying text), the immediate and larger contexts, the text producer(s) and the recipient(s), as well as the lexical content of expressions and the propositional content, all affect what may be perceived as 'the' meaning of a communication in a given instance. Moreover, a reading ('meaning') perceived by an addressee need not be the meaning intended (and perceived) by the speaker/writer. Seeing meaning as *dialogic* shifts the goal of communicators away from making 'the' meaning clear and towards making the *intended* meaning clear, or as clear as possible. This applies equally to communication in questionnaires (Harkness, 1996b; Schwarz, 1996). The success of everyday communication shows that we are adept (ultimately) at getting

intended messages across. But the multiple repairs, repetitions, explanations and expansions we engage in simultaneously underline that meaning is, indeed, dialogic.

This is relevant for survey translation in two main respects. First, given that meaning is not fixed and finite, one of the goals of translation must be to convey the intended and most salient reading of a well-written question. The intended meaning of an item should therefore be documented for translators in the source materials they receive for their task (Hambleton, 1993). Whether this reading can be conveyed by means of close translation (of the moderate kind) is a separate issue (see 4.4), as is whether the salient reading in translation continues to tap the dimension or construct required. The factors which determine a given reading may differ across cultures, thus a close translation in terms of lexical content can conflict with the goal of conveying intended meaning. This poses a real problem for survey translation, since in many instances we currently do not have documentation on intended readings, nor on intended dimensions. We also lack detailed guidelines and examples for what might constitute an acceptable degree of freedom in producing the target text. In this situation, researchers understandably hesitate to experiment.

Significantly, the number of researchers and research bodies suggesting and calling for improved and systematic documentation and guidelines is growing (Hambleton, 1993 and 1994, Prieto, 1992; Guillemin et al., 1993; Acquadro et al., 1996; Van de Vijver and Hambleton, 1996). Only through such documentation and provision of information will modifications to current practice be able to be realised in a consistent fashion. Basic empirical research is needed, too, on how issues of adaptation can best be tackled beyond the modest beginnings of Harkness (1996c), Harkness et al. (1997), Mohler, Smith and Harkness (this volume) and Harkness, Mohler and McCabe (1997).

3.3 The Dual Nature of Questionnaires: Instruments and Texts

A third factor shaping the demands on questionnaire translation is the dual character of questionnaires as *texts "destined for discourse"* (Harkness, 1994, 1995) and as

instruments of measurement. Whether self-administered or read out by interviewers, in principle at least, the questionnaire text determines what is said to or read by respondents. In the closed question format, the questionnaire also basically determines the responses open to respondents. Even if rarely talked about in quite these terms, one of the goals of questionnaire design in the monolingual context is thus to optimise communication of intended stimulus and response. However, optimising communication in the target language may, again, run counter to close translation expectations. For the present, we lack research on how questionnaires as holistic entities (Harkness 1995, 1996b) can best be adapted for other language implementation.

3.4 Translators need Information and Task Specifications

Given the fluidity of meaning and the range of interpretation many texts allow, translators decide what they want (or have) to communicate and then try to do that (Wilss, 1996; Kussmaul, 1986), within the confines of what is possible across the given languages and cultures. These decisions are never made in a vacuum. If not provided with task specifications, translators are forced consciously or unconsciously to provide their own, as is evident from think aloud protocols of survey translators at work (Harkness, 1996c). Translation manuals thus increasingly stress the importance of adequate information and *task specifications* for translators (cf. Wilss, 1996; Kussmaul, 1995; Holz-Mänttäri, 1984; Gutknecht and Rölle, 1996; Gile, 1995). However, task specifications for questionnaires based on requirements, guidelines and standards agreed by the survey research community are not yet available. The recommendations, overviews, and guidelines which have appeared in other areas of research using instruments (Hambleton, 1993, 1994; Prieto, 1992; Acquadro et al., 1996) and, of course, within the field of translation studies (see above) are invaluable starting points for the social sciences.

3.5 Providing Information and Task Specifications

Given the complex dual nature of questionnaires – seemingly simple texts with overt and covert measurement properties – task specifications need to be negotiated between those best informed about textual properties and those best informed about measurement

properties. These task specifications are likely to consist of a compromise between what researchers wish to have and what translation (not translators) can deliver. There can be little doubt that specifying translation tasks will require an exchange of information between researchers, questionnaire designers, target language implementers and translators. 'Rules' of practice in certain fields (Acquadro et al., 1996) suggest that that personal contact between item writers, research teams and translators is assumed to be possible. In our experience, however, implementing situations constantly arise in which a) no individual exchange will be possible, b) people involved in the design are in any case no longer sure what items 'mean' in detail, c) item writers (if, indeed, items are products of individual composition) quickly become anonymous (Harkness, 1994) and d) items move in undocumented journeys from survey to survey, country to country, and formulation to formulation.

Be this as it may, documentation could be organised without undue difficulty to provide the information needed to negotiate a translation. This would need to include information on what is required in terms of measurement, what is intended in terms of textual communication, what is possible in terms of translation versus other forms of adaptation, and where particular language and/or culture problems may arise. Certainly, this kind of documentation is essential to further interdisciplinary understanding of the demands on questionnaires in translation.

4. Some Survey Translation Procedures

In this section we briefly describe the translation approaches most frequently referred to in the survey context. Back translation is discussed under section 5 on *assessment* and not here under *translation,* although it is sometimes referred to as a 'translation method' (Sechrest et al., 1972; Brislin, 1970; McKay et al., 1996) and we briefly outline why.

Back translation involves the translation of a text which itself is a translation *back* into the original language (5.3.1). It is most commonly used and recommended as a way to assess

translation work. (e.g., Werner and Campbell, 1970; Brislin, 1970, 1976, 1980, 1986) but other uses are also suggested. Werner and Campbell (1970) describe a form of decentering which includes back translation steps and assessments – the multistage iterative process. They also suggest back translation can be used for translator assessment. Theoretically, there are as many ways to approach a translation *out* of what was originally a target language *back into* the source language as there are to produce a target language translation in the first place. Descriptions of back translation describe what the (back) translation product can be used for rather than the translatory goals and method involved in producing the back translation text itself. It is not an approach for arriving at a translation in the way that committee (parallel) translation or decentering can be seen to be (see below). We find it helpful, therefore, to maintain a distinction between kinds of translation approaches and uses to which a translation can be put.

4.1 Decentering

Decentering in translation (Werner and Campbell, 1970) is a technique which begins from a *draft* questionnaire in the source language in order to produce final questionnaires in *two* languages (source and target) through a process of paraphrase and translation between source language and target language. Paraphrase is seen as a way of decentering the text in both languages, that is, producing texts which are not 'centred on' or 'anchored to' a specific culture and language.[3] Schoua (1985) reports positively on a Spanish-English decentering experiment, as do McKay et al. (1996); recent psychological test translation work has also shown interest in decentering (Tanzer et al., 1997).

Werner and Campbell (1970) suggest several approaches to decentering including taxonomic decentering, multiple stage translation, mapping of paraphrases across languages, and interview schedule-based decentering. In essence, decentering involves the following (with variations depending on the procedures chosen):

[3] The idea that this is possible (in natural-sounding utterances at least) runs counter to theories in which meaning is determined by use and use is invariably tied to the culture in which it occurs.

- each draft question is reformulated and paraphrased with the goal of eliminating culture-specific aspects and simplifying complex sentences into basic, most simple constructions;
- each item (or set of paraphrases for an item) is translated into the target language. Here the idea is not to translate in any 'close' or literal fashion, but to produce as many paraphrases in the target language of the 'meaning' of the source language text(s) as possible;
- these paraphrases in the target language are translated in comparable 'paraphrase fashion' into the source language;
- the sets of paraphrases for each item/sentence in each language are compared;
- the closest equivalents across the two languages are selected;
- this selection forms the basis of both final questionnaire texts for the item/sentence.

One generally important feature of decentering approaches is that they stand in direct contrast to the 'close' translation described in section 4.4.1, which clings to words or structures across languages and, in doing so, produces unnatural-sounding translation. However, through decentering, the items may also end up sounding odd, an aspect Werner and Campbell (1970:411) consider unimportant. Another important feature of decentering is the centrality it gives to working out different versions in different languages before a 'source' text is fixed for posterity (cf. 4.5).

At the same time, in a world of survey fielding of old 'tried and tested' items, the source text is often not open to emendation. Translation may also be required into many languages. Werner and Campbell focus on two language instrument development and it is difficult to see how a many-to-many matching across, say, twenty languages might be practicable (cf. Werner and Campbell, 1970:406). In addition, the procedures are demanding in terms of time, personnel, qualifications and funding, all real stumbling blocks in the world of survey management and funding. The inherently subjective basis of judgements taken at each of the comparative steps – from identifying paraphrases to be rejected to selecting 'the best' or closest equivalents – is a key factor and, for some, a key

weakness in decentering. Lastly, decentering takes a sentence-based view of meaning as its starting point, with words and grammar interacting to provide sentences with their meaning (Werner and Campbell, 1970:401). More investigation will be needed to assess how successful the procedure can be in different contexts (perhaps in a leaner version) and how it can cater for different notions of textual equivalence, the dialogic view of meaning, and idiomatic-sounding items.

4.2 'Direct' or 'one-for-one' translation

In terms of procedure, the 'simplest' and cheapest translation approach has one translator producing one translation in a traditional manner – the translator simply produces a translation to the best of her/his ability. Sechrest, Fay and Zaidi (1972) call this 'direct translation', a term not to be thought of as in contrast to 'indirect' or 'less straightforward'. References to this kind of approach specify neither the translation process nor the product type envisaged. Limiting the work to one person is attractive in terms of funding, organisation and streamlining of time schedules. The absence of support materials for translators, the low impact so far of translatology findings on theory, practice, and on assessment procedures, the disadvantages of relying on one person's perceptions and skills, the lack of coverage of regional differences (where these are an issue), and, finally, the data quality risks this involves are drawbacks to this approach, at least as frequently implemented (Sechrest, Fay and Zaidi, 1972; Guillemin et al., 1993; Acquadro et al., 1996).

McKay et al. (1996) use the term 'direct translation' for translation from source to target language, that is, 'one way' (forward) translation as opposed to 'two way' (forward and backward or 'double') translation, ie., translation and back translation.

Acquadro et al. (1996:577-578) define direct translation as translation which "comprises borrowings, calques (loan translations) and word-for-word translation". (What is meant by word-for-word translation, which is contrasted with 'literal translation' is uncertain.) They contrast direct translation with indirect translation. This last is characterised as

involving "transposition, modulation, equivalence and adaptation". The processes outlined for indirect translation suggest, broadly speaking, that it pursues (stipulated) goals as a covert translation, whereas direct translations are (among other things) overt translations. Covert translations read 'naturally', overt translations signal that they are translations (see 4.4.1).

4.3 Committee and Modified Committee Translation

Committee approaches are used for translation (discussed here) and for translation assessment (discussed in section 5.3.2). Committee or parallel translation involves several translators who make independent translations of the same questionnaire (Brislin, 1980; Schoua-Glusberg, 1992; Acquadro et al., 1996 (team translation); Guillemin et al., 1993). At a reconciliation (consensus, revision) meeting, translators and a translation co-ordinator compare the translations, reconcile discrepancies and agree on a final version which taps the best of the independent translations or, alternatively, appears in the course of discussion. The committee members should provide competence in whatever varieties of the target language are required for respondents (McKay et al., 1996; Acquadro et al., 1996) and in the various skills required for survey work (Van de Vijver and Hambleton, 1996; Johnson et al., 1997).

The committee approach is fairly labour, time and cost intensive. Schoua-Glusberg (1992) proposes a modified committee approach which involves group work but not parallel translation. Each translator works on a different part of the text rather than the whole text. The committee reviews the text provided in sections by different people and arrives at a final version. The approach can maintain the quality of parallel translation work while cutting some costs and reducing the time needed to arrive at a final version, in particular if the questionnaire is long (Schoua-Glusberg, 1992). Care must be taken to ensure that consistency is maintained across the translation. This applies, of course, for any translation, whether produced in parallel fashion, using the modified approach, or produced by one translator. However, the modified committee approach may require

particular care. For example, it is conceivable that two translators both translate the same expression in the individual parts of the questionnaire each has, and that each comes up with a suitable, but different, translation. Neither translation gives cause for discussion in the committee session. Without consistency checks as a standard part of the process, the term used to refer to 'X' in one part of the questionnaire could unintentionally end up being different from the term used for 'X' elsewhere.

Like any approach which assesses equivalence or appropriateness on the basis of textual evaluation, committee decisions are ultimately based on subjective judgements. Committees are as open to group dynamic drawbacks as other groups. Given individual competence within the group, however, group screening is likely to be effective. While competent translators are necessary (section 6), the role and skills of the committee co-ordinator are crucial, as is an understanding and acceptance of the procedures by all involved.

Institutes or researchers faced with sporadic cross-lingual implementations may find it complicated to maintain a translator committee group who stay 'in practice'. The German ISSP questionnaire is, essentially, the only translated questionnaire the institute involved produces per year. In this context, no group of skilled translators working frequently together on survey translation is available. Instead, the co-ordinator recruits institute researchers with the necessary understanding of survey instruments and grasp of English, student research assistants with competence in English, with and without survey knowledge, and two translators (skilled practitioners) – one external and working regularly as a translator, one internal, working in survey translation research. Germanic language members of the ISSP who need to translate (Austria, Germany, Norway, Sweden) also confer on problems and solutions. In this way, a mix of input from much the same group of people can be maintained from year to year without unreasonable costs or effort. This compromise solution has proved useful, in that it brings together translation drafts guided by instrument knowledge, translations from skilled practitioners and 'fresh' insights from 'outsiders' (students) and includes a degree of consultation across countries and languages.

4.4 Close and Literal Translation

Survey research often favours close renderings of questions as a means to arrive at equivalent measurement. In view of a) the often vague nature of discussions of what 'close' translation is and b) this vagueness notwithstanding, the differing descriptions of 'close' and 'literal' found in instrument literature (and, differently again, in translation literature), we indicate briefly the minimum kinds of 'closeness' we understand to be involved. A close rendering in survey terms would, for example, be expected to refer to the same entities (have the same *referent*) as referred to in the source text (sport, education, TV-watching, God). The entities would also be referred to using lexemes which cover as much of the same *sense*(s) as possible and come as close as possible lexically to the source text choice. The morning star and the evening star – if unlikely candidates for an item – may help us make distinctions here (cf. Lyons, 1977: 197f.). Thus, if the source text mentions the *morning star* (referent Venus), the target text would, if possible, refer to that too, and not, for example, to the *evening star* (referent Venus), nor to Venus with a lexeme like *Venus*. Exceptions to close renderings are what in survey research are sometimes called 'country-specific renderings' for country-specific institutions such *Parliament*, *A-levels* or *Prime Minister* (and, presumably, across cultures and religions, *God*). The idea that this might also apply to *sport, education,* and *TV-watching* is not familiar to survey researchers. If these were felt to require 'country-specific renderings', new items would probably be looked for which avoided these problems. Furthermore, the propositional content of the source text would be expected to be maintained in the target text. In other words, *God created human beings* (X predicate $_{(create)}$ Z) in English would not, for instance, become something more like (Z predicate $_{(create)}$ X) in another language.

4.4.1 Too Close for Comfort?

For survey translations, greatest emphasis is usually placed on avoiding differences in semantic information (lexeme senses) and grammatical information (e.g., number,

tensing, mood). Sticking close to source language (and culture) *ideas* and *concepts* in items is on occasion tricky enough, even with well-designed items. Sticking close to ideas by means of sticking close to lexical senses will at times amount to a lost cause. But even where a close rendering is possible, a 'successful' translation in terms of a close rendering carries with it no guarantee that communicative functions (and with them measurement properties) are equally well retained (Hulin, 1987; Hambleton, 1993; Flaherty, 1988; Johnson, this volume; Van de Vijver and Leung, 1997; Van de Vijver, this volume). In some contexts the need for culture-specific equivalents is apparent, in others rather less so (Harkness and Braun, in preparation).

Survey translations frequently go beyond only trying to convey ideas and concepts from the source text. Following what seems to be a survey understanding of 'close' translation, formulations, words and syntax are copied or imitated across languages (cf. what Acquadro et al. (1996) call 'direct' translation, something like a word-for-word gloss and the 'literal' translation described in McKay et al., (1996). This partly stems from the survey concern to ask the same items in order to compare data. It also reflects survey perceptions of the options available through translation. It may also be related to using back translation as an assessment (Hulin, 1987). An extreme form of close or literal translation is unlikely to result in a covert translation, that is, one which does not signal its foreign origins. It also stands in conflict with the fact that translation involves and requires change, adaptation and compromise.

Covert translation versus overt translation (House, 1977) raises questions related to how respondents perceive the questions and questionnaire. *Overt translation* is the production of a target language text which signals (in a variety of possible ways) that it is a translation. *Covert translation*, in contrast, produces a target language text which reads like an original text of the given text type in the target language and thus does not signal that it is a translation. A considerable body of cognitive research in the monocultural context documents that respondents react to features of questionnaire design which researchers have neglected, and that they do so in predictable ways (reviewed in Schwarz, 1996). By extension, we could expect that questionnaires which signal they *are*

translations (or simply come across as odd texts in some way) will prompt certain responses in respondents. Unless there is a valid reason why respondents should consider the origins of the questionnaire, we suggest that survey translations should be covert translations, ie., should read/sound like original target language questionnaires (cf. Sechrest et al., 1972; Hulin, 1987; Harkness, 1996a). But even if a close rendering of an item results in a) a natural-sounding translation which to all intents and purposes b) fulfils measurement requirements and c) is viewed as a close translation of the source item, *difference*, that is, non-equivalence, is unavoidable.

4.5 Advance Translation

Drafting procedures recommended to the ISSP (Harkness, 1995b) propose that modules are translated while still in the drafting process, before the source questionnaire is finalised. Experience has shown that many translation problems linked to source text formulations only become apparent, even to experienced cross-cultural researchers, if a translation is attempted. As necessary, source formulations can be adapted or annotated on the basis of advance translation feedback and notes for the (annotated) source document can be greatly enriched. This is often particularly relevant for the languages and cultures furthest removed from the models underlying the source text; these are otherwise unlikely to receive much consideration in notes. Nevertheless, without empirical demonstration of the need to translate in advance, the additional effort and costs involved mean it is unlikely to be adopted as a standard practice.

4.6 Passing on the Translation to Fielding Institutes

This is less an approach to translation than a way of dealing with translation as an issue. Research groups sometimes commission the fielding organisation to produce the translated questionnaires required and may or may not be involved in any of the ensuing steps of production and assessment. Fielding institutes may well have more experience in producing different language versions of questionnaires than researchers. Ultimately, however, someone decides on task specifications, guidelines and assessment procedures.

In view of the generally scant exchange of information and research findings on translation procedures and assessments, the requirements, procedures and procedure control measures should be carefully negotiated with the institutes.

4.7 Translation of Finalised Questionnaires 'on the fly'

Translation is sometimes left up to the interviewer or an intermediary. By translating the available questionnaire orally, they are thus able to field with respondents requiring a different language, not an infrequent problem in multi-lingual societies. In the American context, if only a small number of respondents are expected to need a specific language version (not enough to 'warrant' producing a written translation), it is not uncommon for translations to be done 'on the fly', as it is called. Beyond knowing that these translations are made orally, little can be said about the *approach* taken in a specific case (e.g., free or close translation, emphasis on communicative functions, covert or overt translation, etc.).

Some modes of administration make it less likely, in the Western context at least, that translations will be done on the fly. If properly administered, a self-completion format should preclude this. Telephone interviews are more open to translation on the fly, whether as part of the design or not. Under pressure to display good interview achievement rates, interviewers may opt to translate rather than forgo an interview. Importantly, they may also have management permission to do so. Some US research companies use (readily available) bilingual telephone company operators to 'assist' interviewers with respondents unable or unwilling to answer in the language of the questionnaire. *Ad hoc* translation is, of course, used in other countries and continents, too. The general appeal is clear if we consider the obstacles for interviewers fielding, say, in parts of Africa or Asia, loaded down with eight and more language versions of questionnaires, but with never exactly the right version to hand.

The absence of written translations is of import for the data obtained. The relevance of standardly requiring not only a written translation of *question content* but also a *finished questionnaire* in translation is directly related to standard practices and requirements of

monolingual studies. In oral interviewing (of whatever mode) interviewers are trained to avoid providing non-standardised input in the dialogue. Despite problems this raises (Houtkoop-Steenstra, 1995; Cate Schaeffer, 1995; Stanley 1995a, 1995b), recorded questionnaires (paper or computer applications), with integrated, formulated instructions for interviewers, enable interviewers to comply with standard practices. In addition, the hard or soft copy questionnaires thus available are reasonable indications for later reference of what respondents were actually asked and offered as answer options, at least linguistically speaking. None of this follows from translation on the fly. In the worst case, researchers relinquish control of fielding and end up with response data but no record of what was asked and answered in general and in particular.

5. Assessment

The two central issues in translation assessment are what is to be assessed and how this is to be assessed. If the goals to be met by the product are not specified in advance, the criteria of assessment also cannot be specified in any manner fair to translators. In questionnaire translation, they are rarely specified, i.e., articulated, at all. Translation task specification is both a prerequisite for objective assessment of translations and for replication and validation of any assessment made. Without specifications, the usefulness of assessment procedures cannot be evaluated either. Given proper task specifications for translation, forms of assessment can be tailored to fit, within the confines of what can reasonably be expected.

Assessment of translated questionnaires is sometimes tied to the translation procedure adopted and/or to the questionnaire design (e.g., whether old items and old translations are replicated or not). Decentering has translation assessment as integral to producing the final questionnaire in *two languages.* Committee translation has assessment (reconciliation) as a central process in producing the final version of one *translated* questionnaire. Assessment may also be independent of both. Assessment procedures used once translation has been carried out are considered in section 5.3. Assessments of instrument equivalencies are discussed elsewhere (e.g., Hulin, 1987; Hulin, Drasgow and

Komocar 1982; Van de Vijver and Leung, 1997; Van de Vijver, this volume; Saris, this volume).

5.1 Bilingual and Monolingual Feedback

Assessment will either be made by bilinguals, monolinguals or both. Findings from bilinguals are not automatic pointers for findings from monolinguals. The two groups perceive texts, language and cultures differently (cf. Hulin, 1987). This said, bilingual appraisal of translations is an inevitable component of translation productions (with each translator appraising as she/he formulates) and a frequent and useful component of translation product assessment. It is important to avoid pressure on assessors and translators to defend one or the other translation version (issues of criticism of colleagues, superiors, etc.). Independent bilingual assessment of a text may simply mean that people not involved in the translation assess whether they consider the translated text to be 'equivalent' to the source text. Without stipulating what equivalence is understood to involve, this is clearly a hazardous undertaking. Given the subjective nature of textual assessment, even when guidelines are provided, it is important both to ensure a spread of qualified views and to include monolingual feedback.

Monolingual judgements of a translated text should in our view only be made on texts in the language the assessor speaks. In other words, we see little to be gained from having monolinguals compare a source text and a back translation to decide on a text they cannot read procedures. Having monolinguals go through a questionnaire – either as part of a pre-test, a probe interview or simply as copy-editing readers (e.g., interviewers reviewing it for readability) – are very useful. Given that they only know the target text, this group will only be able to comment on things they are asked about or which happen to strike them. This is perhaps less systematic than the comparison which can be made by (suitable) bilinguals accustomed to assessing texts intensively. There are also limits to what can be expected of people who match the target population in terms of textual assessment. Moreover, many questions remain open as to the representativity of the information received.

In general terms, the number of people needed to gain reliable information is a problem. For procedures described in section 5.3.4, large numbers of respondents are required before the data can be considered representative. In the target setting there will presumably be enough monolinguals available. For bilinguals, wherever the testing is being done, this may not be the case. Thus finding enough candidates of suitable language proficiency to participate is one issue. Apart from the double cultural perspective of bilinguals mentioned above, bilinguals may well not match demographic characteristics of the monolingual target groups. In addition, testing along the lines of split ballots, probe interviews, simulated interviews, etc., (section 5.3.4) are all expensive, time-consuming procedures which only produce data which must then be evaluated.

5.2 Assessment Basics

Very little of 'cookbook' nature can be passed on here either about tackling survey translation or assessing survey translation quality. The social sciences have been slower in articulating needs and guidelines than has been the case in psychology and other clinical fields such as medicine or specialised research fields (Hambleton, 1993, 1994; Guillemin et al., 1993; Acquadro et al., 1996; Prieto, 1992; Van de Vijver and Hambleton, 1996). Even in these fields, however, guidelines are still fairly general on translation techniques and assessment. They also seem to imply a greater homogeneity of items and of origins of items and a greater intensity of use and re-use than is common in surveys. Moreover, little research is available on comparative assessment of translation assessment procedures themselves. The following recommendations are therefore necessarily of basic nature:

- In assessing (and in finalising) the translation avoid loss-of-face confrontations. Set up different dynamics from the start to allow open assessment and criticism;
- Assess translated questionnaires (TQs) as covert translations, that is, as texts which read/sound like questionnaires designed in the target language.

- Base TQ assessment on bilingual assessment of SLQs and TQs, defining beforehand which equivalencies are essential.
- Base TQ assessment on monolingual (target population) assessment. (This may be hampered by the same problems on lay person feedback as experienced in monocultural research).
- Keep assessment requirements realistic. A covert translation required to maintain communicative equivalence and measurement equivalence may need to be a rather free translation.
- Choose assessors who understand the mediums involved - questionnaires as instruments and as texts in translation. As need be, find the competencies in several people.
- Even if the assessment is made by one person, extend revision decision-making to a group (which should include translators).
- Budget for assessment and revision (time, people, money).
- Keep in mind that translation assessment is not an assessment of measurement reliability and validity and take steps to assess these.

Research is needed on evaluating assessment procedures. Findings from the last decade of cognitive psychology research on survey design and from translation studies are likely to be valuable here, as is recent research on translation issues in the social sciences, medicine and psychology (Wilss, 1995; Kussmaul, 1995; Dollerup and Lindegaard, 1994; Acquadro et al., 1996; Prieto, 1992; Guillemin et al., 1993; Van de Vijver and Hambleton, 1996). Work is also needed on assessing approaches to survey translation (Sinaiko and Brislin, 1973; Schoua, 1985; McKay et al., 1996; Harkness and Braun, in preparation; Harkness, 1996c) and on investigating survey translation quality (Brislin, 1970; Schoua-Glusberg, 1988; Harkness and Braun, in preparation). Finally, since translation can only deal with some aspects of instrument adaptation, translation procedures and translation assessment have to be coupled with statistical investigations of instrument measurement properties and comparability across versions. Here, too, we need to clarify how best to implement all these in sequence or iteratively.

5.3 Assessment Procedures

General types of assessment which have some currency in discussions of survey translation include *back translation*, *comprehension assessment*, and various kinds of assessments based on *analysis* of response data.

5.3.1 Back Translation[4]

The term back translation is used in survey research literature and in translation studies to refer to the *translation of a translation* back into the source language. Almost without exception in survey work, the purpose of back translation is to compare/contrast the back translation with the source text, usually with a view to assessing the quality of a translation.[5] For survey translation, back translation is seen as offering a solution to the fact that researchers often need information about the quality of translations without being able to read and evaluate these themselves. It operates on the premise that if the translation is good, 'what went in ought to come out', the central idea being that a translation *back* into a language which can be understood allows researchers insight into a text in a language which cannot be understood. The basic steps involved are as follows:

- A source text in one language (Source Language Text One, SLT1) is translated into another language (Target Language Text, TLT).
- The TLT is translated back into the language of SLT1 by a second translator, unfamiliar with the SLT1 and uninformed that there is an SLT1. This second translation, the back translation, is SLT2.

[4] This section draws on material from Harkness (1996c).

[5] There are several references to back translation in translatology. Here authors generally include bilingual perspectives in discussion of the texts. Baker (1992) uses the term to refer to the (natural-sounding) English glosses she gives for texts in 'exotic' languages which themselves have been translated out of English. The purpose is to demonstrate difficulties met with in translations and ways of dealing with these. Vinay and Darbelnet (1958) discuss back translation as a means of assessing accuracy using bilingual insight. They suggest that back translation should really be into some third 'neutral' language, something unlikely to appeal to monolingual survey researchers looking for insight. Uses of back translation suggested in Werner and Campbell (1970) are discussed in section 4. Brislin (e.g., 1970, 1976, 1980, 1986) is most often cited in connection with back translation.

- SLT1 is compared to SLT2.
- On the basis of differences or similarities between SLT1 and SLT2, conclusions are drawn about the equivalence of TLT to SLT1.

The more identical SLT1 and SLT2 are, the greater the equivalence between the TLT and the SLT1 is considered to be. For example, if the source questionnaire in English has *Please enter your nationality* and the back translation in English has *Please enter your nationality*, then the TLT is assumed to say the same, only (somehow) in a foreign language. The frequent references in demographics literature to distinctions and overlaps between *citizenship*, *nationality*, *ethnic membership* and *religion* (e.g., Maier, 1991; Harkness, Mohler and McCabe, 1997) make clear, however, just how fluid the survey overlap for these concepts is. A point in case is the 1997 discussion in Russia about the absence of the fifth rubric in new passports (the rubric for 'nationality') and the positive reaction Russian Jews reportedly had to this. Among the admittedly scant references to translation assessment in study reports in recent years, back translation appears frequently (overview in Guillemin et al., 1993; Acquadro et al., 1996).

In general, back translation can be likened to a primitive metal detector; it can be expected to miss much, but also to pick up some things. It cannot identify what it picks up and neither, unfortunately, can the monolingual researcher. There is no necessary connection between what is 'picked up' (by virtue of being different from the SLT1) and what needs to be picked up. Pragmatically, it is likely but not necessary that major differences between a source text and its translation will also be reflected in a back translation. In saying this, note that the interpretation of what is 'major' is left as much to our readers as it is in the survey context to those deciding whether to change a translation. Brislin (1976) states, moreover, that one of the main disadvantages of back translation is that a good back translator will resolve problems actually present in the TLT (cf. Kussmaul's (1995) recommendations to translators to improve the text). Be that as it may, deciding on the presence or absence of 'difference' raises issues of meaning, appropriateness and equivalence and of how decisions are made on what constitutes a 'major' difference or a 'salient' difference.

A number of general points are noted here by way of clarification:

- Back translation itself does not deal with what, if anything, should be changed in a translation nor, crucially, how to change anything. Monolingual researchers thus come no further than looking at two texts in one language. In order to revise the target language questionnaire, bilingual competence has to be re-introduced, with all the imponderables this involves.
- If the back translation is simply used to make a list of things for *bilinguals* to look at in the target language questionnaire, other procedures which compare the SLQ and the TLT, such as committee discussions (sections 5.3.2 and 4.3), are more efficient. Since bilinguals *are* needed, the notion of monolingual and 'objective' insight often associated with back translation is misplaced (cf. Acquadro et al., 1996).
- At the same time, the goal of providing researchers unable to read the TLT with as much (relatively unfiltered) information as possible on the text is an important one. Research on think aloud survey translation protocols (Harkness, 1996c) suggests that these provide useful information in this situation. Even if researchers do have competence in the target language, they will be likely to welcome additional input of the kind think alouds can offer.
- Finally, we note that researchers using approaches which involve back translation *as one step* frequently describe the entire procedure as back translation. This, we suggest, indicates that researchers recognise more is needed than a back translation. Different approaches to actually producing the (back) translation seem to be involved, ranging from morpheme for morpheme, 'literal'-and-stilted, to quick and free paraphrase (Acquadro et al., 1996; McKay et al., 1996; Werner and Campbell, 1970; Schoua, 1985).

5.3.2 Committee Assessments

Even when the translation has been produced by one translator, committee assessment is recommended (Guillemin et al., 1993; Acquadro et al., 1996; McKay et al., 1996). McKay et al. (1996), for example, describe a variety of group assessments in

experimental contexts, noting the usefulness at different stages of assessment of appraisal by monolinguals from the target population, bilinguals, survey design experts, interviewers, as well as people more narrowly seen as having the necessary translatory expertise to appraise target and source texts. Time, personnel and funds available for translation assessment are usually more restricted than in the McKay et al. experimental setting. A spread of expertise is clearly desirable, the question is what is most effective and viable. This depends on what is to be assessed. If, for example, the questionnaire is required to be a covert translation, understandable to a broad public, and has to follow, say, house-style question formats, it may be better to alternate discussion between those with translatory and survey design expertise on the one hand with feedback from people held to represent the target population. Assessment of instrument equivalence beyond translation adequacy also needs to be incorporated, at least if the intention is to modify the questionnaire on the basis of findings from statistical analysis.

5.3.3 Comprehension Assessment

Comprehension assessments of translations are based on the idea that if people are able to explain, describe or perform accurately on the basis of having read translated material, then the translated material accurately contains the information necessary to perform these tasks. The focus of assessment is thus on the factual information retained in a translation rather than on other aspects of equivalence or translation adequacy. These forms of assessment have been used, for example, to assess translations of instruction materials. In school 'text comprehension' testing, related procedures assess not the texts, but the *recipients* of the texts. This highlights an intrinsic source of potential error when assessment of textual adequacy is based on performance, that of discrepancies between human performance and perceptions on the one hand, and text content on the other. Brislin (1976) outlines further limitations connected to knowledge-testing and performance-testing.

In the survey context, de-briefing sessions with respondents have been used to probe their comprehension of specific items or formulations, as have focus groups in the developmental and translation stages. One advantage of these assessments is that they can be made with monolinguals. Limiting factors are the need to construct tests and questions, the costs involved, the potential impact of social desirability (and knowledge) factors, uncertainties about the representativity of input made, and the limits on the detail which can be pursued due to time, fatigue, or respondent suitability. Beyond this, too, these assessments may provide little information on 'fine tuning' aspects of text formulations. (Acquadro et al., 1996; McKay et al., 1996; Schoua-Glusberg, 1988, 1989).

5.3.4 Statistical Analyses

Statistical analyses take various forms and have different goals, as papers in this volume demonstrate. They investigate aspects of comparability and equivalence inaccessible through assessment of translation quality. Ultimately, what is needed is an approach which neither neglects evaluation of textual and communicative equivalence nor statistical assessment of measurement properties. As mentioned earlier, guidelines are needed on how best to combine these. Statistical analyses of item, battery, construct, or instrument equivalence use data from pre-tests or main study fielding. They investigate instrument quality from various perspectives on the basis of data produced across versions of the questionnaire. Similar distributions or response patterns are taken as evidence of either equivalence between SLQ and translation or as indicative of *instrument* equivalence, validity, reliability, etc. (e.g., Hulin, Drasgow and Komocar 1982; Hulin, 1987; Hazashi, Suyuki and Sasaki, 1992; Davis, 1993; Van de Vijver and Leung, 1997; Van de Vijver, this volume; Saris, this volume). Analyses of unexpected main study results can lead researchers to examine translations as a source of difference (Braun and Scott, this volume), or, indeed, visual representations (Smith, 1995). Facet theory analysis (Borg, this volume; Brislin, 1980) is seen a way of identifying information related to measurement which could help translators.

Procedures used to test translations include the following:

Split ballot assessments One group of bilinguals is administered the SLQ, another comparable group receives the TLQ. If the responses across the two groups have similar distributions or patternings – either marginals or more complex distributions – the questionnaires are considered to function as equivalent instruments. Alternatively, one group completes one half of the questionnaire in translation and the other half untranslated. The other group completes the other half of each questionnaire (source and target) and responses across the groups and the questionnaires are compared (Hulin, 1987; Hayashi, Suyuki and Sasaki, 1992; Acquadro et al., 1996).

Double administration tests Bilingual respondents complete the questionnaire in the SLQ *and* the translated version. Here, again, discrepancies across their responses are taken as indications of differences in the two versions. The remarks made earlier about differences between monolingual and bilingual responses to texts and the problem of assessing text on the basis of performance apply here, too. Moreover, what follows from finding 'differences' or 'similarities' across questionnaires remains open. Presumably, either statistical differences lead to textual examinations and these re-open the imponderables of textual assessment, or the versions are left and the data is adjusted. Double administration tests involve asking people to do something again. However, repetition itself affects responses, as research in the monolingual context has shown. Respondents asked the same questions (or who think they are asked the same questions) try to make sense of the repetition by finding new interpretations for the questions (reviews in Schwarz, 1996). It is quite possible that if asked the same things in two languages, respondents either decide that something different must be meant or decide something is behind being asked 'the same thing' twice. Either way, this may lead to different responses. Differences (and similarities) may thus not be related to features of the translation.

Post hoc analyses which examine translations on the basis of unexpected response distributions across languages are usually intended to help guide interpretation of results rather than the development or assessment of translation. Both the approach and the findings raise new questions about expected versus unexpected results and about

translation differences versus culturally differentiated responses (Braun and Scott, this volume).

6. Organising, Translation and Assessment

Decisions on which translation procedure to adopt and how to assess the translation are influenced by the time, funding, expertise and personnel available, as well as by specific aspects of a given study. Each of these factors impacts on the others. Planning for translation should be made early in the design stage. If translation is known to be a possible (but not certain) factor, contingency plans for this should cover details of people, payment and time schedules for translation and for assessment.

Time Organisation: The time allocation must include time for translation (including 'time off' before revision), assessment, revision, pre-testing, production of the final version of the translated questionnaire. If the SLQ and the translated questionnaire(s) are to be fielded simultaneously, the SLQ must be available early enough to allow for the steps above. In actuality, this is seldom the case and quality, documentation, learning curves and satisfaction suffer.

Funding Allocation: Translators in all fields of work are often poorly thought of and poorly paid. Even if translators are well-paid, translation costs are likely to be low in comparison to other costs in a survey, while poor instrument adaptation can be costly in terms of data quality. Proper selection of translators, appropriate briefing, provision of suitable materials, and adequate assessment to identify problems will contribute significantly to the success of translation products.

People and skills: Survey literature variously advocates that translators should be 'bilinguals', 'professional translators', people with knowledge of empirical social science

research, or combinations of all of these, without much indication of what, concretely, is required in terms of performance. Thus different research groups, while using similar terms, may be referring to different kinds of expertise and knowledge.

Bilingualism, for example, is a term applied to various kinds and degrees of abilities in two languages. One distinction made is between *compound* bilinguals, who learn one language after the other, and *co-ordinate* bilinguals, who learn both more or less simultaneously (cf. Wilss, 1996:206f.). Another distinguishes between bilinguals who learn a language when young and others who learn it when adult. Competencies differ in each case. Moreover, neither the degree of bilingual competence needed for survey translation nor what other competencies are needed has been empirically investigated. The high level of proficiency often glossed as 'first language proficiency' in the target language and 'good proficiency' in the source text language certainly seem to be sound requirements. The problem remains, nevertheless what is meant here by 'high level' and 'good' proficiency and how this can be assessed *before* the work is commissioned. It is important to remember, however, that some kind of 'word perfect' performance in the two languages is neither a necessary *nor* sufficient criterion. Not only is there more to translation than language competence, thinking about bilingual competence in terms of some 'word perfect' performance across languages is based on misconceptions of what is involved.

References to 'professional translators' are equally problematic. Arguably this could refer to anyone who earns their living by translating. However, it is often used or taken to imply skills and experience better associated with expert translators, that is, skilled practitioners. Gile (1995:22-23) gives a definition along 'skilled practitioner' lines (not in connection with surveys). However, this in turn raises the issue of what the yardsticks for translator skills can be (cf. Wilss, 1996:147f.). Essentially, translators should have translating skills and translating experience. However, even translation studies literature debates at length what these involve. And while experience helps develop skills, it is no guarantee for them. We consider translating skills to be more important than survey

translation experience, given that guidelines and examples could be provided for translators (but see below).

In survey literature, as in translation studies, views differ on what translators need to know about a topic in order to be able to translate well. It seems unreasonable to require that translators of philosophy must be philosophers and translators of books on calligraphy, calligraphers. On the other hand, in order to choose well between possible translation options, translators need not only to be proficient in the languages but also proficient in 'reading' the text and the text type. In other words, translators need to understand the material in order to make informed decisions. Survey translators, therefore, need, for example, a basic understanding of the measurement functions of questionnaires to be able to recognise certain problems (Hambleton, 1993; Hulin, 1987; Borg, this volume) – for which translation will or will not offer a solution. From this follows that sufficient and suitable materials should be provided and explained, so as to help translators produce a satisfactory product (McKay et al., 1996). In the field of survey research, little has been done to develop training or informational materials.

For survey translation, especially perhaps in the multi-lingual context, it is currently unrealistic to expect to find translators who have experience in survey translation, a good understanding of the relevant survey practices and are also in command of both translator skills and proficiency in the languages needed. Within translation studies, opinions differ on how best to go about training translators or assessing their work; nevertheless, a number of basic principles are generally accepted. These could be adapted for survey translation and assessment. Training and informational materials can readily be developed from survey work already done, and new (and old) source questionnaires could be annotated without undue difficulty. The modest annotations in ISSP modules, for example, could be developed systematically, as could a framework for annotating translations for posterity (and for secondary analysis).

7. Conclusion

The goals of questionnaire translation are at present under-defined. The criteria of assessment also remain unarticulated and are, it must be assumed, established on the basis of individual perceptions of 'common sense'. Thus undertaking survey translation may well seem more like setting off on an adventure with unforeseen consequences than anything resembling a systematically organised undertaking.

Questionnaires, on the other hand, *look* easy to translate. After all, questionnaire design handbooks recommend that vocabulary and syntax are kept fairly simple, sentence length is also often short, and the item content of many general population surveys refers to well-known, almost everyday, issues, institutions and entities. In certain senses, questionnaires *are* simple texts. In other respects, some of which have been mentioned here, survey translation is fairly complex. The brevity of items and the quick changes between topics across items mean that preceding sections can rarely be utilised to interpret later sections (cf. different comments on brevity in Sechrest et al., 1972; Hayashi, Suzuki and Sasaki, 1992). Cognitive research has convincingly demonstrated, on the other hand, how respondents extend common 'reading strategies' to questionnaires and thus make links between items not intended by researchers (e.g., Schwarz, 1996). In any case, whether survey translation is relatively simple or not, it involves decisions and selection, and it involves difference as well as equivalence.

When discussed, the process of survey translation is talked about in terms of finding appropriate words, phrases, or sentences in a target language, and about handling grammatical and syntactical features of sentences across languages but rarely in terms of conveying communicative functions of a source text – or source text units – in a target text. As suggested earlier, a focus on communicative function is unlikely to be compatible with literal or close translation as implemented in surveys. It is, on the other hand, central to conveying intended meanings. Whether conveying the intended meaning of a source text item results in a target language question which also taps the intended dimension or construct is a separate issue. At the same time, equivalence of dimensions or constructs to

be measured is *the* essential prerequisite in comparative cross-cultural research. Translators (and secondary analysts), therefore, need information on the dimension/construct supposed to be tapped, as well as an indication of the intended salient reading of the text for each item.

Important challenges to be met in questionnaire translation are similar to those faced in formulating monolingual questionnaires. Cognitive survey research has shown how important both the wording and arrangement of questions (item and response options) and instructions are. Designers formulate, pre-test and re-formulate in order to arrive at the most appropriate expression and arrangements for a given audience and study purpose. Optimal expression of items, instructions, and response scales is one of the tasks also faced by translators, in most cases with considerably less information about the communicative intention than in the monolingual context. Since all questions – not just poorly written ones – are open to different readings, this lack of information compromises translators' decisions about which meaning is *salient* and how best to formulate this in a second language. Without advance task specifications, translators are implicitly setting their own specifications. Providing information and documentation on all these aspects is not standard practice in survey research. Some of it would not be difficult to provide. For other information, the cross-cultural research to match available monocultural research is only beginning. Much remains to be done. Almost thirty years ago, Werner and Campbell (1970) offered to set up a clearing house on translation issues, so as to gather information needed by the scientific community. The need to investigate, document, systematise, accumulate and disseminate information is no less acute today, even if modern technology offers us tools for the job. Without this information, it is difficult to see how the high standards demanded of monocultural item formulation can be extended to decisions about and for translation, or, indeed, *against* 'mere' translation.

References

Acquadro, C., Jambon, B., Ellis, D. and Marquis, P. (1996). Language and Translation Issues. In: B. Spilker (ed.), *Quality of Life and Pharmacoeconomics in Clinical Trials* (2nd edition). Philadelphia: Lippincott-Raven.

Alwin, D.F., Braun, M., Harkness, J.A. and Scott, J. (1994). Measurement in Multi-National Surveys. In: I. Borg and P.Ph. Mohler (eds.), *Trends and Perspectives in Empirical Social Research* (pp. 26-39). Berlin: de Gruyter.

Baker, M. (1992). *In Other Words: A Coursebook on Translation.* London: Routledge.

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

Brislin, R.W. (1976). Introduction. In: R.W. Brislin (ed.), *Translation: Applications and Research.* (pp. 1-43). New York: Gardner.

Brislin, R.W. (1980). Translation and Content Analysis of Oral and Written Material. In: H.C. Triandis and J.W. Berry (eds.), *Handbook of Cross-Cultural Psychology,* (vol. 2, pp. 389-444). Boston: Allyn & Bacon.

Brislin, R.W. (1986). The Wording of Translation of Research Instruments. In: W.J. Lonner and J.W. Berry (eds.), *Field Methods in Cross-Cultural Research* (pp. 137-164). Beverly Hills: Sage.

Davis, J.A. (1993). *Memorandum to the ISSP*. Chicago: National Opinion Research Center (mimeo).

Flaherty, J., Moises G.F., Pathak, D., Mitchell, T., Wintrob, R., Richman, J.A. and Birz, S. (1988). Developing Instruments for Cross-Cultural Psychiatric Research. *The Journal of Nervous and Mental Disease 176(5)*: 257-263.

Gile, D. (1995). *Basic Concepts and Models for Interpreter and Translator Training.* Amsterdam: John Benjamins.

Guillemin, F., Bombardier, C. and Beaton, D. (1993). Cross-Cultural Adaptation of Health-Related Quality of Life Measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology 46(12)*: 1417-1432.

Gutknecht, C. and Rölle, L. (1996). *Translating by Factors.* Albany: State University of New York Press.

Gutt, A. (1991). *Translation and Relevance.* Oxford: Basil Blackwell.

Hambleton, R.K. (1993). Translating Achievement Tests for Use in Cross-National Studies. *European Journal of Psychological Assessment 9(1):* 57-68.

Hambleton, R.K. (1994). Guidelines for Adapting Educational and Psychological Tests: A progress report. *European Journal of Psychological Assessment (Bulletin of the International Test Commission) 10:* 229-244.

Harkness, J.A. (1994). Who is Who in Questionnaires? Paper presented at the World Congress of Sociology, Bielefeld.

Harkness, J.A. (1995). Getting a Word in Edgeways. Questionnaires as Texts and Discourse. Paper presented at the annual meeting of the American Association for Public Opinion Research, Fort Lauderdale, Florida.

Harkness, J.A. (1995b)

Harkness, J.A. (1996a). Minding one's P's and Q's and one's P's and C's: Gender issues in questionnaire translation. Paper presented at the World Association for Public Opinion Research meeting, Salt Lake City, Utah.

Harkness, J.A. (1996b). The (Re)Presentation of Self in Everyday Questionnaires. Paper presented at the International Sociological Association Conference on Social Science Methodology, Colchester.

Harkness, J.A. (1996c). Thinking Aloud about Survey Translation. Paper presented at the International Sociological Association Conference on Social Science Methodology, Colchester.

Harkness, J.A. and Braun, M. (in preparation). Text-based and Data-based Perspectives on Questionnaire Translation.

Harkness, J.A., Mohler, P.Ph. and McCabe, B. (1997). Towards a Manual of European Background Variables. ZUMA Report on Background Variables in a Comparative Perspective. In: J.A. Harkness, P.Ph. Mohler, and R. Thomas, *General Report on Study Programme for Quantitative Research (SPQR).* Report to the European Commission. Mannheim: ZUMA (mimeo).

Harkness, J.A., Mohler, P.Ph., Smith, T.W. and Davis, J.A. (1997). *Final Report on the Project: Research into Methodology of Intercultural Surveys (MINTS).* Mannheim: ZUMA (mimeo).

Hayashi, C., Suzuki, T. and Sasaki, M. (1992). *Data Analysis for Comparative Social Research: International Perspectives.* Amsterdam: North-Holland.

Holz-Mänttäri, J. (1984). Sichtbarmachung und Beurteilung translatorischer Leistungen bei der Ausbildung von Berufstranslatoren [The elucidation and evaluation of translation performances in translator training]. In: W. Wilss and G. Thome (eds.), *Die Theorie des Übersetzens und ihr Aufschlußwert für die Übersetzungs- und Dolmetschdidaktik* [Translation theory and its relevance for the teaching of translation and interpretation]. Tübingen: Narr.

Hönig, H. and Kussmaul, P. (1984). *Strategie der Übersetzung*. Tübingen: Narr.

House, J. (1977). *A Model for Translation Quality Assessment*. Tübingen: Narr.

Houtkoop-Steenstra, H. (1995). Conversational Problems in Telephone Survey Interviews. Paper presented at the International Conference on Survey Measurement and Process Quality, Bristol.

Hui, C.H. and Triandis, H.C. (1985). Measurement in Cross-Cultural Psychology. A Review and Comparison of Strategies. *Journal of Cross-Cultural Psychology 16(2):* 131-152.

Hulin, C.L., Drasgow, F. and Komocar, J. (1982). Applications of Item Response Theory to Analysis of Attitude Scale Translations. *Journal of Applied Psychology 67(6):* 818-825.

Hulin, C.L. (1987). A Psychometric Theory of Evaluations of Item and Scale Translations: Fidelity across languages. *Journal of Cross-Cultural Psychology 18(2):* 115-142.

Johnson, T.P., O'Rourke, D., Chavez, N. et al. (1997). Social Cognition and Response to Survey Questions Among Culturally Diverse Populations. In: Lyberg, L., Biemer, P., Collins, M. et al. (eds.), *Survey Measurement and Process Quality* (pp. 87-113). New York: Wiley & Sons

Kiraly, D. (1995). *Pathways to Translation.* Ohio: Kent University Press.

Kussmaul, P. (1986). Übersetzung als Entscheidungsprozeß. Die Rolle der Fehleranalyse in der Übersetzungsdidaktik. In: M. Snell-Hornby (ed.), *Übersetzungswissenschaft - Eine Neuorientierung* (pp. 206-229). Tübingen: Francke.

Kussmaul, P. (1995). *Training the Translator.* Amsterdam: John Benjamins.

Lyons, J. (1977). *Semantics*, Vol. 1. Cambridge: Cambridge University Press.

Maier, M.H. (1991). *The Data Game: Controversies in Social Science Statistics*. Armonk, N.Y.: Sharpe.

McKay, R.B., Breslow, M.J., Sangster, R.L., Gabbard, S.M., Reynolds, R.W., Nakamoto, J.M. and Tarnai, J. (1996). Translating Survey Questionnaires: Lessons Learned. *New Directions for Evaluation 70*: 93-105.

Newmark, P. (1988). *A Textbook of Translation*. London: Prentice Hall.

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

Przeworski, A. and Teune, H. (1970). *The Logic of Comparative Social Inquiry.* New York: John Wiley & Sons.

Reiss, K. and Vermeer, H.J. (1984). *Grundlegung einer allgemeinen Übersetzungstheorie*. Tübingen: Niemeyer.

Schaeffer, N.C. (1995). Negotiating Uncertainty: Uncertainty proposals and their disposal in standardised interviews. Paper presented at the International Conference on Survey Measurement and Process Quality, Bristol.

Schoua, A.S. (1985). *An English/Spanish Test of Decentering for the Translation of Questionnaires.* Unpublished dissertation. Northwestern University.

Schoua-Glusberg, A.S. (1988). A Focus-Group Approach to Translating Questionnaire Items. Paper presented at the annual meeting of the American Association for Public Opinion Research, Toronto.

Schoua-Glusberg, A.S. (1989a). The Spanish Version of the NHSB Questionnaire: Translation Issues. Paper commissioned under Order 263-MD-835040 for presentation at the National Institute of Health AIDS and Sexual Behavior Change Conference.

Schoua-Glusberg, A.S. (1989b). Using Ethnoscience and Focus Group Techniques in Translation. Paper presented at the annual meeting of the Central States Anthropological Society, South Bend, IN.

Schoua-Glusberg, A.S. (1992). *Report on the Translation of the Questionnaire for the National Treatment Improvement Evaluation Study.* Chicago: National Opinion Research Center (mimeo).

Schwarz, N. (1996). *Cognition and Communication. Judgmental Biases, Research Methods, and the Logic of Conversation.* Mahwah: Lawrence Erlbaum.

Sinaiko, H.W. and Brislin, R.W. (1973). Evaluating Language Translations: Experiments on three assessment methods. *Journal of Applied Psychology 57(3):* 328-334.

Snell-Hornby, M. (1988). *Translation Studies. An Integrated Approach.* Amsterdam: John Benjamins.

Sechrest, L., Fay, T.L. and Hafeez Zaidi, S.M. (1972). Problems of Translation in Cross-Cultural Research. *Journal of Cross-Cultural Psychology 3(1):* 41-56. Reprint in: L.A. Samovar and R.E. Porter (eds.) (1988), *Intercultural Communication: A Reader* (5th edition). Belmont: Wadsworth.

Smith, T.W. (1995). Little Things Matter: A sampler of how differences in questionnaire format can affect survey responses. Paper presented at the annual meeting of the American Association for Public Opinion Research, Fort Lauderdale, Florida.

Stanley, J.S. (1995a). *Results of Behavior Coding of December Agricultural Survey Interviews. Introduction, Acres Operated, and Hogs and Pigs Sections*. DCB Staff Report Number DCB-95-01. Washington, DC: United States Department of Agriculture. National Agricultural Statistics Service.

Stanley, J.S. (1995b). *Results of Behavior Coding of December Agricultural Survey Interviews. Corps Section.* DCB Staff Report Number DCB-95-02. Washington, DC: United States Department of Agriculture. National Agricultural Statistics Service.

Tanzer, N.K., Ellis, B.B., Zhang, H., Sim, C.Q.E., Broer, M. and Gittler, G. (1997). Cross-Cultural Decentering of Test Instructions in a Letter-Cancellation Test: A field test of the ITC Guidelines for test adapations. In: N.K. Tanzer (Chair): *Advances in Test Translation Methodology.* Symposium conducted at the 5th European Congress of Psychology, Dublin.

Triandis, H.C. (1994): *Culture and Social Behavior*. New York: McGraw-Hill.

Van de Vijver, F.J.R. and Hambleton, R.K. (1996). Translating Tests: Some practical guidelines. *European Psychologist 1(2):* 89-99.

Van de Vijver, F.J.R. and Leung, K. (1997). Methods and Data Analysis of Comparative Research. In: W. Berry, Y.H. Poortinga and J. Pandey: (eds.), *Handbook of Cross-Cultural Psychology* (2nd edition, vol. 1, pp. 257-300). Boston: Allyn & Bacon.

Vinay, J.P. and Darbelnet, J. (1958, translated 1995). *Comparative Stylistics of French and English. A Methodology for Translation.* Amsterdam: John Benjamins.

Werner, O. and Campbell, D. (1970). Translating, Working through Interpreters and the Problem of Decentering. In: R. Naroll and R. Cohen (eds.), *Handbook of Cultural Anthropology*. New York: American Museum of National History.

Wilss, W. (1996). *Knowledge and Skills in Translator Behavior*. Amsterdam: John Benjamins.

Multidimensional Scaling and Equivalence: Is *having a job* the same as *working*?

MICHAEL BRAUN AND JACQUELINE SCOTT

The question of functional equivalence in internationally comparative surveys is discussed from the viewpoint of secondary analysis. A number of data-analytical procedures – ranging from a comparison of means over establishing correlations with third variables in individual countries to Multidimensional Scaling (MDS) – are discussed and used to check problems in functional equivalence in an item battery on gender roles. The data base consists of several national samples (Italian, German, American, British and Hungarian) of the 1988 International Social Survey Programme study on Family and Changing Gender Roles.

It is concluded that (1) ex post checking of functional equivalence is useful and necessary because (2) the related problems might have less to do with translation than with ambiguities of the concepts and/or formulations used in source instruments and with differences in the social and economic realities in the respective countries, but that (3) even secondary-analytic strategies are as a rule not conclusive and should be supplemented with methodological experiments.

1. Introduction

Inadequacies in translations leading to a lack of functional equivalence of indicators, i.e., that the measures do not relate to the same underlying concepts, may be due to a variety of causes. These are related to different possible remedies, and they may appear in very different forms: from translation errors to slight variations in connotations. Some problems may arise due to a lack of carefully conducted translations. At first glance, this problem seems to be quite easy to handle by devoting more resources to this crucial step of the research process. *Ex ante*, i.e., in the context of questionnaire construction, the procedure of backtranslation seems to be appropriate. The basic idea is to translate the master questionnaire into the other languages and then to re-translate the translations. Any deviations between the original master questionnaire and the backtranslations are then discussed. Ideally, this is not only done to improve the translations themselves, but if

translation difficulties are found to be due to a general ambiguity of concepts/formulations used in the master questionnaire, also to modify the latter (Scheuch, 1968). In doing this, it would be possible to eliminate ambiguities which could be problematic even in the source language. However, it is not possible to guarantee functional equivalence by this method, only simple translation errors could possibly be eliminated by this procedure (see Harkness and Schoua-Glusberg, this volume).

Another, probably only partial remedy during the phase of questionnaire construction is to make the facet structure (Borg and Shye, 1995) of items explicit and to document the intended meaning of the questions, in order to give translators the necessary guidance and orientation (see Borg, this volume). Ultimately, however, the need to reappraise the problems of functional equivalence in the phase of data analysis remains (Scheuch, 1968; Smith, 1988). *Ex ante* procedures do not do away with the need for *ex post* procedures, i.e., detecting problems with functional equivalence during data analysis. As a rule, more than one indicator for a theoretical concept is necessary to do this.

A first step then is to check whether the ordering between different countries – with respect to the marginals or means (or any other adequate summary measure) – is similar between different indicators. If this appears not to be the case, especially if the mean differences have different signs or dramatically different magnitudes for items that are supposed to tap the same underlying aspect, the measures have to be treated with care. In this case, further inquiry is necessary, because that casts some doubt on whether the assumption of unidimensionality is fulfilled in all of the countries.

In addition, the regression of attitudes on demographic variables such as age, education or income can be used to help clarify the meaning of an attitudinal question. High correlations with age and education often give some support to an ideological interpretation, while correlation with income (net of education) make a self-interest-based understanding more likely. The question remains whether more ideological and more interest-based variables can be unambiguously singled out by this procedure. It is clear that a distinction between ideological and interest-based variables is likely to vary from

country to country or, more precisely, by the 'meaning' of the socio-demographic explicative variables. This is because it may be unclear whether what we see is generated by comparable groups giving different responses or by basically different groups in the single countries (Scheuch, 1968). Therefore, using variables like age, education or substantive variables known to measure values to inquire into the nature of an item will not always yield conclusive results. For example, in East Germany, different educational groups are closer to each other with regard to whether a pre-school child is likely to suffer if his or her mother works than are the respective groups in West Germany. (There is a 7 percentage points difference between lowest and highest educational qualifications in the East and a 15 percentage points difference in the West (cf. Braun and Bandilla, 1992). This could be due either to a different meaning of *suffering* in the two parts of Germany or to a difference with respect to the relationship between education and the attitudinal item (in the language adopted here, the latter would indicate different 'meanings' of the variable education), or both.

A wide range of more sophisticated statistical techniques (for an overview see Van de Vijver and Leung, 1997) have been proposed and used in the area of internationally comparative research, such as confirmatory factor analysis (Watkins, 1989) and psychometric approaches (Hulin, 1987; Van de Vijver and Poortinga, 1997). In this paper, in addition to simple procedures such as comparing marginals, we use Multidimensional Scaling (MDS), which has at least two advantages. First, MDS is easy to use, so that it can be applied routinely in analyzing item batteries. Second, MDS provides an instructive impression of the structure of the data. We outline details of this approach in the next section.

We apply this method to some potential problems of functional equivalence in the analysis of gender roles. The examples come from our substantive work (Alwin, Braun and Scott, 1992; Braun, Scott and Alwin (1994); Braun, Scott and Alwin, 1993). Part of the problems documented here became visible in the course of a substantive analysis of the data. Others were discovered by carrying out a backtranslation (admittedly while also knowing the source text). We also point out several ambiguities in the formulations of the

British English master questions, even if potential problems resulting from them could not be tracked in the data. Most of the issues raised here have to be tested, they should be understood as hypotheses on the effects of question wording.

2. Data and Methods

The data come from the 1988 International Social Survey Programme (ISSP) module on Family and Changing Gender Roles (see Braun, 1994; Davis and Jowell, 1989). Eight countries participated in this study, but for our purpose, it is sufficient to concentrate on the Italian, the American and the German data, treating the British and Hungarian data as ancillary and ignoring the other countries. The 1988 questionnaire includes a battery of nine items designed to measure gender-role attitudes. These can be conceived of as tapping three different aspects: consequences of women working, gender ideology and the economic importance of work (Braun, Scott and Alwin, 1994). In the following, we concentrate on four items which present some peculiarities, especially in the Italian data. The other items are explained as needed. The items we focus on are related to the consequences and gender-ideology aspects. In the table below, italics are used to highlight the part of the items focused on.

First we consider the item *A woman and her family will all be happier if she goes out to work*. It is likely to tap several aspects: the consequences of women working for family life, gender-role ideology, and the possible contribution of a double income to the economic well-being of the family. Thus, in terms of the English text, it is obvious that this item might be highly ambiguous. Under certain conditions, ambiguity as such does not pose a big problem, provided people nevertheless understand the item in the same way, perhaps even mixing up different interpretations in a kind of summary statement (constant ambiguity mix). The problem arises if different subgroups in society or people in different countries understand the question in a different way. The problem with international comparability may be exacerbated because translations of ambiguous items are likely to emphasize some aspects more than others.

Table 1: Four language versions of four items from the 1988 ISSP module

Item 1	**Family suffers**
English source	All in all, family life suffers when the woman *has a full-time job*.
Italian	Tutto considerato la vita familiare risente negativamente se la madre *lavora a tiempo pieno*.
German	Alles in allem: Das Familienleben leidet darunter, wenn die Frau *voll berufstätig ist*.
Hungarian	A család élete megsínyli, ha a feleség *teljes munkaidöben dolgozik.*
Item 2	**Woman happier**
English source	A woman and her family will all be happier if she *goes out to work*.
Italian	Una donna e i suoi familiari sono più sereni se la donna *ha un lavoro*.
German	Wenn eine Frau *berufstätig ist*, wird sie und ihre Familie glücklicher sein.
Hungarian	A nöknek is és a családnak is job, ha a nök eljárnak dolgozni.
Item 3	**Children better**
English source	A job is all right, but what most women really want is a home and children.
Italian	Un lavoro è una buona cosa ma quello che realmente vuole la maggioranza delle donne è una casa e dei bambini.
German	Einen Beruf zu haben ist ja ganz schön, aber das, was die meisten Frauen wirklich wollen, sind ein Heim und Kinder.
Hungarian	Állásban lenni is fontos lehet, de a legtöbb nönek az az igazi vágya, hogy otthona és gyermeke legyen.
Item 4	**Housework as fulfilling**
English source	Being a housewife is just as fulfilling *as working for pay*.
Italian	Essere una casalinga è altrettanto soddisfacente quanto *avere un lavoro retribuito*.
German	Hausfrau zu sein ist genauso erfüllend wie *gegen Bezahlung zu arbeiten*.
Hungarian	A háziasszonyi teendöket jól ellátni legalább akkora teljesítmény, mint a fizetésért végzett munka.

We found that the Italian translation diverges from the English source rendering *if she goes out to work* by *se la donna ha un lavoro* (translated word-by-word: 'if the woman

has a job'). The Italian formulation may prompt, for some respondents, a comparison of employment and unemployment rather than a comparison with the voluntarily chosen homemaker role. Italian colleagues felt that in Italian 'to work' and 'to have a job' are treated and understood as equivalent expressions. However, that does not rule out the possibility that, at least for some respondents, a comparison of employment with unemployment may come to mind, thus moving the meaning of the item away from the consequences/gender-ideology aspect in the direction of the economic-function aspect. Thus, given that the original English formulation is itself ambiguous, the Italian translation only changes the ambiguity mix. There is an additional peculiarity with the Italian rendering which might or might not have an effect in the same direction. While the English formulation alludes to work outside the home *(goes out)*, some translations (both the Italian and the German versions) are more likely to include work that is done at home. This may make respondents more likely to agree to this item.

In a second item, the Italian translation deviates in a similar way from the English master questionnaire. In the item *Being a housewife is just as fulfilling as working for pay*, *working for pay* is again rendered by *avere un lavoro retribuito* (word-for-word: 'having a paid job'). Here, too, economic consequences might come to mind more easily in the Italian version.

Thus we have two pairs of items which, on the surface, appear to be very similar with regard to their intended meanings. The first pair is:

* *All in all, family life suffers when the woman has a full time job.*
* *A woman and her family will all be happier if she goes out to work.*

And the second pair is:

* *A job is all right, but what most women really want is a home and children.*
* *Being a housewife is just as fulfilling as working for pay.*

The English version uses *to work* for some items and *to have a job* for others. It is only in the second item of each pair that the Italian version deviates in the described way from the English, i.e., refers in Italian to 'to have a job' instead of 'to work'.

MDS (see Borg and Groenen, 1997) can help us to track the effects of these translation problems in the data. Technically, what MDS does is represent the intercorrelations of the items in a multidimensional space. Correlations correspond to the distances between the items which are drawn as points. The interpretation of the MDS representation focuses on the correspondence between geometrical characteristics of the configuration and the substantive characteristics of the items. The lines entered in the MDS representations are theoretically derived dividing lines between items; ideally the partitions of the space should be achievable by straight lines (only the MDS representation of the Hungarian data below does not meet that criterion).

3. Results

Do the empirical data support the assumption that the Italian item is understood in a different way from the English or the German? Let us examine the marginals first. While the Italians show quite traditional gender-role attitudes in general, on the items 2 and 4 in Tables 1 and 2, where the Italian version is different ('Woman happier' and 'Housework as fulfilling'), they turn out to be the least traditional nation (see Table 2).

Table 2: Means of gender-role items in the US, Germany and Italy (items recoded such that high values correspond to non-traditional attitudes)

Items	USA	Germany	Italy	North Italy	South Italy
Family suffers (1)	3.2	2.5	2.4	2.5	2.1
Woman happier (2)	2.7	2.6	3.3	3.3	3.4
Children better (3)	3.1	2.9	2.7	2.9	2.4
Housew. as fulfilling (4)	2.6	2.7	3.1	3.1	2.9

Moreover, the same trend can be seen within Italy, with Southern Italians appearing to be slightly less traditional in terms of these items than the Northern Italians, while for the other two questions Southern Italians are clearly more traditional than Northern.

The next question is: Apart from the marginals, is the cognitive representation of the items different in Italy, too? To discuss this aspect we turn to 2-dimensional Multidimensional Scaling. The graph for Northern Italy (Figure 1) shows that the items can be easily grouped preserving the theoretically postulated partitioning of the items into consequences, gender-ideology and economic-function regions (the partitioning is not produced by the program, but by the researchers reflecting theoretical considerations). The items in bold letters in the graph refer to the items affected by the deviating translation, the items in italics are those described above which do not suffer from this problem. The different partitionings are underlined. The additional items of the *consequences class* (on the left-hand side) are:

* Child suffers: *A pre-school child is likely to suffer if his or her mother works*.
* Warm relation: *A working mother can establish just as warm and secure a relationship with her children as a mother who does not work*.

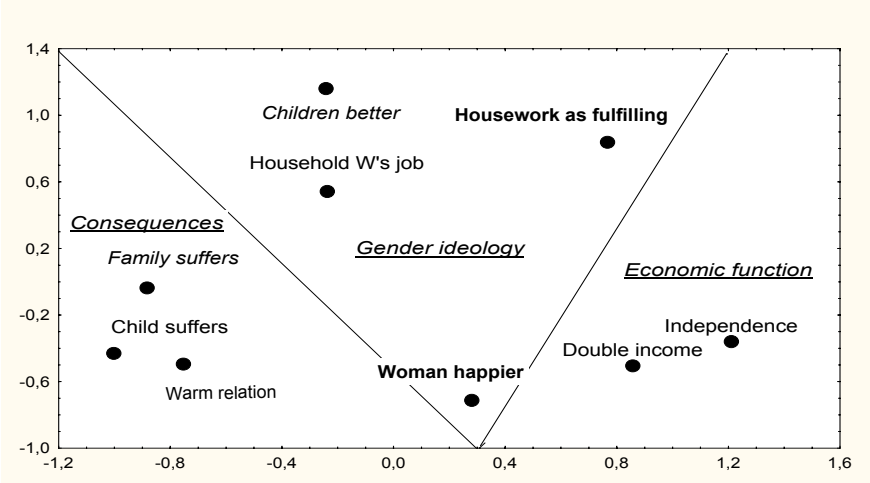
The additional item of the *gender-ideology class* which appears in the middle is:

* Household W's job: *A husband's job is to earn money; a wife's job is to look after the home and the family*.

Finally, the two items of the *economic-function class* on the right-hand side are:

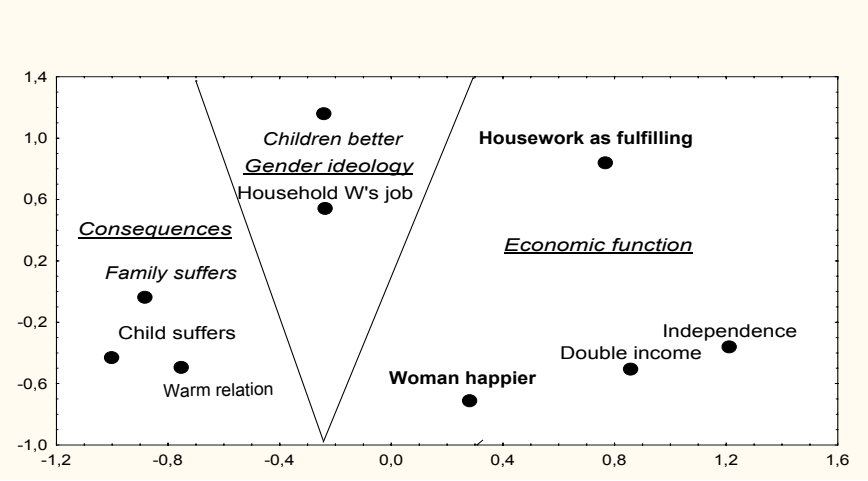
* Independence: *Having a job is the best way for a woman to be an independent person*.
* Double income: *Both the husband and wife should contribute to the household income.*

Figure 1: 2-dimensional MDS for Northern Italy



However, as Figure 2 shows, it is also possible to draw a dividing line between the consequences and gender-ideology regions on the one hand, and the economic-function region on the other, with the two items affected by the deviating translation located in the latter region. In fact, the two affected items are closer to the items of the economic-function class than to the remaining items of the consequences and gender-ideology class.

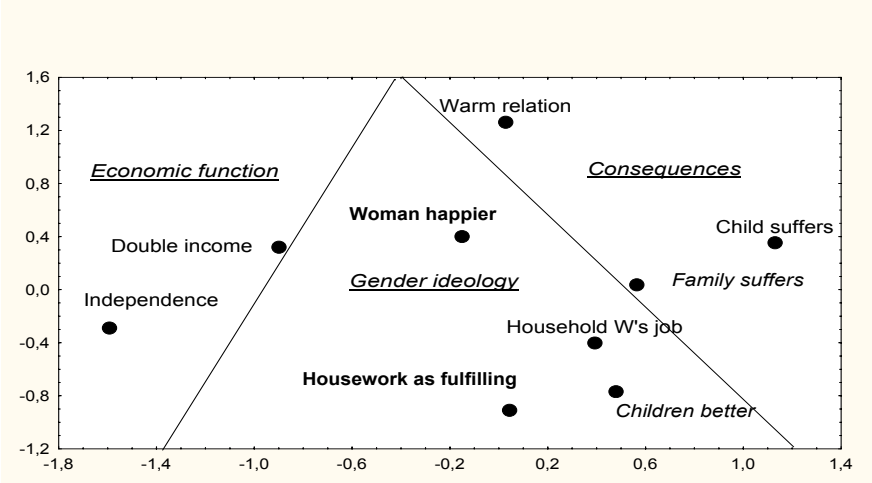
Figure 2: 2-dimensional MDS for Northern Italy, alternative partitioning



Thus, the problematic items do seem to belong more in the economic than the gender-ideology region. This leads us to ask whether this tendency is more pronounced in Southern Italy, where unemployment is higher and economic aspects might come more easily to mind than in Northern Italy. Though the pictures for Southern and Northern Italy look roughly the same, the correlations between the two items forming the above-mentioned pairs are somewhat higher in the North (for ‘Woman happier’ and ‘Family suffers’, .25 in the North and .22 in the South, and for ‘Housework as fulfilling’ and ‘Children better’, .38 in the North and .21 in the South.)

As argued above, the ambiguity of the source questionnaire items suggests that there could be a similar effect in other countries, too, with the problematic items being closer to the economic-function region. The graph for Germany (Figure 3) illustrates that the problematic items are closely related to the economic-function region, despite their 'natural home' for Germany being closer to the consequences and gender-ideology regions.

Figure 3: 2-dimensional MDS for Germany



Unfortunately, the graph for the United States (Figure 4 below) does not show this as clearly. The problematic items turn out to be located somewhere in the middle, between consequences items and gender-ideology items on the one hand, and items from the economic-function class on the other. This, of course, makes our argument much less conclusive. Our *post hoc* conjecture is that, apart from language, the interpretation of the problematic items is also affected by social reality. In a society where female labor-force participation is perceived as 'normal', the gender-ideology interpretation of the problematic questions might be less salient than an economic interpretation. On the other hand, in a society where paid work is seen as contrary to the nature of women, a gender-ideology interpretation is more likely. Because attitudes towards female labor-force participation are, on the whole, more liberal in the US than in the other countries which participated in the 1988 survey, the American interpretation of the two items might be more in terms of economic function. In countries like Italy, where attitudes are in general more traditional, the gender-ideology interpretation should have been more dominant - had the translation not encouraged an economic interpretation.

Figure 4: 2-dimensional MDS for the United States

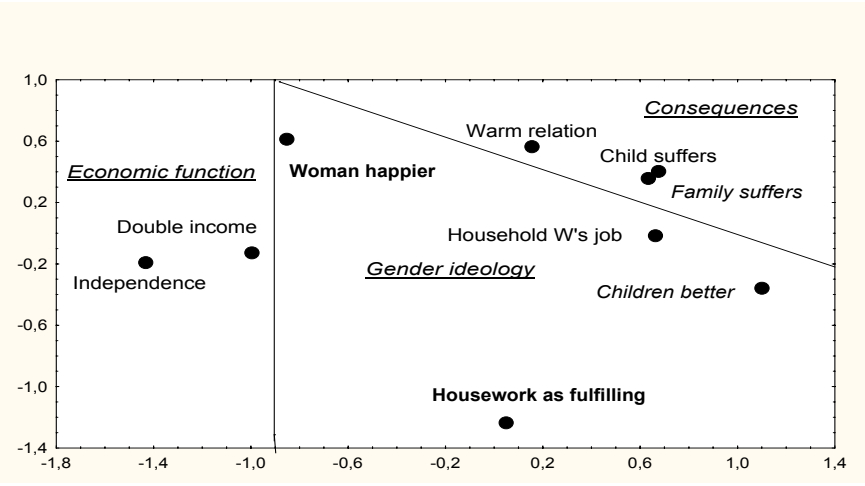
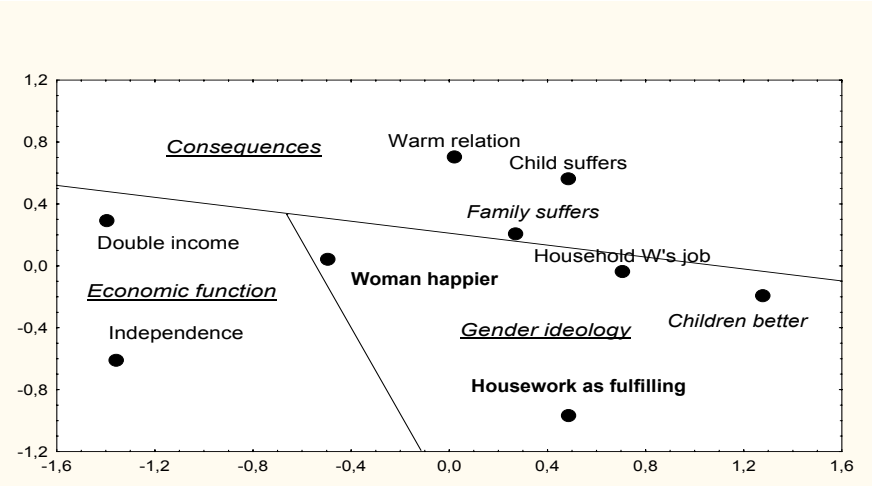


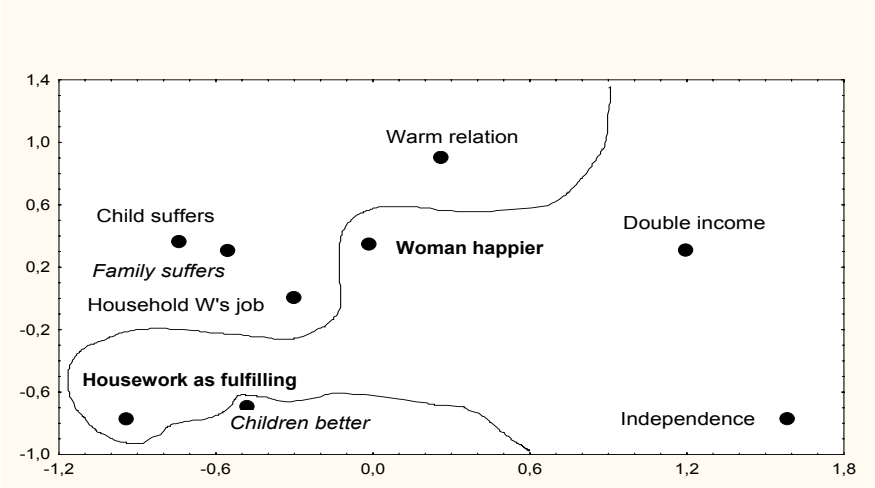
Figure 5: 2-dimensional MDS for Great Britain



The structure for the second English-speaking country, Britain, illustrated by Figure 5 below, lies somewhat in between those for the United States and Germany. The British show a level of acceptance for female labor-force participation similar to that of Americans. On other aspects, however, such as the perceived economic necessity of work for women and the actual level of female labor-force participation, Britain is somewhat between the United States and Germany.

As yet, we have little conclusive empirical evidence to add in support of our *post hoc* hypothesis. The Hungarian data would be an ideal candidate for further clarification, because in 1988, when the data was collected, women were obliged to work and there was virtually universal female labor-force participation. Unemployment was virtually non-existent and the attitudes of the population were extremely traditional with regard to female labor-force participation. Under these circumstances we would not expect the two problematic items to receive an economic interpretation. However, the Hungarian data is affected by a serious translation error with one of the crucial items (see below). Nevertheless, the Hungarian results are worth presenting.

Figure 6: 2-dimensional MDS for Hungary



The graph for Hungary (Figure 6) fits perfectly with our expectations. No sensible dividing line could be drawn to bring the problematic items into the economic-function region (Figure 6 shows how a dividing line would have to look). As mentioned, there is a problem with the Hungarian translation: In Hungary *Being a housewife is just as fulfilling as working for pay* was translated to something like 'To do the job of a homemaker well is at least as big an achievement as paid work'. The translation focuses on *fulfill* in the sense of 'carry out a task' while the source item uses *fulfill* in emotional psychological terms. Correspondingly, for this item, it does not make much sense to compare the attitudes of the Hungarians with those of the remaining nations. Nevertheless, on the basis of an analysis of the correlations, the bias in the Hungarian translation cannot be detected (i.e., the correlations between this wrongly translated item and other items is similar to those obtained from other countries). The most likely reason for this is that attitudes measured by the original source questionnaire question and the wrong translation may be highly correlated: Respondents who think that looking after the family and house is as big an achievement as paid work should also be more likely to think that this is just as fulfilling as working for pay. What should remain different, however, are the marginals – and they are. To conclude, while the marginals for the Hungarians are incomparable with the other countries, the concept measured is very likely to be the same. Therefore, we contend that the Hungarian case lends at least some support to our interpretation.

4. General Discussion

We have demonstrated the importance of checking for problems of functional equivalence at the data-analysis stage. We advocate the use of MDS at different stages in data analysis. It should be used either if the marginals yield unexpected results or to check whether problems in translation are evident in the data. However, it could also be routinely used before the substantive analysis of any battery of items because functional equivalence could be violated, even if the marginals do not look suspicious and translation problems are not detected.

It will be clear that some of the usual remedies for problems of functional equivalence might have been of little help. In the present case, giving more attention to translation and using backtranslation would not have guaranteed functional equivalence, because here we are not dealing with simple translation errors which could have been eliminated by this procedure. Moreover, the English formulations are, themselves, vague or ambiguous. As a result, even in English-speaking countries, they may be interpreted in different ways by different respondents. Perhaps the most important aspect, however, is the way social reality and language interact. Identical questions might be understood differently against different backgrounds. A reading of *to have a job* as referring to a contrast between a voluntarily chosen role as a worker and unemployment requires that unemployment is seen as a possible outcome. The degree to which such notions exist in different cultures varies considerably and it is very difficult to anticipate the effects of such variations when drafting a questionnaire and adapting it for different cultures. Thus, exploring problems of functional equivalence is a necessary part of data analysis; otherwise substantive interpretations may well be misleading. Such investigations will lead, incrementally, to improving the validity of the data and, as a by-product, investigators can gain new insights into the methodology of intercultural research. However, there is no guarantee that secondary-analytic techniques will bring conclusive results. The present discussion illustrates only too well that alternative interpretations may remain. Therefore, we advocate that split-ballot methodological experiments are also used so as to identify effects resulting from question format and wording and also to help disentangle possible artefactual and substantive explanations.

References

Alwin, D.F., Braun, M. and Scott, J. (1992). The Separation of Work and the Family: Attitudes towards women's labour-force participation in Germany, Great Britain, and the United States. *European Sociological Review 8*: 13-37.

Alwin, D.F., Braun, M., Harkness, J.A. and Scott, J. (1994). Measurement in Multi-National Surveys. In: I. Borg and P.Ph. Mohler (eds.), *Trends and Perspectives in Empirical Social Rresearch* (pp. 26-39). New York: de Gruyter.

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*. New York: Springer.

Borg, I. and Shye, S. (1995). *Facet Theory: Form and Content*. Newbury Park: Sage.

Braun, M. (1993). Potential Problems of Functional Equivalence in ISSP 88 (Family and Changing Gender Roles). Paper presented at the International Social Survey Programme Scientific Meeting, Chicago.

Braun, M. (1994). The International Social Survey Programme (ISSP). In: P. Flora, F. Kraus, H.-H. Noll and F. Rothenbacher (eds.), *Social Statistics and Social Reporting in and for Europe* (pp. 305-311). Bonn: Informationszentrum Sozialwissenschaften.

Braun, M. and Bandilla, W. (1992). Familie und Rolle der Frau. In: Statistisches Bundesamt (ed.), *Datenreport 1992. Zahlen und Fakten über die Bundesrepublik Deutschland.* Bonn: Bundeszentrale für politische Bildung.

Braun, M., Scott, J. and Alwin, D.F. (1993). Attitudes on Marriage and Divorce in Comparative Perspective. Paper presented at the International Social Survey Programme Scientific Meeting, Chicago.

Braun, M., Scott, J. and Alwin, D.F. (1994). Economic Necessity or Self-Actualization? Attitudes toward Women's Labour-Force Participation in East and West Germany. *European Sociological Review 10*: 29-47.

Davis, J.A. and Jowell, R. (1989). Measuring National Differences - An Introduction to the International Social Survey Programme. In: R. Jowell, S. Witherspoon and L. Brook (eds.), *British Social Attitudes - special international report* (pp. 1-13). Aldershot: Gower.

Hulin, C.L. (1987). Psychometric Theory of Item and Scale Translations: Equivalence across languages. *Journal of Cross-Cultural Psychology 18*: 115-142.

Küchler, M. (1987). The Utility of Surveys for Cross-National Research. *Social Science Research 16*: 229-244.

Scheuch, E.K. (1968). The Cross-Cultural Use of Sample Surveys: Problems of comparability. In: S. Rokkan (ed.), *Comparative Research across Cultures and Nations* (pp. 176-209). Paris: Mouton.

Smith, T.W. (1988). The Ups and Downs of Cross-National Survey Research. GSS Cross-National Report No. 8. Chicago: National Opinion Research Center.

Van de Vijver, F.J.R. and Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research*. London: Sage.

Van de Vijver, F.J.R. and Poortinga, Y.H. (1997). Towards an Integrated Analysis of Bias in Cross-Cultural Assessment. *European Journal of Psychological Assessment 13*: 29-37.

Watkins, D. (1989). The Role of Confirmatory Factor Analysis in Cross-Cultural Research. *International Journal of Psychology 24*: 685-701.

A Facet-Theoretical Approach to Item Equivalency

INGWER BORG

Abstract. Three notions of item equivalency are distinguished. They correspond to the back-translation approach, the psychometric IRT approach, and the facet-theoretical approach. The latter defines equivalent item as items that answer the same questions. The question, then, is explicated in terms of its design. This yields the item's blueprint. One can extract such blueprints by studying given items, but the result is generally not unique. Nevertheless, it makes it possible to predict empirical regularities for the items and, therefore, tests for equivalency. If the tests fail, however, item non-equivalency is just one possible explanation. Design-equivalency is, on the other hand, a definitional issue, not an empirical one. The enmpirical issue is the design's usefulness for a particular purpose, usually for answering the research question.

1. Definitions of item equivalency

What one ideally wants in cross-cultural surveys are items that are equivalent in the different language versions of the questionnaire. What does that mean? One rather obvious approach to item equivalency is the operational requirement to first translate an item into the other language and then translate this item back into the original language. The backtranslation should be highly similar to the original item. One of the problems with this approach is it merely guarantees a one-to-many mapping of item wordings. That is, there may be more than just one proper translation of the item, even though they all translate back to the original item. For example, I have been told (Hess Medler, 1993) that if one translates the question 'How are you doing these days?' into Spanish, one has two options: one that asks about the respondent's emotional well-being, another that asks

about his or her objective-material well-being. Translated back into English, they both lead to a question similar to the one we started with.

Moreover, there is, of course, always the possibility that translations are difficult or even impossible because the item addresses issues or concepts that have no meaning in the other language or culture. A less dramatic but common case is the challenge of translating a Likert rating scale, where one wonders whether 'strongly agree' is properly translated into German by 'stimme voll und ganz zu'. This brings in a new, and deeper, issue, one that is addressed by the psychometric approach to item equivalency, which requires that "equivalent items ... evoke a specified response, from the set of permissible responses, with the same probability among individuals with equivalent amounts of the characteristic assessed by the item" (Hulin, 1987, p. 123). The extent to which this is true can be checked via (logistic) regression of the observed response scores for an item onto the estimated 'amount of the characteristic' or by simply comparing item statistics (e.g., the mean or the rank order of mean values of a homogeneous battery of items).

Yet, this IRT (item-response theoretical) approach rests on a statistical model, where "one of the critical assumptions ... is that the latent trait space is unidimensional" (Hulin et al., 1982, p. 823). Hence, the assessment of equivalency is conditional to the validity of the assumed model. The issue is addressed in great detail in the IRT literature, and a variety of models have been proposed. All models, however, are dimensional ones. More importantly, the "motive" of the empirical inquiry does not play any role. In that sense, the IRT approach resembles almost all schools of measurement. They conceive of measurement as a process where one first builds an all-purpose measurement "instrument" or "scale". The instrument is put on the shelf, ready for future utilization. There are a number of tomes in which one can look up such instruments and their psychometric properties; one is the three-volume "ZUMA-Skalenhandbuch" (Allmendinger et al., 1983) which is now under revision with an important change of emphasis, i.e. turning it into a handbook of "items" rather than "scales".

Collections of measurement instruments are useful in applied research. They represent, in a sense, an engineering approach, providing "tools" that have been shown to "work". In science, however, one wants more than just predictive validity. One really wants to establish (empirical) laws that relate to theories. Thus, whether implicitly or explicitly, instruments and items in basic science are never formulated in isolation: the researcher formulates items with some lawfulness in mind, a hypothesis that relates observations on these items to other items and to definitional systems. The hypothesis precedes the items and the particular items used in the empirical investigation are almost always only a sample from a huge *universe* of items. What governs the construction or selection of items is the structural hypothesis.

Hence, I propose that an important aspect of item equivalency should be whether corresponding items *both answer the same substantive-scientific question*. Equivalency of items should therefore be considered in the context of *what it is that the researcher wants to know*. In the above example of translating 'How are you doing these days?' into a Spanish language item, for example, one would have to know whether the researcher wants to assess emotional or material well-being, and, indeed, whether assessing this particular issue is important for the hypothesized lawfulness. If this is known, the translations can be checked against this criterion. In fact, we may decide not to translate the item literally but to rewrite it in a particular subcultural jargon. As long as it assesses the particular type of well-being we want to assess, the phrasing of the item does not matter.

Without knowing the intent of the question, translating and back-translating items may only preserve equivalency of words. And items with similar statistical properties, while both satisfying the same formal model, usually ignore the issue of the universe of content and, in any case, the wider structural hypothesis. This amounts to a sterile form-precedes-content approach to item construction, where the formal machinery is guaranteed to generate a battery of one- or multi-dimensional scales simply by trimming content to the

statistical model. In the end, it remains unclear what exactly is being assessed by such items.

Yet, viewing item equivalency from this perspective, one notes immediately that neither statistical models nor linguistic theories nor any other extrinsic scaffolding suffices in providing good translations. What is needed is that the researcher esplicates his or her research question.

2. Explicating the item's blueprint

The above argument may seem artifical and exaggerated, because in most research in the social sciences, the overall research question is stated quite explicitly. However, it is also true that what each individual item is supposed to assess is almost never explicated – except in experimental research! In experiments, the items are experimental conditions which are typically well-designed. Each such condition asks a particular research question and formulates what is to be recorded as an answer.

In survey research, in contrast, items are typically constructed by mixing intuition, factor-analytic thinking, and, possibly, empirical evidence on certain item statistics. Nevertheless, there is always an *implicit* item design. It may have gotten blurred by statistical tinkering with the item pool - such as rephrasing items that are "factorially ambigious" or even eliminating items that are not well-"explained" by the space spanned by the first few principal components - ,but one can often uncover an implicit item design by carefully studying the items with respect to the semantic variables that are systematically varied throughout the items. Such analyses may even help the researcher to come up with items that focus more sharply on what he or she wants to know.

Consider the following case. Bastide & van den Berghe (1957, p. 690) set out "to determine the patterns of race relations in the white middle class of Sao Paulo". They collected empirical data using items that they categorized into four types:

(a) A list of "stereotypes" where the respondent was asked whether he considered Blacks inferior, equal, or superior to Whites in some sense;
(b) Items on social norms of behavior, such as 'Should Whites and Blacks exchange courtesy visits?'
(c) Items on "actual behavior" of the subjects;
(d) Items of "hypothetical personal behavior", such as 'Would you go out with a black person?'

These four types imply, of course, rules for constructing or culling relevant items. However, these rules are not further explicated, and so translating items such as those shown above in (b) and (d) is unnecessarily difficult. What is meant by 'courtesy visits', what behavior does 'go out with' refer to? These are rather obvious problems, but what is more important is that it remains unclear what roles 'courtesy visits' and 'go out with' play in the context of what the researchers sets out to study, i.e. “patterns of race relations”. In other words, is it important that the items contain these semantic elements? As a further example, consider the item 'Are Whites more intelligent than Blacks?'. This seems relatively easy to translate, but is it important that we explicitly refer to 'intelligence'? Or does this item attempt to provide just one piece of evidence in an effort to assess general feelings of superiority of white persons relative to black persons?

Guttman (1959), reanalyzing the Bastide & van den Berghe study, attempted to abstract some of the distinctions made by the item classes (a)-(d). The approach for doing this is actually quite simple. The first step consists of writing each item type as a complete sentence so that the different sentences are structurally as similar as possible. For the Bastide & van den Berghe items, Guttman proposed the following scheme:

1. Belief of a white person that Whites are superior to Blacks on desirable traits.
2. Belief of a white person that Whites should socially interact with Blacks.
3. Belief of a white person that he or she would socially interact with Blacks.
4. Overt action of a white person in the domain of social interactions with Blacks.

The four sentences can be interpreted as rules for allocating given items to a particular item type or as rules for constructing particular types of items. For example, one finds that the item 'Would you go out with a black person?' belongs to class 3.

Figure 1. A mapping sentence for the Bastide & van den Berghe items on patterns of race relations in the white middle class of Sao Paulo.

		A=modality							
	(	a_1 = belief	)						
The	(	a_2 = overt action	)	of a white person (p) vis-à-vis					
	(	a_3 = emotion	)						

	B=relation			*C=reference group*		
(	b_1=comparison	)		(	c_1=Whites in general	)	
(		)	of	(		)	with Blacks
(	b_2=interaction	)		(	c_2=himself/herself	)	

		range		
	(	very positive	)	
➔	(	to	)	towards the relation of the reference
	(	very negative	)	group to Blacks

The next step is to analyze in what ways the four classes of items differ among each other by asking what semantic dimensions are *systematically* varied over the item classes. The semantic material contained in the items that is unsystematic is either likely not to be of direct importance to the scientific question addressed by those who formulated the items or it may reflect an unsystematic item design.

Let us extract what is varied systematically over the four item types. We note, first of all, that the first three item classes assess 'beliefs', the fourth 'overt action'. This constitutes

the first *facet* of the item's blueprint. In facet theory, we write this facet in set-theoretical notation as A = {belief, overt action} and assign to it the name 'modality (of attitudinal behavior)'. Then, we note that the first two item types refer to Whites in general, the latter two to the respondent him/herself. This constitutes facet B = 'reference group' = {Whites, respondent him/herself}. Finally, the first item class assesses comparison behavior, the other item classes refer to interaction behavior. Hence, facet C = 'relation (of respondent to reference group)' = {compare (with respect to desirable traits), interact}. Note that the facets thus extracted reflect a particular perspective, namely the perspective of a psychologist who uses a particular technical language. Notions such as 'belief' or 'overt action' have technical meanings in psychology.

Conceptual clarity can be further enhanced by not only listing the various facets, but by interrelating them within a particular framework, a mapping sentence. This also forces one to explicate the range of the items ("the response scale"). In the given case, one such mapping sentence is shown in figure 1the following one:

The mapping sentence shows that the items of this study all assess attitudes of the respondent towards different forms of behavior of Whites towards Blacks, because they all assess the extent to which a behavior of a reference group is positive or negative towards a common object (Borg & Shye, 1995). Therefore, we may immediately extend facet A to include the usual third "component" of attitudes, i.e., emotions.

Promoting the formality of the item's design reveals, moreover that the item types are not well-designed in one important aspect. Item type 2, by referring to "should" behavior, refers to norms on interracial behavior, while the other items refer to actual behavior or to behavior that is probable. This is a theoretically important distinction, and the translator must know whether it is also important to the researcher. If the original item is vague, and if its measurement intention remains hidden, the translated item is likely to be unclear in the desired research sense. One might decide, therefore, to express the additional distinction just noticed by introducing a fourth facet, 'factuality of the interracial relation' = {certainly exists, presumably exists, is desirable}. The translator, of course, cannot

figure out for him/herself what role this facet plays in the empirical inquiry. It is the task of the researcher to clarify this issue.

Let us now return to the question of what the researcher wants to know. Bastide & van den Berghe wanted to study “patterns of race relations”. Our analyses show that their item typology constrains this research question effectively to a study of patterns of attitudes on race relations. It thereby builds a bridge to what is already known about attitudes in general. Thus one “pattern” hypothesis is that such attitudinal items are positively intercorrelated, reflecting the first law of attitudes (Borg & Shye, 1995). Another “pattern” hypothesis is that the facets built into the items or, expressed differently, projected into these items by a psychologist’s interpretation, are reflected in the structure of the data, in the sense that the items can be statistically discriminated along these facets. One way of testing this discriminability is to ask if the items form non-overlapping regions in a multidimensional scaling representation (Borg & Groenen, 1997).

3. Some comments on mapping sentences

Asking the translator (and not the researcher)to explicate the item’s design is, of course, not the ideal way to proceed. Yet, from experience, I know that this situation is not as unlikely as it may seem. I have been asked on several occasions to provide a facet analysis for a set of items given to me, without being told the purpose of the items in any but an exceedingly vague way. We saw in the above that such a facet analysis is possible. However, it should be obvious that the mapping sentence we came up with is not the only one that is conceivable. Indeed, we pointed out that this particular mapping sentence is one that relates to a psychological background, with technical notions that are obvious only to the psychologist. But even psychologists would, of course, not always arrive at the same facets, because different psychologists operate within different theories.

Borg (1991), for example, classified work value items ('How important is work outcome XYZ for you?', with the range 'not important ... very important') in a variety of different

ways, each reflecting one particular theory. The mapping sentence used in this analysis distinguished two facets of the work outcomes; 'need served by work outcome' and 'performance-dependency of work outcome'. Since various classification schemes for needs exist (e.g., the Maslow hierarchy, Alderfer's ERG theory, and Herzberg's dichotomy), there are also different ways to facetize work values. Similar arguments hold for the second facet. Which of these possible facetizations is to be preferred depends on the purpose of the study. Indeed, depending on the purpose, one may opt for different sets of facets. An analogy in this context is the classification of matter, where the purpose at hand determines whether Mendeleev's periodic table of chemical elements is better than, say, the archaic earth-wind-fire-water distinction.

Apart from the purpose, however, a number of general criteria can be formulated for judging the goodness of a mapping sentence. Since the mapping sentence is a definition and not a hypothesis, "truth" is not an issue. Clearness, however, *is* relevant. Further criteria are its reliability for classifying items, for constructing items, and for communicating about items (among experts). Ideally, a mapping sentence should also be empirically useful. Empirical usefulness is the testable hypothesis associated with the mapping sentence definition. It predicts, among other things, that the conceptual structure induced by the mapping sentence into a pool of items is mirrored in a corresponding structure of the observations.

One cannot expect that a translator by him/herself will, in general, come up with anything else but a mapping sentence that is "superficial", focusing on rather concrete distinctions made by the items and considering their *apparent* purpose only. A "deeper" mapping sentence usually involves considerable expertise. Moreover, good mapping sentences typically develop over time in bidirectional, mutually constraining interaction between conceptual-theoretical work and empirical testing, a cooperative alternation which almost always involves many mapping sentence modifications and item reformulations. Advanced mapping sentences, therefore, become rather abstract and hard to understand for the uninitiated. Translators must ultimately not only be knowledgable about the

languages involved in the translation task, but also at least be able to understand what the researcher wants to know. This requires substantive expertise.

The mapping sentence is the items' blueprint, but this blueprint is often not fully developed before one begins to construct items. Typically, one begins with a vague notion of commonality and then writes down a set of items. One then studies these items – similar to what we did above – to find facets that are likely to make a difference, conceptually or empirically. Then, a first mapping sentence is sketched. This mapping sentence is best tested against new items: They often cannot be reliably classified by the first-draft mapping sentence, and so more conceptual work has to be done on this mapping sentence (sharper definitions, additional facets, better "grammar", etc.). If, after some such iterations, one arrives at a conceptually sufficiently clear mapping sentence, data are collected and the mapping sentence is tested for its empirical usefulness. But the empirical structure of the data may also suggest conceptual structure, as we all know from exploratory data analysis. So the mapping sentence and data related to it are related in some kind of "partnership", i.e. in the basic scientific ping-pong relation of theory and observation.

4. Predictions from facet-designed items and assessing the effects of bad translations

For the Bastide & van den Berghe items, a whole set of predictions can be derived from their mapping sentence. We noted above that one can predict positively intercorrelated items or a regionality of MDS representations that reflects the facets. A more intricate prediction is that the whole set of items forms a system of interrelated cumulative scales, a partial order of Guttman scales (Borg, 1994). These issues are described in detail elsewhere. They are not of particular importance to the topic of this paper, but it should be pointed out that structural hypotheses are an automatic by-product of mapping sentence designs. Hence, mapping sentence designs allow one to check the equivalence of

different-language versions of the items empirically in the usual sense of construct validity.

Several related examples for concrete cross-cultural applications of this approach are the studies by Borg (1986, 1991), Elizur et al. (1991), Borg & Braun (1996). They studied work value items, i.e. items such as 'How important is it to you to make a lot of money in your job?' or 'How important is it to you to have interesting work?'. The mapping sentence for these items distinguished two facets. One facet classified items according to the need that a particular outcome relates to (e.g., in one particular formulation, whether the outcome satisfies an existential-material need, a social-emotional need, or a growth need). The other facet distinguished whether the outcome is performance-dependent or system-dependent. Work value items that were used in surveys conducted in countries such as China, West-Germany, East-Germany, Israel, and the USA were all classified by these facets in the same way. It was found that the data from all countries could be structured equivalently by this facet design. That is, the various items could be statistically discriminated by each facet in turn. Moreover, the pattern of discrimination was the same for each country, i.e., a so-called radex structure in two-dimensional MDS space. No further detail on what exactly this means is needed to see that the data analysis is driven by the content facets, not by a preconceived formal notion such as unidimensionality as in ICC. Indeed, even a multidimensional analysis that concentrates on interpreting dimensions of the items (such as factor analysis) would not have revealed the facets' roles in the data (Borg & Groenen, in press).

Yet, this leads to the question how one should evaluate the situation if no such structural similarities are found. In particular, is it possible to separate effects of bad translation from other effects, such as systematic differences of the samples or non-validity of the design facets in certain cultures? It seems to me that this is not possible. That is, only if structural similarity is given, may one then conclude that the items are equivalent, too. Otherwise, it is a challenging task to disentangle the confounding of effects. An easy solution is to eliminate the problem *by assumption*. This is what is done in traditional

psychometrics: if the underlying true structure of the construct assessed by the items is assumed to be identical, and if the samples are assumed to be homogeneous, then any systematic differences in item statistics that remain after admissible fitting transformations are due to bad translation.

5. Establishing item equivalency independently of the research question?

The equivalency issue is not confined to cross-cultural research. It is equally relevant when one wants to construct parallel tests, for example, or when one considers replications with "similar" items. In a sense, even replicating an empirical investigation with the same items may raise the equivalency issue. In the end, it is not difficult to see that it is a fallacy to believe that one may be able to resolve the issue independent of the research problem, by first establishing instruments with equivalent items before turning to the research question one is really interested in.

The main reason is that in the traditional psychometric approach, items are first selected on the basis of a substantive rule. They are then studied empirically for certain formal properties, in particular for their dimensional structure. Items that do not fit this structure are eliminated or rewritten. However, this also affects, indirectly, the initial rule for constructing or selecting items. One cannot, for example, first pick items because they belong to the domain of attitude items on “race relations in the white middle class of Sao Paulo”, and then decide on statistical grounds that some of them have to be eliminated afterwards. What belongs or does not belong to a universe of items is not a statistical but a definitional issue. The data only reveal the structure of this domain, but cannot affect its content

Making the common blueprint of the items as clear as possible, establishes *one* feature of equivalency, i.e., *design equivalency*. It makes it possible to map items into *design-equivalent* items rather than to translate them literally and trying to preserve irrelevant or

even distractive semantic material. Further advantages of this approach are that it facilitates the identification of item types; it helps modeling the conceptual structure of items; it systematically lays out the universe of items, not just ad-hoc collections of items with a vague notion of communality; it suggests structural laws; and it thus enables one to see common content-driven (not statistically induced) structure in one or several sets of items.

Other approaches to item equivalency may or may not be compatible with or complementary to the facet-theoretical approach. The psychometric method, obviously, does not belong to this set. Both methods are, in a sense, opposites of each other, one starting with content and then proceeding to data and models, one starting with models and fitting content to the models.

An obvious special case of the facet approach is the MTMM approach. 'Method' and 'trait' are just two facets that distinguish among different items. Usually, the MTMM approach is also special in a statistical sense, because researchers who use MTMM these days also use particular (usually linear) statistical methods to analyze the data. Yet, there is no compelling reasons for combining the MTMM approach with such statistical models. In fact, the original work by Campbell & Fiske (1959) looked for certain patterns in the MTMM matrix rather than attempting to fit a particular statistical model. Borg & Groenen (19##) showed, moreover, that the traditional models may be easily replaced with the usual content-driven techniques such a regional MDS, where the regions, of course, relate to 'method' and to 'trait'.

References

Allmendinger, J., Schmidt, P. & Wegener, B. (eds.), *ZUMA-Handbuch Sozialwissenschaftlicher Skalen*. Mannheim und Bonn: ZUMA und IZ.

Bastide, R. & van den Berghe, P. (1957). Stereotypes, norms and interracial behavior. *American Sociological Review*, *22*, 689-694.

Borg, I. (1986). A cross-cultural replication on Elizur's facets of work values. *Multivariate Behavioral Research*, *21*, 401-410.

Borg, I. (1991). Multiple facetizations of work values. *Applied Psychology: An International Review*, *39*, 401-412.

Borg, I. (1994). Evolving notions of facet theory. In I. Borg & P.Ph. Mohler (Hrsg.), *Trends and Perspectives in Empirical Social Research* (178-200). New York: DeGruyter.

Borg, I. & Groenen, P.F.J. (1997). *Modern Multidimensional Scaling*. New York: Springer.

Borg, I. & Groenen, P.F.J. (in press). Regional interpretations in MDS. In M. Greenacre & J. Blasius (eds.). *Visualizing Categorical Data*. New York: Academic Press.

Borg, I. & Shye, S. (1995). *Facet Theory: Form and Content*. Advanced Quantitative Methods in the Social Sciences, Vol. 5. Newbury Park, CA: Sage.

Elizur, D., Borg, I., Hunt, R. & Magyari-Beck, I. (1991). The structure of work values: a cross cultural comparison. *Journal of Organizational Behavior*, *12*, 21-38.

Guttman, L. (1959). A structural theory for intergroup beliefs and action. *American Sociological Review*, *24*, 318-328.

Guttman, L. (1971). Measurement as structural theory. *Psychometrika*, *36*, 329-347.

Hess Medler, S. (1993). Personal communication at the Fourth International Facet Theory Conference. August. Prague, Czech Republic.

Hulin, (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of cross-cultural psychology*, *18*, 115-142.

Hulin, C.L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, *67*, 818-825.

Respondents' Ratings of Expressions from Response Scales: A two-country, two-language investigation on equivalence and translation[1]

PETER PH. MOHLER, TOM W. SMITH AND JANET A. HARKNESS

The paper presents German-American research on expressions from response scales used in cross-national and cross-lingual survey research. Respondents in the United States and Germany were asked to rate expression for the degrees of intensity they were held to express. The scales used were scales of agreement, importance and for/against. The findings of the study raise as many questions as they answer. Translation-based pairings of expressions across English and German work well but not perfectly. Symmetrical response scales often lead to artificial-sounding 'scalespeak' constructions: their effect on scale responses is unknown. Well-matched translation pairings were sometimes differently scored across the populations. Germans and Americans differed in the range of scale points they employed and in the range of vocabulary used to 'explain' expressions. The study is seen as a first step towards understanding cross-national response scale issues.

1. Introduction

Cross-national survey research usually takes translated instruments as their route to 'equivalent instruments' (Acquadro, 1996; Van de Vijver, this volume). A number of authors have discussed issues of equivalence and non-equivalence of translated instruments. Others demonstrate and discuss the fact that translation equivalence is only

[1] This research was supported by a grant from the Humboldt Stiftung Transcoop Programme.

one of the equivalencies to be considered in questionnaires (Van de Vijver, this volume; Hui and Triandis, 1986; Hulin, 1987).

The MINTS project (Research into Methodology of Intercultural Surveys) investigated expressions used in response scales in cross-cultural research. The project is the first step in a research programme aimed at exploring the limits and potential of translation with respect to response scales. One of the questions of interest was whether even 'good' translations of expressions used in response scales means that the expressions matched in translation across languages do indeed capture the 'same' or comparable degrees of differentiation. Another was which, if any, of translations already in use for an English response scale would match up best. A further aim was to compare the ratings respondents assigned terms actually used in scales (see section 5.2) with ratings they assigned to terms not used, but potentially usable in scales. And finally, the project investigated what respondents understood various expressions to mean. This is relevant of itself and, we hope, can be linked to corpora research on the various expressions (lexemes) involved.

The MINTS project investigates expressions frequently used in cross-national survey response scales, specifically, expressions used in English and German ISSP[2] response scales. The most commonly used ISSP scales were taken: *agreement/disagreement*; *for/against*; *important/unimportant* (Davis, 1993). Other expressions which are not used in ISSP response scales but are comparable in the degrees of importance, agreement, etc., they express were also investigated (Smith, 1997).

Survey questions generally consist of a (fairly restricted) number of parts which can include an introduction or pre-code, a question-asking part and, in closed format questions, a response scale and instructions such as *Please tick one box*, etc. In the monocultural context, considerable research (not reviewed here) has appeared on almost

[2] The International Social Survey Programme (ISSP) has conducted annual surveys since 1995. Twenty-nine countries are currently members of the ISSP. The data from the survey are distributed by the Zentralarchiv für empirische Sozialforschung in Cologne, Germany.

every aspect of questionnaire design, for example, on *item wording* (e.g., Hippler et al., 1987; Bradburn and Sudmann, 1991; Converse and Presser, 1994; Sudman et al., 1996; Schwarz, 1996), *introductions to questions* (e.g., Cannell et al., 1979; Schumann and Presser, 1981; Converse and Presser, 1994), *length of questions* (e.g., Payne, 1951; Cannell et al., 1979; Converse and Presser, 1994), *question ordering* (e.g., Schumann and Presser, 1981; Hippler et al., 1987; Converse and Presser, 1994; Wänke and Schwarz, 1997; Sudman et al., 1996), and *response scale designs* (e.g., Schumann and Presser, 1981; Presser and Schumann, 1980); Converse and Presser, 1994; Schwarz, 1996; Krosnick and Fabrigar, 1997) to the *interaction between response scales and items* (e.g., Hippler et al., 1987; Schwarz et al., 1991; Schwarz and Hippler, 1991; Schwarz, 1996).

While questions cover a vast range of topics, and there are numerous, albeit 'standard' formats for constructing question-asking parts, response scales, once chosen, tend to be used time and again in identical format. Davis's (1993) review of circa 300 ISSP questions shows the ISSP agreement scale was used 92 times in modules from 1985-1993, the ISSP importance scale, 23 times, an *allow/not allow* scale, 22 times and an *in favour/against* scale, 11 times. Other research programmes, such as the American GSS[3], also repeatedly use the same scales from year to year. Response scales therefore seemed a most useful starting point for our research programme.

2. Agreement Scales across Institutes and Countries

Many major surveys use *agreement* scales and, as just mentioned, often consistently use one or the other format. Where variation occurs, this is often due to taking over questions from other surveys. Across programmes, however, both within one country and across countries, differences in the formulation of a particular scale are frequent. In different

[3] The American General Social Survey (GSS) is an annual survey conducted by the National Opinion Research Center (NORC) in Chicago. The first survey took place in 1972. Further information at the web site (www.norc.uchicago.edu/gss.htm).

programmes in English, for example, one finds the following variations of an agreement scale:

1) A 'forced choice' response scale with only the first two options read out to respondents.

Agree
Disagree
Don't know
No answer
Not applicable

Source: American General Social Survey (GSS), Cumulated Codebook, Q.357a, 1972-1993.

2) A 'forced choice' design using a four-point scale, with two 'agreement' points, two 'disagreement' points and no middle option:

Strongly agree
Agree a little
Disagree a little
Strongly disagree
DK
NA

Source: British Social Attitudes (BSA), Cumulated Sourcebook. K-15 (1987/1989).

3) Seven- or five-point scales provide mid-points and some differentiation in degrees of agreement and disagreement. In addition, the following British scale has the reverse order of modifier to that of the previous scale (italics added here).

Agree *strongly*
Agree
Neither agree nor disagree
Disagree
Disagree *strongly*
DK
NA

Source: British Social Attitudes (BSA), Cumulated Sourcebook. K-15 (1987/1989).

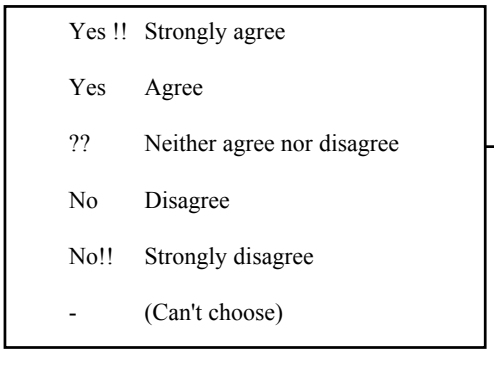
4) The 'standard' ISSP format is as follows (italics added here):

Strongly agree
Agree
Neither agree nor disagree
Disagree
Strongly disagree
Can't choose, Don't know
NA, Refused

Source: ISSP 1993 - GSS (USA) Q 542 A.

5) An Australian version of the standard ISSP scale in example 4) which is used in mail surveys presents the agreement scale and then re-formulates it in terms of *Yes* and *No* and exclamation and question marks, while *Can't choose* seems to become a dash:

To begin with we have some questions about (topic). Do you agree or disagree...(topic)



			Please circle a word			
a. text first item	Yes!!	Yes	??	No	No!!	-
b. text second item	Yes!!	Yes	??	No	No!!	-
c. text third item	Yes!!	Yes	??	No	No!!	-
d. text fourth item	Yes!!	Yes	??	No	No!!	-

Source ISSP 1988 Australia Q 1.

Here both the pre-code wording (*Do you agree or disagree...*) and the scale offered respondents alongside the items differ importantly from the standard ISSP scale.

Cognitive survey methodology research findings show that any one of these differences can affect how respondents react to a scale and the question(s) accompanying it. Numerous findings have demonstrated, for example, that respondents use response scales to interpret questions and questions to interpret scales; that distributions of responses to the 'same' question differ depending on characteristics of the response scales offered; and that the presence or absence of verbal labels or numeric labels, as well as the individual choice of labels, also affect respondents' selection of response options (see Schwarz, 1996, for a review and further references).

Issues of equivalence and the effects of different response scales and scale designs multiply in the cross-national context, in particular when response scales require to be translated. Moreover, the 'close' translation approach often adopted in survey research (Harkness and Schoua-Glusberg, this volume) quickly meets with obstacles in response scale translation. Research on the issues involved is only beginning (Harkness, 1993; 1997; Van de Vijver and Leung, 1997).

3. Measuring the Intensity of Response Categories

The first goal in our research was to establish the *degree* of acceptance, agreement, importance, etc., respondents ascribed to expressions. To do this we needed to measure the degree of intensity respondents assigned to each.

In the monocultural context several approaches have been used to measure the strength of response categories along an underlying response scale. One approach is to have respondents rate the strength of terms defining each point on the scale. There are three standard variants of this approach.

First, one can rank the terms from weaker to stronger or from less to more, or along any similar continuum (cf. Spector, 1976). This, of course, only indicates their relative

position and not the absolute strength or distance between terms. Second, one can rate each term on a numerical scale, usually with 10 to 21 points; (Wildt and Mazis, 1978; Worcester and Burns, 1975; Myers and Warner, 1968; Cliff, 1959; Jones and Thurstone, 1955; Mittelstädt, 1971). This allows the absolute strength or distance between each term to be known and thus facilitates the creation of equal interval scales. It is also possible to use an alphabetical scale or unlabelled spaces, rungs, or boxes, as in a semantic differential scale (Osgood et al., 1957). The letters or spaces are then transformed into their numerical equivalents. Third, magnitude measurement techniques can be used to place each term on a ratio scale (Lodge et al., 1971; 1982; 1992; Wegener, 1991; Hougland et al., 1992). The magnitude measure technique requires that the investigator (sometimes the respondent) give an arbitrary value to a reference term and has respondents rate other terms as ratios to the base term. Typically, respondents have to scale each term by two modes, say, numbers and length of lines. The resulting scales can be calibrated for each individual as well for the whole group of respondents. This allows more precision than the numerical approach, since the terms are not constrained by the artificial limits of the bounded number scale.

Of these three variants, the second seems most useful. On the one hand, the ranking method fails to provide the numerical precision that is necessary to calibrate terms across languages. On the other hand, the magnitude measurement technique is much more difficult to administer and quite difficult for respondents to do, with about 15% of an average population being unable to produce reliable scaling. In addition, the extra precision that a magnitude measurement procedure can provide over that achievable using a 21−point scale approach seems, in our case, to be marginal and thus not needed.

The direct rating approach has been used to rate words along various dimensions. Of most interest here are those that either rate terms along a general good–bad or positive–negative dimension or which rate the intensity of modifiers (Worcester and Burns, 1975; Wildt and Mazis, 1978). Similarly, other studies have rated probability statements (Lichtenstein and Newman, 1967; Wallsten et al., 1986); frequency terms (Simpson,

1944; Spector, 1976; Schaeffer, 1991; O'Muircheartaigh et al., 1993); and terms used in reports to describe percentages from public opinion (Crespi, 1981).

The studies generally show that:

- the tested population (most often American college students) can perform the required rating tasks;
- ratings and rankings are highly similar across different studies and populations (if other than college students);
- there is a high test–retest reliability;
- several different treatments or variations in rating procedures yield comparable results;
- some qualifiers need to be considered differently, as, for instance, vague frequency terms (Schaeffer, 1991; Bradburn and Sudman, 1979).[4]

A second approach for assessing the intensity of scale terms and response qualifiers is to measure the distributions generated by using different response scales (Smith, 1979; Laumann et al., 1994). One version is an *across respondents* design, where two randomly selected groups of respondents get different response scales. With some modelling around what the two observed distributions suggest concerning the supposed underlying distribution, it is possible, within the limits of this approach, to estimate at what point each term cuts the underlying scale (Clogg, 1982; 1984). The assumptions needed for this kind of modeling, namely an underlying 'true' distribution is actually not in line with the more recent literature on judgements and decisions (Schwarz, 1996) or Facet Theory (Borg, 1996; Borg and Groenen, 1997). An alternative version of this approach uses a *within subjects* design. In this, respondents are asked the same question two or more times with different response scales offered (Orren, 1987).

The advantage of the distribution approaches is that they ask respondents to do what they

[4] Experimental settings show systematic differences and artefacts, these seem to vanish or at least to become much smaller in most cases in general population samples and surveys (Weller, 1996).

are normally required to do in the questionnaire context, that is, to answer substantive questions with a standard and typical set of response scales. However, the disadvantages are clear:

– only a very limited number of response scales can be used;

– the statistics need a relatively high number of respondents for each stimulus;

– the implicit model of an underlying 'true' distribution requires detailed analyses.

Since the direct rating approach (asking respondents to rate terms on a 21–point scale) provides the quantified intensity scores needed in the most straightforward manner, this was adopted as the main technique for the MINTS study. At the same time, using a numerical approach in a cross-cultural experiment assumes that respondents in both cultures will respond to and employ numerical values in comparable fashion. While this may be unproblematic for a USA–Germany comparison, in other parts of the world problems are likely, related, for example, to lucky and unlucky numbers, standard (and internalised) rating scales used in education and other spheres, different degrees of familiarity with assessment tasks using more than single digit numbers, etc. These considerations will need to be controlled for in extending our research further.

4. The Study Setting

Experimental pilot studies were carried out in the United States and Germany in 1995 using the direct rating approach described in section 3 to evaluate the equivalence of response scale expressions. The American pilot study was carried out with a sample of adults living in households. Ten sample points were selected to represent all four Census regions (West, South, Midwest, and Northeast). Interviewers had quotas to fill based on gender, age, and employment status. They proceeded through neighbourhoods in the selected communities until the quotas were completed. In contrast to test populations of

college students commonly used in other studies, the respondents of the American pilot study represented the American adult population, according to the stratification variables used for the quota and with respect to marital status and race as a by-product of the selection procedure. Under-represented are, as in many other surveys, the less educated segment of the society. The study was designed and carried out by the National Opinion Research Center at the University of Chicago. Fielding was done in July and August of 1995 with 117 interviews successfully completed (Smith, 1997).

The German experiment was designed as a stand–alone study. By selecting 60 interviewers from different regions, the sample covered all 15 federal states and two main regional substrata, metropolitan regions (100,000 inhabitants and more) and small towns. Within these regional strata, respondents were selected according to a threefold quota table (gender x two age groups x two education groups). The quota cut the population at about the mid-point. As in the American case, the respondents represent the adult population. The sample was split at random to cover two linguistic variants (see below). The study was designed at the German Centre for Survey Research and Methodology (ZUMA); fieldwork was carried out by Infratest-Burke Sozialforschung, Munich. Fieldwork started on September 7 and ended on September 22, 1995. Each interviewer administered only one of the two split-versions; 221 interviews were successfully completed (split 1: 113; split 2: 108).

4.1 Splits

United States: The two American questionnaires differed in question 4 by using *important/unimportant* in one split, and *important/not important* in the other.

Germany: The two German questionnaires differed in all the questions using *agree/disagree* (Q 2, 3, 6, 7, and 8). In split one, *disagree* was translated as *ablehnen*, a verb covering much of the meaning of *disagree/reject*; in split two, *disagree* was translated as ‘not agree’, that is, with *zustimmen* (‘agree’) and a negative particle, *nicht* (‘not’).

4.2 Pairing of English and German Expressions for the Experiment

Selection and pairing of expressions in German to match the English expressions was made on the basis of a) current usage in German surveys, which is itself either based on translations made at some point in time or based on preferred institute or country style, b) translator judgements of appropriateness, and c) formulations which maintained response scale symmetry (Harkness and Mohler, 1997; Harkness, 1993). The experiment was thus able to investigate expressions based on current practice in survey translation and also to expand on this in two relevant directions. All three bases of pairings should be kept in mind when looking at what in some instances might otherwise be surprising alignments.

4.3 Respondents' Ratings of Expressions on a 21-Point Scale

One of the central tasks in the experiment had respondents rate 28 expressions of agreement (26 in English) on a 0 to 20-point scale. Apart from introductory material which contained survey question and answer formats, respondents worked with the expressions outside the survey question-and-answer context. This was important in order to be able at a later stage in research to distinguish between how respondents react to expressions in the questionnaire setting and how they react to these expressions outside of a response scale. Respondents rated each expression in terms of the degree of *agreement/disagreement, importance/unimportance* or 'support for' (in terms of *for/against)* each was felt to express. Theoretically, respondents might be expected to rate *completely agree* somewhere near to 20 and an expression like *completely disagree* near to 0. Respondents were also given the opportunity to adjust their ratings of the agreement expressions once they had completed this task. This revision step was seen as both psychologically useful and informational. It provided some indication of respondent certainty of assessment, gave respondents a chance to look back over the longest and perhaps most demanding task before moving on, and afforded a break in a long sequence of interviewer-respondent dialogue. Respondents did not use this as an opportunity to change ratings to rankings.

4.4 Respondents' Own Definitions of What *Agreement* Means

After the rating part of the experiment, respondents were asked to indicate what they understood the various terms to mean. In English, they were asked the following for agreement: "Now, I'm going to ask you about some of words we've just been discussing. What does the word *agree* meant? What does it involve?" Similar probes were made for *disagree*, *neither agree nor disagree, important*, and *unimportant*. The German respondents were asked as follows: „Im folgenden geht es um einige der Begriffe die Sie gerade eingeordnet haben. Was bedeutet das Wort *stimme zu?* Was heißt das?"

Table 1a below contains the nouns, verbs, adjectives, etc., used by American respondents in their definitions of the meaning of *agree.* Table 1b contains the words used by German respondents in explaining *zustimmen*. Eighty different words were provided by the sixty-one USA respondents taking part in this task. Interviewer records indicate that a fair number of USA respondents used the word asked for as a description of its meaning (e.g., "agree means to agree"). Thirteen of the words used (16,25%) can be seen as variations of the word asked for *(agreement, agreeing),* twenty-five of the words used (31%) can be seen as paraphrases. Of the words used offered by 218 German respondents, 90% of the words chosen can be seen as paraphrases, 6% (fourteen expressions) as repetitions of the word stem of *zustimmen*.

Table 1: 1A - Words used by German Respondents for *zustimmen,* 'agree'

Word	Frequ.	Categ. Frequ.
Akzeptieren	1	
Akzeptabel	1	
Akzeptiere	1	3
Anerkennen	1	
Befürworte	1	
Befürworten	2	
Befürwortung	1	4
Bejahe	1	
Bejahen	3	
Bejahung	4	8
Dafür	42	42
Einverstanden	47	
Einverständnis	7	
Einverständniserklärung	1	55
Gleiche	12	
Gleichen	2	
Gleicher	9	23
Große	1	
Grund	1	
Grunde	3	
Gut	8	8

Word	Frequ.	Categ. Frequ.
Identisch	1	
Positiv	3	
Positive	4	7
Richtig	7	
Richtige	1	
Richtigkeit	1	9
Selbe	4	
Selben	1	
Selber	1	6
Soll	7	
Volle	3	
Volles	1	11
Zustimme	1	
Zustimmen	2	
Zustimmung	11	14
Zutreffend	1	
Zuverlässig	1	
Übereinstimmen	1	
Überzeugt	9	
Übereinstimmung	6	
Überzeugung	3	

1B - Words Used by US-Respondents for *agree*

Word	Frequ.	Categ. Frequ.
About	2	
Accept	3	
Acceptance	2	6
Accomplish	1	
Accord	1	
Accordance	1	2
Admit	1	
Against	1	
Agree	8	
Agreeable	2	
Agreeing	1	
Agreement	2	13
Alike	1	
Approve	3	
Congenial	1	
Consensus	2	

Word	Frequ.	Categ. Frequ.
Consent	2	4
Disagree	1	
Favor	7	
For	5	
Harmony	2	
Like	2	
Liking	1	
Line	2	
Mutual	1	
Ok	1	
Okay	1	
Same	16	
Similar	1	
Support	2	
True	2	
Valid	1	

The readiness of the German respondents to paraphrase or provide alternative expressions and that of Americans to offer the word probed as an explanation of itself can be a reflection of various culturally determined factors (Johnson et al., 1997).

5. Selected Results from the Rating of Agreement Expressions

5.1 *In the Middle* is in the Middle

In the middle and *in der Mitte* both have a mean about the mid-point of the rating scale used. Respondents in both countries not only located expressions such as *neither/nor* and the corresponding German *weder/noch* close to the middle of the scale range, but also placed the so-called 'off-scale' response option of *can't choose* (and *kann ich nicht sagen*

– 'I cannot say') around this middle area, too. Off-scale options are generally understood in survey research as recording the *absenc*e of opinions. It is also sometimes argued that middle categories are used to record non-opinions. Rather than supporting the suggestion that middle options are in fact off-scale options, our findings suggest that middle options, at least in the experimental context, are precisely that. Moreover, expressions implemented in surveys as off-scale options (e.g. *can't choose* and *kann ich nicht sagen)* are in this context close to the centre of the scale, not off-scale (cf. Smith, 1997:13).

Table 2 shows the respective ratings for this middle group of expressions. D stands for the German questionnaire, USA for the American questionnaire, the letters and numbers (e.g., A13 and c in column one) are the respective expression IDS in the two experiments.

Table 2: *In the Middle* and *in der Mitte*

Item IDs D/USA	German Expressions	Mean D	Mean USA	American Expression
A13/c	Stimme ein bißchen zu	12,46	12,10	Agree a little
A26/m	**In der Mitte**	**10,02**	**10,10**	**In the middle**
A22/z	Unentschieden	10,00	9,60	Undecided
A4/p	Stimme weder zu noch lehne ab	9,77	9,90	Neither agree nor disagree
A9/e	Kann ich nicht sagen	9,42	9,80	Can't choose
A7/u	Lehne teilweise ab	6,77	6,60	Somewhat disagree

5.2 'Equivalent' translations do not always have equivalent ratings

Table 3 below shows that *in the middle* and *agree a little*, as well as the German counterparts, *in der Mitte* and *stimme ein bißchen zu,* are rated closer to one another (mean value difference: USA: 2.00 and D: 2.44) than *in the middle/in der Mitte* and the next closest 'disagreement' expression in each language (*disagree a little* (difference

3.00), *lehne teilweise ab* (difference 3.25). Moreover, the distance between *in der Mitte*, *in the middle* to the disagreement expressions which are ‘equivalent’ in terms of word symmetry (*disagree a little* and *lehne ein bißchen ab*) is greater for German (3.05) than for the USA (3.00). The 'structurally equivalent' translation pairing here is not supported by the respondents' ratings. This is suggestive evidence of the dangers of equating linguistic similarity and/or expression symmetry with measurement properties. It may also be related to scalespeak effects, in as much as *disagree a little* is normal English and *lehne ein bißchen ab* is constructed, artificial German.

Table 3: Mean Values of Agree/Disagree Expressions

Item IDs D/US	German Expressions	Mean D	Mean USA	American Expression
A20/v	Stimme voll and ganz zu	19,87	18,80	Strongly agree
A27/f	Stimme völlig zu	19,55	19,40	Completely agree
A17/h	Stimme bestimmt zu	19,22	19,00	Definitely agree
A16/b	Stimme zu	19,05	16,00	Agree
A12/aa	Stimme sehr zu	17,77	18,50	Very much agree
A28/d	Stimme ziemlich zu	16,33	17,20	Agree a lot
A1/a	Stimme im Grunde zu	14,93	13,80	Basically agree
A25/y	Stimme eher zu	13,99	13,50	Tend to agree
A6/r	Stimme wahrscheinlich zu	13,93	13,60	Probably agree
A18/t	Stimme teilweise zu	13,37	12,90	Somewhat agree
A11/n	Stimme mäßig zu	12,49	13,30	Moderately agree
A13/c	Stimme ein bißchen zu	12,46	12,10	Agree a little
A26/m	In der Mitte	10,02	10,10	In the middle
A22/z	Unentschieden	10,00	9,60	Undecided
A4/p	Stimme weder zu noch lehne ab	9,77	9,90	Neither agree nor disagree
A9/e	Kann ich nicht sagen	9,42	9,80	Can't choose
A7/u	Lehne teilweise ab	6,77	6,60	Somewhat disagree
A10/s	Lehne wahrscheinlich ab	6,66	6,20	Probably disagree
A21/o	Lehne mäßig ab	6,63	6,40	Moderately disagree
A24/k	Lehne ein bißchen ab	6,57	7,10	Disagree a little
A19/y	Lehne eher ab	5,82	6,40	Tend to disagree
A15/l	Lehne ziemlich ab	3,91	3,00	Disagree a lot
A14/q	Stimme nicht zu	3,32	3,50	Not agree
A2/i	Lehne bestimmt ab	2,42	1,00	Definitely disagree
A3/j	Lehne ab	2,41	3,50	Disagree
A23/bb	Lehne sehr ab	1,77	1,40	Very much disagree
A5/w	Lehne stark ab	1,21	1,50	Strongly disagree
A8/g	Lehne völlig ab	0,67	0,80	Completely disagree

5.3 Sample Error Variance of Mean Values

In statistical terms, mean values resulting from samples may vary from sample to sample.

Possible variations around a 'true' mean value in the population from which the sample was drawn can be estimated, however. In Figure 1 below, the mean values from our sample are surrounded by vertical lines indicating the band width of stochastically possible variation (variation due to sampling and measurement error – 95% confidence interval). In other words, if the experiment were repeated many times, the expectation is that 95% of the respective mean values would fall within the band width indicated.

Figure 1: Comparison of Means - *agree a little - somewhat disagree* and German counterparts – 95% Confidence Interval

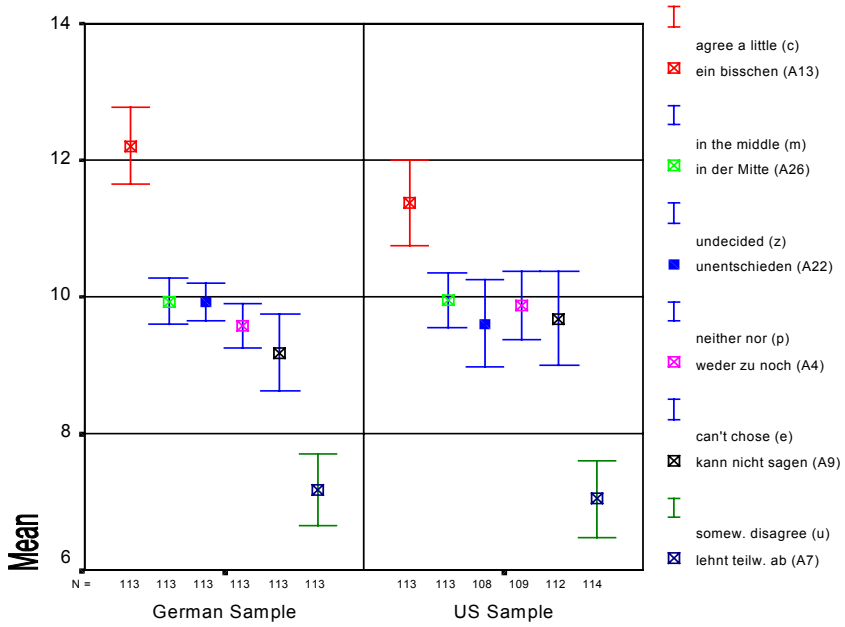


Figure 1 presents these band widths for *agree a little/stimme ein bißchen zu* over *in the middle/in der Mitte* to *somewhat disagree/lehne teilweise ab*. Each vertical bar above and below the boxes indicates the band width of the respective mean value. German results

are plotted on the left, American on the right. A horizontal overlap of the bars indicates that the mean values of the respective expressions are statistically indistinguishable.

The topmost expression here is the agreement pair respondents rated lowest but still above *in the middle/in der Mitte*, that is, *agree a little, stimme ein bißchen zu.* Response pair *lehne teilweise ab/somewhat disagree* is the first *pair* rated below *in the middle*, *undecided*, *neither/nor*, *can't choose* and their German counterparts. We took the German order of mean values here. The first American *expression* in terms of rating is *disagree a little,* as can be seen in Table 3.

The four expressions in each language lexically referring to a mid-point, a non-decision, or an inability to choose, are clustered around the mid-point 10 on the scale. The confidence intervals of the means overlap within countries as well as across, but are distinct from the next 'agreement' and 'disagreement' expressions. In short, the four expressions indicate a mid-point with the same accuracy; they are statistically indistinguishable.

5.4 US and German Differences in Range of Scales

Table 3 findings indicate that the range of scale points American respondents used to rate English expressions is narrower than that used by the German respondents for German expressions. The highest German mean value is 19.87 for *stimme voll und ganz zu*, the American corresponding highest mean value, of 19.40, is for *completely agree*.

On the disagreement ratings, we find a similar pattern. *Lehne völlig ab* is rated as 0.67, while *completely disagree* is located at 0.80. However, inspecting the median values shows this result holds for the top of the scale only (Table 4). This indicates differences across the experiments in dealing with agreement and disagreement which require further investigation.

Table 4: Median Values for *Strongly Agree* and *Strongly Disagree* and *Stimme Voll und Ganz Zu* and *Lehne Völlig Ab*

German expression	Median, Germany	Median, America	American expression
Stimme voll and ganz zu	20,0	19,0	Strongly agree
Stimme zu	18,0	16,5	Agree
Weder zustimmen/noch ablehnen	10,0	10,0	Neither agree nor disagree
Lehne ab	03,0	03,5	Disagree
Lehne stark ab	01,0	01,0	Strongly disagree

6. Summary of Main Findings

The rating experiments showed in general a high correspondence between the *a priori* pairings of expressions by researchers in the United States and Germany. Most means are close and not statistically different from one another (Mohler et al., 1996). Despite this extremely high correspondence, expressed in correlation coefficients above 0.9 (Smith, 1997), there are, nevertheless, some important differences in the mean values. First, the simple base terms such as *agree – stimme zu*, *disagree – lehne ab/stimme nicht zu* are rated more extremely by German respondents than their English counterparts are by American respondents. It remains to be seen whether this means the German expressions involve greater intensity of agreement/disagreement, etc. or whether, independent of this, German respondents differ in rating behaviour. Certainly, in other languages and cultures, response behaviour and the intensity of agreement/disagreement associated with unmodified base terms do seem to differ (Johnson et al., 1997).

Some expressions rank differently across the two countries. Thus in the US experiment, respondents gave the following order to expressions (in the middle = 1):

US Sequence No.1	2	3	4
in the middle	*disagree a little*	*somewhat disagree*	*moderately disagree & tend to disagree*
'German Pair' Sequence No.1	5	2	4

In Germany the expressions paired to the above by researchers were ordered by respondents as follows:

German Sequence No.1	2	3	4
in der Mitte ('in the middle')	*lehne teilweise ab ('disagree/reject in part')*	*lehne wahrscheinlich ab ('probably disagree/ reject')*	*lehne mäßig ab ('moderately disagree')*
'US Pair' Sequence No.1	3	5	4

Differences in ranking in the two populations can be noted for scalespeak pairs such as *disagree a little* and (scalespeak) *lehne ein bißchen ab* but also for expressions which, at face value, are well-paired, ordinary translatory equivalents *(lehne wahrscheinlich ab, probably disagree).*

7. The Next Steps

Assessment of response scales in translation can neither be limited to assessment of translating equivalence (however defined, cf. Harkness and Schoua-Glusberg, this volume; Harkness and Braun, in preparation) nor assessment of measurement properties.

For instance, the effects of scalespeak characteristics across languages have, to our knowledge, never been investigated. If, for example, symmetrical scalespeak designs skew response scales, then other expressions which do not observe scale symmetry might be preferable. Moreover, linguistic corpora could be used to provide researchers with a wider range of expressions to choose from; these could, moreover, be evaluated in their habitual or preferred contexts. In this way, researchers would have concrete evidence of whether, for example, a modifier is usually used with positive or negative headwords or whether headwords are gradable (potentially a part explanation of why *a little bit unimportant* is unusual).

Our findings are based on respondents' reactions to expressions removed from the answer scale context. It remains to be seen to what extent these carry over to a response scale context. On the basis of our findings, for example, the expressions used in the English ISSP agreement scale are not equidistant from one another in the degrees of agreement/disagreement respondents felt they expressed. The same applies to the expressions used in German as standard response scale translations of these. We now plan to test respondents' reactions to standardly used response scales against their reactions to response scales using other expressions which our findings indicate might signal more equidistant intervals.

References

Acquadro, C., Jambon, B., Ellis, D. and Marquis, P. (1996). Language and Translation Issues. In: B. Spilker, (ed.), *Quality of Life and Pharmacoeconomics in Clinical Trials* (2nd edition). Philadelphia: Lippincott-Raven.

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling. Theory and Applications.* New York: Springer.

Bradburn, N.M. and Sudman, S. (1979*). Improving Interview Method and Questionnaire Design.* San Francisco: Jossey-Bass.

Bradburn, N.M. and Sudmann, S. (1991). The Current Status of Questionnaire Design. In: P.P. Biemer, et al. (eds.), *Measurement Errors in Surveys* (pp. 29-40). New York: John Wiley & Sons.

Cannel, Ch.F., Oksenberg, L. and Converse, J.M. (1979). *Field Experiments in Health Reporting 1971-1977.* ISR Research Report Series. Ann Arbor: ISR.

Cliff, N. (1959). Adverbs as Multipliers. *Psychological Review* 66: 27−44.

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

Clogg, C.C. (1984). Some Statistical Models for Analyzing Why Surveys Disagree. In: Ch.F. Turner and E. Martin (eds.), *Surveying Subjective Phenomena*. Vol. 2. New York: Russel Sage.

Converse, J.M. and Presser, S. (1994). Survey Questions: Handcrafting the Standardized Questionnaire. In: M.S. Lewis-Beck (ed.), *Research Practice* (pp. 89-162). London: Sage/Toppan.

Crespi, L.P. (1981). *Semantic Guidelines to Better Survey Reportage.* Office of Research, International Communication Agency, Memorandum.

Davis, J.A. (1993). Memorandum to the ISSP, Chicago: NORC (mimeo).

Harkness, J.A. (1996a). Mountains and Molehills - Equivalence in Cross-Cultural surveys: the Case of Response Scales. (Based on a paper first presented at the American Association of Public Opinion Research, St Charles, 1993).

Harkness, J.A. (1996b). The Representation of Selves in Everyday Questionnaires. Paper presented at the 4th International Sociological Association Conference on Survey Methodology, Colchester, England.

Harkness, J.A. and Mohler, P.Ph.(1997). Towards a Manual of European Background Variables. ZUMA Report on Background Variables in a Comparative Perspective. ZUMA: Mannheim (mimeo).

Hippler, H.-J., Schwarz, N. and Sudman, S. (1987). *Social Information Processing and Survey Methodology.* Heidelberg: Springer.

Hougland, J.G., Johnson, T.P. and Wolf, J.G. (1992). A Fairly Common Ambiguity: Comparing Rating and Approval Measures of Public Opinion. *Sociological Focus 25*: 257-271.

Hui, C.H. and Triandis, H.C. (1985). Measurement in Cross-Cultural Psychology. A Review and Comparison of Strategies. *Journal of Cross-Cultural Psychology 16(2)*: 131-152.

Hulin, C.L., Drasgow, F. and Komocar, J. (1982). Applications of Item Response Theory to Analysis of Attitude Scale Translations. *Journal of Applied Psychology 67(6)*: 818-825.

Hulin, C.L. (1987). A Psychometric Theory of Evaluations of Item and Scale Translations: Fidelity Across Languages. *Journal of Cross-Cultural Psychology 18(2)*: 115-142.

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

Johnson, T., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R., Lacey, L. and Horn, J. (1997). Social Cognition and Responses to Survey Questions among Culturally Diverse Populations. In: L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewim (eds.), *Survey Measurement and Process Quality* (pp. 87-113). New York: John Wiley & Sons.

Krosnick, J.A. and Fabrigar, L.A. (1997). Designing Rating Scales for Effective Measurement in Surveys. In: L. Lyberg, et al. (eds.), *Survey Measurement and Process Quality* (pp. 141-164). New York: John Wiley & Sons.

Laumann, E.O., Gagnon, J.H., Michael, R.T. and Michaels, S. (1994). *The Social Organization of Sexuality: Sexual Practices in the United States.* Chicago: University of Chicago Press.

Lichtenstein, S. and Newman, J.R. (1967). Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities, *Psychon. Sci. 9*: 563-564.

Mohler, P.Ph., Harkness, J.A., Smith, T.W. and Davis, J.A. (1996). Calibrating Response Scales Across Two Languages and Cultures. ZUMA: Mannheim (mimeo).

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

Myers, J.H. and Warner, W.G. (1968). Semantic Properties of Selected Evaluation Adjectives. *Journal of Marketing Research 5*: 409-412.

Payne, S.L. (1951). *The Art of Asking Questions.* Princeton/NJ: Princeton University Press.

O'Muircheartaigh, C.A., Gaskell, G.D. and Wright, D.B. (1993). The Impact of Intensifiers. *Public Opinion Quarterly 57*: 552-565.

Orren, G.R. (1978). Presidential Popularity Ratings: Another View. *Public Opinion 1*: 35.

Osgood, Ch.E., Suci, G.J. and Tannenbaum, P.H. (1957). *The Measurement of Meaning.* Urbana, IL: University of Illinois Press.

Presser, S. and Schumann, H. (1980)

Schaeffer, N.C. (1991). Hardly Ever or Constantly? Group Comparisons Using Vague Quantifiers. *Public Opinion Quarterly 55*: 395-423.

Schönemann, P.H. (1994). Measurement: The Reasonable Ineffectiveness of Mathematics in the Social Sciences. In: I. Borg and P.Ph. Mohler (eds.), *Trends and Perspectives in Empirical Social Research*. (pp. 149-160). Berlin: de Gruyter.

Schumann, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context.* New York: Academic Press.

Schwarz, N. (1996). *Cognition and Communication.* Mahawah/NJ: Lawrence Erlbaum.

Schwarz, N. and Hippler, H.-J. (1991). Response Alternatives: The Impact of their Choice and Presentation Order. In: P.P. Biemer et al. (eds.), Measurement Errors in Surveys. (pp. 41-56). New York: John Wiley & Sons.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E. and Clark, L. (1991). Rating Scales: Numeric values may change the meanings of scale labels. *Public Opinion Quarterly 55*: 570-582.

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

Smith, T.W. (1979). Happiness: Time trends, seasonal variations, intersurvey differences, and other mysteries. *Social Psychology Quarterly 42*: 18-30.

Smith, T.W. (1997). Improving Cross-National Survey Research by Measuring the Intensity of Response Categories. GSS Cross-National Report No. 17. Chicago: NORC (mimeo).

Spector, Paul E., (1976). Choosing Response Categories for Summated Rating Scales. *Journal of Applied Psychology 61:* 374−375.

Stone, L. and Campbell, J. (1984). The Use and Misuse of Surveys in International Development: An Experiment from Nepal. *Human Organization 43*: 30-37.

Sudman, S., Bradburn, N.M. and Schwarz, N. (1996). Thinking About Answers – *The Application of Cognitive Processes to Survey Methodology.* San Francisco: Jossey-Bass.

Van de Vijver, F.J.R. and Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research.* Newbury Park/CA: Sage.

Wänke, M. and Schwarz, N. (1997). Reducing Question Order Effects: The Operation of Buffer Items. In: L. Lyberg et al. (eds.), *Survey Measurement and Process Quality* (pp. 115-140). New York: John Wiley & Sons.

Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R. and Forsyth, B. (1986). Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology 115*, 348–365.

Wegener, B. (ed.) (1991). *Social Attitudes and Psychophysical Measurement.* Hillsdale/NJ: Lawrence Erlbaum.

Weller, I. (1996). Kontexteffekte in Eurobarometer Umfragen - Theoretische Implikationen und praktische Bedeutung. Unpublished dissertation University of Heidelberg.

Wildt, A.R. and Mazis, M.B. (1978). Determinants of Scale Response: Label vs. Position. *Journal of Marketing Research 15*: 261–267.

Worcester, R.M. and Burns, T.R. (1975). A Statistical Examination of the Relative Precision of Verbal Scales. *Journal of the Market Research Society 17*: 181–197.