Warning Concerning Copyright Restrictions

The Copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyright material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction not be "used for any purposes other than private study, scholarship, or research." If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use," that user may be liable for copyright infringement.
5

THE WORDING AND TRANSLATION OF RESEARCH INSTRUMENTS

RICHARD W. BRISLIN

The purpose of this chapter is to provide very specific guidelines which will allow researchers:

1. to select or to develop measuring devices which will be readily translatable into other languages;
2. to write items for measuring instruments which will be easily understood by respondents;
3. to insure that the instruments measure concepts which are important in the cultures under study;
4. to insure that there is no imposition of concepts from the researcher's own culture; and
5. to supervise adequate translations of research instruments.

In addition, a few suggestions will be given concerning the administration of research instruments to actual respondents.

Many researchers find that their well-laid research plans must be modified after arrival in another culture. Some opportunities for data collection disappear and others arise; topics originally considered important seem less so upon first-hand examination; political changes make investigations of some topics unacceptable; pre-test data show that changes must be made; and researchers develop new interests given the stimulation of life in another culture. In many cases, then, researchers must start anew but without the benefit of their libraries and technological back-up, which are most often left home. The guidelines to be presented here place few demands on hardware or on extensive library resources, and instead place demands on the researcher's most valuable commodity: careful thought.
APPROACHES TO INSTRUMENTATION

In any research involving the use of questionnaires, interview schedules, or tests, a major decision must be made concerning the relative usefulness of existing instruments, newly designed instruments, or a combination. "Existing instruments" will be a shorthand term to designate measures which were developed and standardized in one culture and which can possibly be used for data gathering in another culture. The advantages of using existing instruments are considerable. There are usually a number of published studies, or at least unpublished but available data, which can be compared with newly acquired data. Such comparisons allow a literature to be built up around a commonly shared set of concepts and operational definitions, an advantage admittedly more difficult to achieve if different researchers use different instruments. Time and expense can often be conserved if existing instruments are used. There is always limited time, energy, and funding for cross-cultural research projects, especially for members of professions who do not yet have a tradition of taking long periods of time away from their home base to do field research in other cultures. Cost/benefit considerations in both economic and scientific matters should become part of the researcher's planning.

Another advantage will be felt more by neophyte than experienced researchers. There is a sense of security when using existing instruments, a sense that if some established researcher has used a certain measure and obtained respectable publications, then it must have merit. In the case of very frequently used instruments such as the Minnesota Multiphasic Personality Inventory (MMPI) or the Internal-External Locus of Control scale, there is often the sense that these edifices cannot possibly be the target of criticism or improvement. Further, to modify the instrument would be like tampering with some sacred writing, rather like trying to improve the King James version of the Bible through additions, deletions, and modifications. A general caution, then, is that one should not be lulled into a false sense of security through choice of a popular existing device.

There are disadvantages, as well, to the use of existing instruments. Researchers run the risk of missing aspects of a phenomenon as viewed by (and seen as important by) people in other cultures. Further, they risk imposing conclusions based on concepts which exist in their own cultures but which are foreign, or at least partially incorrect, when used in another culture. Put another way, existing instruments provide operational definitions of certain concepts. There is no automatic guarantee that those concepts, or those same operational definitions, exist in other cultures.

An important point to keep in mind is that items on existing scales were tried out with people from one culture, most often in a highly industrialized nation such as the United States. Items which "did not work," as shown by low item-to-total correlations, would be discarded. Consequently, items were purified so that they measure a phenomenon as seen by people in the culture-of-standardization. There is no reason to suppose that these purified items will also measure the phenomenon as experienced by people in another culture. The clearest example involves tests of intelligence, and no thoughtful researcher gives existing I.Q. tests to people from other cultures and makes conclusions about mental capacities (see Chapter 7 by Irvine, this volume). Questions from such tests, asking for an estimate of the population of the United States or the meaning of the word "esoteric" are obviously culture bound. The fact that psychology's history does include cross-cultural use of I.Q. tests to rank cultures on intelligence is a black mark for our discipline. But similar thinking about the contribution of individual test items to a conclusion about psychological processes should be done for any existing tests before its cross-cultural application. Users of the MMPI for cross-cultural studies, for example, should be able to answer a query as to why the item asking about a preference between Presidents Washington and Lincoln should contribute to conclusions about personality traits or syndromes.

Another disadvantage is that the intellectual timidity of researchers can be reinforced if existing tests only are used. Rather than trying to make an original contribution, with its inherent discomfort because of uncertainty regarding the acceptance or rejection of one's work, some researchers busy themselves by using existing tests, reaping the benefits of security already discussed. The exact tests, then, are used not to discover important conclusions but rather to insure that something happens during one's research stint in another culture. The unimaginative use of tests should be discouraged.

My recommendation is that existing tests be used only after very careful thought has been given to the meaning of any scores that the test may yield. For example, intelligence tests cannot be used to
make conclusions about underlying competencies in other cultures. But if people from other cultures are attending school administered by members of the majority culture for which an I.Q. test was designed, then the test might give valuable information to teachers regarding skills (vocabulary development, reading comprehension) which need more attention. There is a large difference between conclusions about mental capabilities and conclusions about skills which need improvement in a particular school system. As another example, consider use of the MMPI. If a researcher could make a good case for the possible universality of schizophrenic etymology and expression (as discussed by Draguns, 1980, who also reviewed the evidence for culture-specific manifestations), then cross-cultural use of those test items related to schizophrenia might be used in diagnosis. Another caveat for such test use is that there is little chance for discovering new insights into the nature of intelligence or schizophrenia.

Another recommendation is that users of existing tests be willing to modify items and to add new ones which should tap additional aspects of a phenomenon in addition to those indicated by the original test. In this procedure, researchers use the emic-etic distinction to their benefit (Berry, 1969a; Brislin, 1983). Take the example of a two-cultural comparison. The etic refers to a phenomenon, or aspects of a phenomenon, which have a common meaning across the cultures under investigation. This can be called the core meaning. For intelligence, such a core meaning might be the ability to give correct answers to problems, the exact form of which has not been seen before by test takers. Eemic aspects are different in the two cultures, but each emic is related to the shared etic core. In some African societies, for instance, intelligence includes the ability to get along with kin (Serpell, 1976). In the United States, intelligence includes quickness in problem solving, as shown by the frequent use of timed tests. Eemics are not necessarily shared: the United States does not share the concern with kin, and there is not a value placed on quickness in the African societies studied by Serpell (1976). A description of the phenomenon for each culture, then, would consist of the etic core plus the culture's emic aspects. This is a good model for actual research projects and the creation of measuring instruments, and so an example will be reviewed in some detail.

Miller, Slomczynski, and Schoenberg (1981) were interested in the measurement and meaning of authoritarian-conservatism in the United States and Poland. The researchers had the goal of assessing shared meanings of authoritarian-conservatism, for etic aspects, as well as culture-specific aspects necessary for a fuller understanding of the concept. Items from the best known test of authoritarianism, the California F scale, were used, with the addition of other items (based on the ideas of Pearlin, 1962) concerned with submission to authority. The authors felt that these latter items would be useful in more fully documenting the meaning of the concept in the two cultures. The study thus had the benefits of possible comparisons with other data sets through use of the F scale, as well as the opportunity to add new insights through the use of additional items. The researchers also modified the classic F scale items when there were “colloquial expressions” or where there was need for “substitution for situations that do not arise in Poland” (Miller et al., 1981, p. 180).

Factor analysis demonstrated that there was a shared core meaning as well as country-specific aspects of authoritarian-conservatism. The core meaning was measured by the following five items:

1. The most important thing to teach children is absolute obedience to their parents.
2. In this complicated world, the only way to know what to do is to rely on leaders and experts. (This was the wording for the US data collection. The phrase “experts and advisors” was a better wording for the Polish data collections.)
3. It's wrong to do things differently from the way our forefathers did. (This was the US wording. The Polish wording was, “It's wrong to do anything in a different way from past generations.”)
4. Any good leader should be strict with people under him in order to gain their respect.
5. A decent man can respect a woman who has had sex relations before marriage.

These items seem to indicate that the core meaning includes deference to respected others (parents, forefathers, leaders, and experts), strict treatment of subordinates, and conventional sexual attitudes. Another advantage of this approach to data collection, analysis, and presentation is that topics for further investigation can be readily identified. Strict treatment of subordinates and conventional sexual attitudes are each measured by only one item, making an in-depth analysis of these aspects of authoritarian-conservatism impossible. Further research can investigate these aspects through the development of multi-item scales.
The country-specific meaning for the United States was indicated by four items:
1. Prison is too good for sex criminals; they should be publicly whipped or worse.
2. Young people should not be allowed to read books that are likely to just end up causing trouble.
3. There are two kinds of people in the world; the weak and the strong.
4. People who question the old and accepted ways of doing things usually just end up causing trouble.

These items measure two aspects: "the use of stereotypes and the endorsement of public intervention in matters ordinarily addressed elsewhere (e.g., in the family or courts)" (Miller et al., 1981, p. 187).

The country-specific meaning for Poland was indicated by three items:
1. One should always show respect to those in authority.
2. You should obey your superiors whether or not you think they're right.
3. Every legal action is proper or are some actions wrong even if legal? (Respondents chose one of the two positions implied in the question. This item is the English translation from the Polish measuring device, and the item undoubtedly reads more smoothly in the original language.)

These items add detail to the core meaning (above) of deference to respected others. The addition is that the deference is also given to "hierarchical authority which is bureaucratically or legally legitimized" (p. 186).

It is no accident that the labels for the country-specific items are long and convoluted. By definition, since the *emic* aspects are not shared, there will not be a readily available, familiar language with which to describe another country's *emics*. The task for the researcher is much like that of introducing a visitor to one's country to unfamiliar customs. Long explanations are demanded since the visitor has little in his or her background which can be related to the newly encountered custom. Similarly, describing another country's *emics* demands introducing unfamiliar ideas, and this cannot be done with the short, precise, sometimes pithy labels researchers use to communicate the results of factor analysis to like-minded peers.

The work of Miller and her colleagues, then, provides a good example of the approach to cross-cultural questionnaires advocated here. To summarize the steps taken, they used an existing instrument so as to allow comparisons with previous studies, but were willing to modify items when necessary. They added new items to expand the scope of inquiry into the nature of authoritarian-conservatism, and they presented their data in terms of shared meaning and country-specific meaning. Countries can then be compared on the items contributing to the shared or *etie* meaning after the caveats suggested by cross-cultural psychometricians (e.g. Irvine & Carroll, 1980; Poortinga, 1975; this volume) are satisfied. In any interpretation of such comparisons, however, researchers must be mindful that a full understanding of a concept within a culture demands the addition of *emic* results to the shared or *etie* meaning. Various researchers have given practical advice regarding two of these stages, the modification of existing items and the creation of new items, and these "lessons from experience" will now be reviewed.

### GUIDELINES USEFUL BOTH FOR WRITING NEW ITEMS AND MODIFYING EXISTING ITEMS

A set of guidelines useful for any sort of item preparation, whether for new items or modification of existing items, has been presented in the cross-cultural literature as "guidelines for writing material which is readily translatable" (Brislin, Lonner, & Thorndike, 1973; Brislin, 1980). The purpose of these guidelines, based on experiences preparing research instruments in over twenty languages, is to assure that translators will:
1. have a clear understanding of the original language item;
2. have a high probability of finding a readily available target language equivalent so that they do not have to use convoluted or unfamiliar terms;
3. be able to produce target language items readily understandable by the eventual set of respondents who are part of the data-gathering stage of the research project.

The guidelines were written with "translatable English" in mind, but they should be of more widespread use since they emphasize clear communication of concepts. While written to help with the methodological step of producing good translations, then, these guidelines should also help in the creation or modification of items which will be readily understandable to respondents. Another benefit is that the items finally produced will be understandable to researchers in other cultures. An important point to keep in mind constantly is that items on a questionnaire or survey instrument constitute the operational definitions of concepts, and clear communication of concepts to
colleagues is essential for the growth of knowledge. As more and more research is done by members of cultures who previously were only the hosts for visiting research teams, communication among cross-cultural researchers will demand more and more attention. In addition to these general points, there are reasons for specific knowledge about translatable English. English will most likely remain the most frequently used international language of communication in the behavioral and social sciences. The same English easily used by translators may also be helpful for researchers attempting to communicate their findings to widely dispersed colleagues.

This presentation is not being presented as a plea for greater use of English in cross-cultural research. The international use of English (Smith, 1981) is a fact, and some readers will bemoan this state of affairs. But given that communication with a broad, world-wide audience is dependent upon publication in English, knowledge of guidelines for writing readily understandable English may improve cross-cultural communication of research results.

In writing or modifying items, the suggested guidelines and rationales are:

1. Use short simple sentences of less than sixteen words. Use of this guideline leads to items with one dominant idea per sentence. With sentences longer than sixteen words, ideas become difficult to disentangle, with lack of clarity regarding which subordinate clause refers to which idea. This guideline does not have to lead to short items. Items can be composed of more than one sentence, each sentence following the suggested sixteen-word limit.

2. Employ the active rather than the passive voice. With the active voice, the translator can more easily identify the subject, verb, object, and can match adjectives and adverbs to the appropriate nouns and verbs. Even though there are languages where a corresponding form of the passive is frequently used (e.g. Tagalog), the active voice is still clearer for translators reading the original English. They may use the passive in their translation, with the comment that what is clearest through use of the active voice in English is communicated best through use of the passive in the target language. Such a difference becomes, then, just one example of the sort of modification necessary for good instrument preparation in other cultures.

3. Repeat nouns instead of using pronouns. With more than one noun per sentence, use of pronouns leads to the risk of unclear references due to vague noun-pronoun links. In many languages, there are far more pronouns than in English, leading to chances for mistakes (e.g. does “you” refer to one, two, or more people? There are different forms for the three meanings in some languages.)

4. Avoid metaphors and colloquialisms. Such phrases are least likely to have equivalents in the target language. Standardized tests from one country are likely to have many colloquialisms since such terms are very good at communicating within a community sharing the same language (e.g. a stitch in time saves nine). Such items survive item purification, as previously discussed. But such terms are very difficult if not impossible to translate for cross-language comparisons. Examples and suggested solutions will be given in a forthcoming section in which cross-cultural use of the MMPI, a coloquialism filled test, is discussed.

5. Avoid the subjunctive, for example verb forms with “could,” “would,” “should.” The rationale here is very pragmatic: other languages rarely have readily available terms for the various forms of the English subjunctive. Researchers who use the subjunctive force the translator to make the best guess or the best approximation. Assuring clear communication is the researcher’s job, and it should not be carelessly delegated to translators.

6. Add sentences to provide context for key ideas. Reword key phrases to provide redundancy. This rule suggests that longer items and questions be used in single-country research. There is no contradiction with guideline number 1, above, on sentence length since items can have several sentences. At times, the context is provided to a translator who does not necessarily have to provide target language equivalents for every word, but should have the context as background. For instance, an item asking about “avoiding stepping on cracks in the sidewalk” would talk about the origin of this superstition. For countries in which sidewalks are continuous flows of concrete and are not interrupted by cracks, the context would include the reason why this item is on the original test. This information maximizes the probability that a cultural equivalent of avoiding sidewalk cracks can be found. Other contextual information will be demanded by knowledge of the published literature and by pre-test data. Questions asking about whether a gentleman should give up his seat on a bus to a lady might have to be contextualized in Asian countries according to the relative age of the people involved, social
class as shown by dress, degree of acquaintance, and so forth. The fact that most Americans can readily provide a straightforward answer to the inquiry about giving up their seat but that Asians demand additional contextualizing information should not lead to researcher frustration. Rather, such information is often important data, and it provides eclectic coloring to cross-cultural comparisons, as previously discussed. The reason for redundancy is to help translators catch mistakes. If they are unsure of the meaning of an item from one phrase, they may find it in the redundant information found in another phrase.' Such items may be marked “wordy” and “repetitive” by high school English teachers, but these people are not the target audience for cross-cultural research instruments.

7. Avoid adverbs and prepositions telling “where” or “when” (e.g. frequently, beyond, upper). Again, the reason for this guideline is pragmatic: there are often inadequate direct equivalents of these words, and so the meaning of an entire sentence can be changed. For instance, how often does an event have to occur for the word “frequently” to be used? Such a term demands knowledge of base rates and subjective reactions to deviations from the base rates, and this information is not always available to researchers. In some cases, cross-cultural analyses have been done of topics which are reflected in adverbs. In analyzing cultural differences in perceptions of time, for instance, Hall (1959) found that concepts summarized by words like “punctually” or “promptly” have very different meanings across cultures. A method for analyzing the connotations of potentially troublesome words and phrases encountered in translation will be reviewed in a subsequent section of this chapter.

8. Avoid possessive forms where possible. Use of possessive forms in English is relatively easy for native speakers since they have a basic understanding of concepts concerning ownership. But to assume that this knowledge will be shared by native speakers of a different language is an example of cultural imposition. The term “his land” may have seemingly adequate linguistic equivalents in another language, but the connotations may be quite different. For instance, the land may be tied to a family or clan, or may be considered “owned” only in the sense that a family is using it at the present time. Another reason for care regarding possessives is that, especially in long sentences, translators may find difficulty in matching exactly what is “possessed” with who is “doing the possessing.” Finally, another difficulty is similar to that presented in guideline number three regarding pronouns. What is presented in one possessive form in English (yours) is presented in one of three forms in many other languages (corresponding to single, dual, and multiple referents).

9. Use specific rather than general terms (e.g. the specific animal such as cows, chickens, or pigs rather than the general term “livestock”). One of the conclusions from the literature on cross-cultural studies of cognition is that people in various language communities do not categorize specific items in the same manner (e.g. Cole & Scribner, 1974). Rather, the items are grouped in different ways across cultures, and consequently any generalized label (when there is one) is likely to refer to collections of items which are not exactly equivalent. Using the above example, for instance, what an English speaker means by “livestock” is unlikely to contain the same specific animals as compared to someone from another language community. Further, people in that other community may not have a general term. For instance, an item asking about good or bad interpersonal relationships with one’s brother, while reasonable for speakers of English, would not be appropriate for research in Japan. There, speakers make a distinction between “older brother” and “younger brother,” and there is no general term to incorporate both.

A careful literature review prior to research, or good pre-testing, would bring out this fact prior to actual data collection. Upon discovery of this fact, researchers should then ask English-speaking respondents about interpersonal relationships with older and younger brothers. This is an example of decentering (Werner & Campbell, 1970; Brislin, Lonner, & Thorndike, 1973). No one country is the “center” of the research project and consequently no one country is the source of instrument development. Rather, data from both countries are used to design instruments which are sensitive to cultural phenomena. If English speakers do not make major distinctions between relations with older and younger brothers, while Japanese do, this becomes an important finding which can then be related to other facts established in the actual data-collection efforts. But unless the distinction related to age were made, the finding will not be forthcoming from the data. To quote from a wise senior professor during my graduate student era, “If you don’t ask, you won’t get.” Another reason for the recommendation regarding specificity of items is related to progress in attitude measurement. Ajzen and Fishbein (1977, 1981) have concluded that behavior is best
predicted, and consequently best understood, when specific rather than general questions are asked about behavior and behavioral intentions. Behavior is very specific: people act in a certain manner, in a certain situation, with other specific actors present or not, and so forth. Using the example of domestic animals again, people do not behave in ways best predicted by responses to inquiries about livestock. Rather, since they behave differently regarding treatment of chickens, cows, sheep, and pigs, behavior is best predicted and understood by inquiries about those specific animals. Another benefit of this concern with specific questions is that researchers have to do great amounts of reading, pre-testing, and listening to people before they are able to formulate specific questions. Such efforts prior to instrument development will surely benefit the growth of cross-cultural research.

10. Avoid words indicating vagueness regarding some event or thing (e.g. probably, maybe, perhaps). This guideline is another reminder of the benefits stemming from specificity in items, discussed above. Even when there are seemingly equivalent terms to words like “probably” or “maybe”, the number of times an event has to occur to be labeled “probable” may differ from culture to culture. When doing survey research within one’s own country, interviewers can call upon large amounts of shared information and shared experiences with respondents. When a respondent demonstrates a blank stare as a response to a vague inquiry, an interviewer can sometimes break the ice by the phrase, “Oh, you know what I mean!” But his phrase cannot be used when interviewer and respondent are from different cultures. Even if indigenous interviewers are used, often a wise practice, the appeal to shared meaning cannot be used if the interviewer is working from an instrument prepared by the outsider-researcher. The best approach is to specify the referent for the subjective experience of concepts like “probably” through such inquiries as the number of times per hour, day, month, or year, as appropriate.

11. Use wording familiar to the translators. There are a number of benefits to keeping this guideline in mind. One is that if wording is familiar to translators such that they can create a well-worded target language version, then that version will most likely be readily understandable to the eventual set of respondents in the data collection effort. Another benefit is that translators are treated more like colleagues than hired-help. Researchers should sit down with translators and go over the materials to be translated, line by line. When translators point out that such and such a term has no direct equivalent, then the researcher should consider this information as good data. The researcher might then change the original language wording, keeping in mind the philosophy of decentering, previously discussed. Reasons for lack of translation equivalence, such as no single word for “brother” in Japan, can point to a way to fruitful hypotheses. If translators have a difficult time with a phrase or a sentence, the researcher can say, “Well, here is what we are trying to get at.” With the underlying purpose of an item in mind, the researcher and translator together can sometimes create a culturally equivalent item.

12. Avoid sentences with two different verbs if the verbs suggest two different actions. If they encounter two verbs in one sentence, translators sometimes have a difficult time attaching the relevant subject to the appropriate verb. The difficulties of translation are extensive enough without burdening research colleagues with problems which can be prevented through good preparation of materials. Another reason for the avoidance of dual verbs in a sentence is that interpretation of respondents’ answers is difficult. With dual verbs indicating dual actions, researchers will have difficulty ascertaining why respondents endorsed or rejected a given item. The reason could involve the thought indicated by one of the verbs, the other, or a combination. For instance, here is an item from the California Ethnocentrism (E) Scale which is a negative model for materials prepared according to these guidelines. “If there are enough Negroes who want to attend dances at a local dance hall featuring a colored band, a good way to arrange this would be to have one all-Negro night, and then the whites could dance in peace the rest of the time” (used as an example by Ashmore, 1970).

Translators will have a difficult time ascertaining who wants to dance, who is doing the arranging, and who will finally dance in peace. Other violations of the guidelines are that the sentence is far too long, and a few of the phrases are colloquial (“all-Negro night,” “dance in peace”). One of the pieces of incidental learning from cross-cultural research is to discover how many colloquial phrases native speakers of a language use on an everyday basis. Even if there happened to be an adequate translation, the researchers eventually will have a difficult task pinpointing the reasons for item acceptance or rejection. The reason could be the concept of an all-Negro night,
the fact that whites can dance without Negroes present, or others. The wisest practice is to have one concept per item so that the respondents' feeling about the concept can be unambiguously measured.

I have found that a good exercise for classes in cross-cultural research methods is to present students with items like that from the California E scale. They are then given the task of rewriting the items using the twelve guidelines. The independent efforts of the students can then be compared and discussed.

**Guidelines Useful for Item Modification**

To obtain good translations, and thus good terms for data gathering, modifications of existing instruments often have to be made. Butcher (1982; Butcher & Garcia, 1978; Butcher & Clark, 1979) has prepared a very extensive and sophisticated treatment concerning the use of existing instruments for cross-cultural use. He has been actively involved in cross-cultural applications of the MMPI, and many of his examples are drawn from this work. There is no attempt to endorse the MMPI in this treatment (see also Chapter 8 by Guthrie & Lonner, this volume). The fact is that a great deal of sophisticated thinking has been invested in how best to modify the MMPI for cross-cultural use. The resulting suggestions for cross-cultural instrumentation are treated here, not the rationale behind the original development of the MMPI.

The most generally helpful guideline for item modification was formulated by Gough (in Brislin, Lonner, & Thorndike, 1973, p. 26) who drew upon experiences with translators of the test he developed, the California Psychological Inventory (CPI).

Most personality assessment material has some sort of diagnostic rationale which may or may not be apparent from content, and which can be quite different from what content suggests. In translation it is this intent that must be maintained, not the content. Thus, translators must know the infrapsychology of the tools they are converting and they must know the empirical connotations of an item as well as its linguistic and literal referents. The CPI abounds in items of this kind, that is those that need intuitive rather than linguistic conversion. Of course, the person who instructs the translator about the “infrapsychology of the tools” and “empirical connotations” is the researcher in charge of the project.

Gough used the example of a CPI item: “Every family owes it to the city to keep its lawn mowed in summer and sidewalks shoveled in winter.” The problem for translation is that people in many countries do not experience harsh winters and consequently the item has no personal relevance. This item is from the socialization scale of the CPI, designed to measure people’s integration into their society, their concern with being good citizens, and their acceptance of widespread social norms which contribute to the smooth functioning of everyday life. To design a cultural equivalent, activities are needed which mark the concerned, well-socialized citizen. In Mediterranean countries, the following is a suggested functional equivalent: “The good citizen does not throw his garbage down the stairwell.”

Butcher and Garcia (1978, p. 474) use an example from the MMPI which has been referred to earlier: “I think Lincoln was greater than Washington.” They argue that “It is necessary to substitute two historical figures who were influential in the historical and cultural development of the target country. The subtle differences between the images of these two historical figures (i.e. one a humanist and the other a militarist or liberator) needs to be maintained.”

A more difficult example is provided by Butcher and Clark (1979): “I like poetry.” This item is designed to reflect feminine versus masculine interests, but in Japan both men and women enjoy poetry. In addition, there is no general word for poetry: there has to be a distinction made between Japanese poetry and Western-style poetry. The word for the latter was chosen for the translation since Japanese women are thought to enjoy Western poetry more than men do. Again, the important point to keep in mind is the detailed knowledge of the target culture necessary to modify items and to insure good, usable translations.

Another example (Butcher & Garcia, 1978, p. 473) from the MMPI asks about “hobbies.” For the Spanish translation, the seeming equivalent of “distracciones” was not used because it had “vague connotations and was thus too imprecise.” The solution in this case was to retain the English word “hobbies” since it is a familiar word to the target population and is frequently used by native speakers of Spanish. Another frequently encountered difficulty is that items formulated through negative wording (e.g. “I can read a long time without tiring my eyes”) do not read smoothly when translated into other languages. One suggested solution is to rephrase the item and to reverse the scoring key for the item. One such rewording is: “My eyes get tired when I read for a long time.”
Eventually, the results from the data-collection efforts with the translated instruments should be subjected to multivariate statistical analyses. Such activities will almost always be done “back home,” with the assistance of personnel in computer centers. Conclusions about adequacy of the translated instruments will be made based upon such results as similar factorial structures, item analyses, inclusion of specific items as indicators of specific factors, and so forth (for general discussions see Brislin, Lonner, & Thorndike, 1973; Irvine & Carroll, 1980; Miller et al., 1981; Irvine, this volume). To aid in these complex data analyses, Butcher and Clark (1979) suggest that detailed information concerning translation efforts accompany each item on the final versions of the data-collection instruments. This information is then taken into account during data interpretation. The information would include the following:

1. Indications of which items were difficult to translate, and for what reasons. Reasons can include grammatical problems; problems stemming from seemingly good equivalence, but with the necessity of choosing a lower frequency-of-use word in the target language; good linguistic equivalence, but probable differential experience across cultures with regard to a concept or activity.

2. Indications concerning which items were modified to obtain cultural in contrast to linguistic equivalents, as discussed above with the “shovel snow” and “Lincoln-Washington” items. The reasons for the need for cultural modification should be given, as well as the rationales for the final choices of target-language wording.

3. Indications concerning which items were translated literally, despite careful thought given to the possibility of cultural equivalents. Again, reasons for final choices should be given. In some cases, the reason will be that respondents will be familiar with the meaning of an item even though they do not directly experience the activities indicated in the item. For instance, people in the deep South of the United States can respond to the “shovel snow” item because they are familiar with the activity through pictures in the mass media, stories they read as children, and so forth.

4. Indications concerning which items needed grammatical modification. Often the tense of original language items must be modified. As previously discussed, many languages do not have good equivalents of the English subjunctive.

5. Indications concerning which items contained idioms in the original language version which were rendered into standard, nonidiomatic phrases in the target language. “Feeling blue” from an English language test might best be translated as the more direct “depressed,” especially when the word for the latter is used frequently in everyday conversations. Since many readers of such explanations will be from nonEnglish-speaking countries, researchers should also explain the meaning of the original language version. Such explanations are especially important since the meaning of items, and the diagnostic inference from item endorsements, changes over time. The MMPI item involving a choice concerning whether or not to label oneself as a personal representative of God may be far less diagnostic of bizarre thinking in the 1980s given the widely disseminated phrases used by devout (sometimes called “born again”) Christians.

6. Indications concerning which items were changed from negative to positive wording, or vice-versa, necessitating changes in the scoring key.

7. Indications concerning which items were easily translated, with no need for modification, but about which researchers have doubts. The doubts could be due to relevance of the concepts or activities designated in the items, the relative frequency or infrequency of activities across cultures (e.g. people may have little time for “hobbies” in subsistence-level societies), possible inappropriateness due to trampling upon cultural taboos, and so forth.

As more and more such information is made available about translation procedures with existing tests, future efforts to develop new translations should be far easier and more psychometrically sound. It is hard to imagine a translation effort of an existing test not benefiting from Butcher’s (1982; Butcher & Clark, 1979) analyses of cross-cultural investigations using the MMPI.

Guidelines for Developing New Items

If researchers choose to use existing tests and to add to these tests for the purpose of investigating emic aspects of a concept in other cultures, then they must concern themselves with developing new items. There are very definitely two schools of thought on new item development: some researchers point to its ease, and others point to its difficulty: Burisch (1984, p. 219) presents arguments from the standpoint of economics, an important consideration given the time...
needed for, and expense of, cross-cultural studies.

Not too much need be said about economy in the sense of construction economy. Some of the better known inventories, like the MMPI or the Sixteen Personality Factor Questionnaire (16PF), have taken years to build, not to speak of the funds spent and the man-and-woman power consumed in the process. In contrast, it cost me two hours and a bottle of wine to write an aggression and a depression scale that turned out to be of equal or superior validity, compared to much more sophisticated instruments [Burisch, 1984, Table 3]. As I learned later, even weeks of discussion over item formulations did not lead to scales that were appreciably better than that.

Another proponent of the idea that item development is a straightforward task undemanding of massive resources is Campbell (1968, p. 255). Given Campbell's visibility in cross-cultural research, and the provocative nature of these ideas, his position has been surprisingly undercited.

[Researchers should not] feel that tests are well constructed just because they have been expensively conceived or are much used. The methodology in test construction is in such flux that any well trained graduate student today can construct a better test: (for cross-cultural or intracultural purposes) than the Manifest Anxiety Scale, the MMPI, the F-scale measure of authoritarianism, etc. For the applied practitioner it may be well to use a test rigidly in its original format as a magical bundle of rituals of which he does not dare disturb any part. But [this] is not science, and one does not then know what aspect of the test is producing its correlates. For the scientist's test of intersubjective communicability, loyalty is required only to the theoretically relevant aspects of the tests—all the rest of the vehicular specifics should be free to vary.

On the other hand, Gorsuch (1984, p. 235) is of the opinion that the development of new measuring devices is very difficult. His analysis is specifically concerned with the measurement of religious phenomena:

Even a single, simple scale requires a considerable amount of time and empirical analysis to have any possibility of competing on psychometric grounds with previously existing scales. This means that scale development in the psychology of religion has progressed sufficiently so that it should generally be left to those with advanced training in the psychology of religion, in scale development, and in related topics such as factor analysis. Certainly, it is not a task for the average master's level project.

As with all quotes taken from longer articles, the authors' thoughts are contextualized. Campbell (1968) distinguishes between diagnosing or demonstrating the presence of a phenomenon and understanding it, and his argument is that new scales are necessary for theory development to incorporate data gathered in other cultures. Burisch (1984) argues the merits of the deductive approach to scale development in which item generation follows after very careful definitions of explanations of a phenomenon. To use his examples, the straightforward creation of items could take place only after very careful definitions of depression and aggression were written. Such definition, when written about concepts in other cultures of which researchers are not members, is undoubtedly a difficult but not insurmountable step. Jackson (1975) has shown that educated members of a culture, in his research college students untrained in psychometrics, can write usable items as long as they are provided with good definitions of the researchers' concepts under study. Gorsuch (1984) argues that many good scales exist for the study of various well-defined religious phenomena and the researchers must make a case that there is a need for new ones. The resources needed for the development of new scales, which are extensive in his opinion, could go into more data gathering with existing instruments so as to obtain the benefits discussed previously.

Even understanding the context of these positions concerning scale development, disagreements will exist among researchers for many years to come. Such disagreement is part of the "flux" to which Campbell refers. For readers who decide to take up the call for new scale development, or emic additions to etic scales, the following guidelines are offered. These suggestions are meant to help answer these queries:

I am in another culture working on a research project, but had to leave my library behind. I have identified an indigenous concept which I would like to operationalize by means of a measuring device. In another case, I found an existing test useful, but there seemed to be aspects of the concept as exists in this culture which are not picked up...
by the existing test. What are some hints for writing new items for these two types of problems?

One guideline is to have previously followed the Boy Scout motto, "Be prepared." Prior to field work, there should be a careful literature review, not just of the psychological literature but also the writings of anthropologists, sociologists, political scientists, missionaries, governmental officials, and so forth. Fortunately, much of the relevant materials have been efficiently catalogued by the developers of the Human Relations Area Files (HRAF: see Barry, 1980, and Chapter 3 of this volume for descriptions concerning its use), a resource available either in printed form or on microfilm in many large university libraries. Cataloguing has been done not only by culture, author of materials, and the author's disciplinary background, but also according to concepts covered. Use of HRAF is mandatory since it is one of the few places where researchers can obtain excellent but out-of-print materials on various cultures. One use of the HRAF would be to designate the research topic (e.g., mental illness), examine the materials with potentially applicable etic concepts in mind (e.g., the nature of intervention by outsiders), and to search for emic colorings to the etic. One difficulty with HRAF is that coverage reflects the interests of the people who lived in the cultures. There is more information of use to anthropologists than to psychologists. Still, for many topics which are the focus of research in several disciplines (e.g., alcohol use, the family, sex role distinctions), much valuable information will be found. Another benefit, in my experience, is that if researchers are familiar with the written ethnographic materials about a culture, they are respected by indigenous people. If researchers show that they have a good knowledge of local customs, as learned through their library study, then indigenous hosts are quicker to cooperate in the researcher's own investigation.

If a number of perplexing problems can be overcome, collaboration with colleagues from other cultures can yield tremendous benefits (Brislin, 1979; various chapters in Bulmer & Warwick, 1983). Difficulties stem from decisions regarding the relative status levels of research team members, the dangers of imposition by the best-funded members, decisions concerning publication credit and order of names on the final book or journal article, as well as the typical personality clashes which exist in any team effort. If such problems can be overcome, one way to benefit from the varied cultural backgrounds of team members is to ask each person to be a spokesperson from his or her culture. They could examine a concept, such as depression, and ask themselves, "How would a person from my culture manifest this psychological state? What might they say about themselves, and why?" Those statements might then be good items which the eventual target population can endorse or not. Another approach might be to examine items on established tests and to ask, "Which of these might people in my culture endorse if they are depressed? Do these established items remind me of other things depressed people might say about themselves, or about things they would deny? Do any of the established items demand modifications before they are usable?"

Pre-test work with bilingual and bicultural people can also yield good items (Wesley & Karr, 1966). Such people have lived in two or more cultures, perhaps as a function of extensive study abroad, traveling with their parents as a result of their job assignments, or a personal history of immigration (Brislin, 1981). These people can be asked to think about a concept, and then asked to reflect upon it from the viewpoint of one culture, then the other. Researchers should be careful about the potential imposition of etics by starting with emic concepts and then working up and down through various levels of abstraction (e.g., a culture's broad labels as well as very specific behaviors through which a concept is manifested).

Another approach is to take advantage of the research literature concerned with bilingual individuals. Ervin-Tripp (1964) showed that people are able to tell quite different stories when asked to reflect upon pictures from the Thematic Apperception Test (TAT), depending upon which language they use. Similarly, researchers can present bilinguals with a variety of stimuli (pictures, items, stories, experimental situations) and ask them to comment, first in one language, then the other. Controls could be added, such as counterbalancing for order of language; and use of one language only for a percentage of the stimuli to guard against simple repetition of what was already said. Responses could then be analyzed in terms of what is common across all the bilinguals' verbalization, and what is specific to each language.

Researchers can also use the Human Relations Area Files in innovative ways. A commonly shared goal in item generation methods under discussion here is that researchers want people to be
verbal so that their responses can later be content analyzed. The key, then, is to find approaches which bring out large amounts of verbal output. Some questions put to respondents bring blank stares, such as "Tell me about your culture." Such a question is too broad and has too little focus to help the respondent know what is desired. But if the question is specific, or if the respondents are asked to comment about specific stimuli, then there is a much greater chance of helpful replies. Brislin and Holwill (1977) found that 105 educated respondents from 24 different countries, or distinct cultural groups within a political entity, had no trouble reading and commenting upon written materials about their cultures drawn from the Human Relations Area Files. They were able to label certain points as correct or incorrect; suggest corrections; make expansions on certain points; bring other points up to date; and make recommendations for better research methodology. Brislin and Holwill found that the respondents enjoyed their task, finding the concept of insiders commenting on the work of outsiders provocative. In most cases, the insider-respondents knew that certain anthropologists had done work in their cultures, but few of the 105 respondents had read the materials prior to participation in this research project.

If researchers find that respondents are skillful at giving answers to questions which demand that they fill in blank spaces, or complete sentence stems (e.g. Phillips, 1960), or place marks on semantic differential scales, then the antecedent-consequent method can be used (Triandis, 1977). Using this approach, researchers phrase questions in the following form to obtain perception of what is necessary for a phenomenon to occur (the antecedents), and what happens once the phenomenon does occur (consequences). Using the example of depression again, the questions would take the form:

(Antecedents)
If you have _______, then you have depression.
If _______ happens, then you are depressed.
If _______ happens, others will know that I am depressed.

(Consequences)
If I am depressed, then I _______.
If others think that I am depressed, then _______.
If I am depressed, I then try _______.

Like the debate concerning the straightforwardness versus difficulty of generating new scales, these and similar approaches which ask respondents to complete sentences or to make check marks on scales divide researchers into various camps. Anthropologists, especially, are wary of questionnaire approaches, fearing imposition of concepts and preferring long-term participant observation. There may be more disagreement among neophytes, however, than among experienced researchers. Often, judgments about another discipline’s methods seem based on the worst examples, not those studies which have made important contributions. Once researchers become secure in their identity and do not have to establish their credentials by criticizing other disciplines, they often draw from a variety of approaches in their actual research endeavors (e.g. Howard, 1974, trained in anthropology but sophisticated in questionnaire development). Irvine (1968, p. 3), trained in psychology, has pointed to the benefits of participant observation: “to collect valid data the psychologist needs adequate and sympathetic training in understanding a society system that is alien, complex, and conceptually different. Participant observer research will acquire greater scientific status, it seems, as a result.”

**TRANSLATION: A RECOMMENDED PROCEDURE**

A close relation exists between the modification of existing instruments, and the creation of new items, and subsequent translation into other languages. If materials are prepared according to the suggested guidelines, they are more likely to be translatable than if the guidelines were ignored. Further, much information can be gained about desirable wording in both the target and original languages through examinations of the translators’ efforts.

The recommended procedure is back-translation, and its close relative, decentering (Werner & Campbell, 1969; Brislin, 1970; Brislin, Lonner, & Thorndike, 1973). In back-translation, one bilingual translates from the source to the target language, and another blindly translates back to the source. The procedure can be repeated for several rounds, as different bilinguals work with the efforts of their predecessors. The work of the team of bilinguals, each working independently on a different step, can be depicted as follows:
Moving back and forth between languages in this way is the basis of decentering, since no one language is the “center” of attention. The procedure can be repeated for several more rounds. The researcher then compares the last back-translated version with the original version. If a concept “survives” the decentering procedure, it is assumed to be *emic* since there must be readily available words and phrases in the two languages which the translators could use. If a concept is not in the final back-translated version, the reason could be that it is *emic*. That is, the concept might be readily expressible in only one of the languages. At this point, there should be extensive discussions with the bilinguals who can indicate reasons why materials were and were not translatable. For instance, Phillips (1960, p. 302) found it very difficult to have this sentence stem translated into the Thai language: “Sometimes a good quarrel is necessary because ...” He wrote, “After much discussion the translators decided that although it was conceivable that an American might enjoy a quarrel for its cathartic effect, the notion would be incomprehensible to a Thai.” To use the terms already introduced, “having a good quarrel” is not an *etic*. Rather, the translation procedure has pointed up a very important difference which can be built into cross-cultural investigations of interpersonal relationships.

Another benefit to this procedure is that it guides final decisions about the wording of the original language version which will eventually be used in actual data collection. That version will often be the final back-translation. The reasoning for this suggestion is that the final back-translation is most likely to be equivalent to a target language version, probably the version immediately preceding the final step in the decentering process (see diagram). For instance, Brislin (1970) wanted the following item from the Crowne and Marlowe social desirability inventory (1964) to be translated into Chamorro, the language of Guam and the Marianas Islands. “I have never intensely disliked anyone.” The final back-translated version was “I have never really disliked someone.” Discussions with translators yielded the facts that it was difficult to translate “intensely” (translators often have difficulty with adverbs), and the equivalent of “someone” was more readily available than for “anyone.” It was this back-translated version which was later used in data-gathering with English-speaking respondents. The position defended by Campbell (1968) cited previously was taken very seriously. The topic of investigation was the need for approval, and it is measured by a set of statements which allow people to present themselves in a positive light. If changes from “intensely” to “really,” and from “someone” to “anyone,” change the meaning so much that there will be a differential response, then the underlying concept is weak. Still, changes in stimuli cause discomfort since Campbell’s position is not universally accepted. If researchers are worried that these sorts of changes can lead to noncomparability with previous studies which have used the original version, they can gather data using both versions. English-language respondents can see versions which contain one-half original items, and one-half revised based on decentering efforts. Basic controls can be added to counterbalance the order of original and revised items. Data analysis can then indicate whether or not different conclusions should be made about respondents based on their reactions to original versus revised items. Brislin (1970), for example, found no differences and concluded that the items on the Marlowe-Crowne scale are robust enough to allow modifications based on what is easily expressible in other languages. This conclusion seems similar to that of Butcher (1982), who has written most extensively about cross-cultural use of the MMPI.

Like any single method or approach, back translation is no panacea. All materials should be pre-tested with respondents similar to those in the proposed main sample since there will always be items which simply do not work well in actual use. The major advantage of back-translation is that it gives researchers some control over the instrument development stage since they can examine original and back-translated versions and make inferences about the quality of the translation. Researchers rarely know the target language well enough to do their own translations. Even researchers who are native speakers, because of the large number of years they have devoted to their formal education, may use phrases which are unfamiliar to the sample of respondents. At times, they can prepare first drafts and then ask *monolinguals* similar to the respondents to rewrite the
material so that it will be clear to native speakers. In such a case, the sequence would be:

original to target to target rewrite to original

bilingual #1 a monolingual bilingual #2

This sequence can also be used if the researcher is concerned that the back-translation is better than the target-language version. This problem can occasionally arise if a highly-skilled bilingual can make sense out of a mangled target version. Just as all readers of this chapter can make sense out of broken English when interacting with visitors from other countries, back-translators can sometimes take very rough materials and polish them so much that they appear as “diamonds” in the final version. Adding the target rewrite step minimizes this possibility, but even here pre-testing is necessary.

I have frequently given this advice regarding pre-testing, but have been met with this complaint: “It took me a great deal of effort to receive permission to do research in such-and-such an organization in another country. I have limited time there, and have no facilities to duplicate a revised instrument in any case. How can I pre-test?” The answer is that pre-testing in the other country may not be possible on any large scale, but work with approximately ten respondents is still wise to determine what questions they will have and what items need oral explanation in the instructions-to-respondents step. I once advised a researcher who was about to work in a Malaysian factory, but who would have no opportunity for extensive pre-testing, to do the pre-test work in his own country using the back-translated English items. He listed key aspects of his eventual sample in Malaysia: limited education, not familiar with filling out questionnaires, not much experience with researchers, limited reading skills in their native language, and so forth. He then sought out a factory in his own country which employed similar people. He administered the English-language version there, identifying a number of places where there were communication difficulties with respondents. He then made changes, incorporating them into the Malaysian version. He reported to me that he had no problems in the actual data gathering in Malaysia.

ADMINISTRATION OF RESEARCH INSTRUMENTS

Although space limitations prevent an extensive discussion, a few guidelines can be suggested for the administration of measuring instruments. Usually, administration will be in the form of interviews, questionnaires, or formal tests. Various biases in instrument administration have been given names (e.g., Pareek and Rao, 1980). The rudeness bias occurs when researchers forget that they are requesting the valuable time of people in other cultures and plunge ahead with their questions without knowledge of local norms. The sucker bias occurs when respondents give nonsensical answers in the spirit of fun, to see how much of their silliness the researcher is willing to record. The “I-can-answer-any-question-bias,” more prevalent in some parts of the world than others, refers to the deep-seated belief among respondents that no question should go unanswered. People thus have answers, however poorly formulated, on topics ranging from the most appropriate age at which children should start writing, to fighting inflation through the planned introduction of an economic recession.

The courtesy bias occurs when respondents try to discover what the questioner wants, and then direct their answers accordingly. The hidden premises bias occurs, more frequently than many researchers believe, when respondents try to discover who the researcher really represents and what the researcher is really trying to learn. Often, researchers are judged based on the source of funding for their studies, and any government-funded investigation is seen as an attempt to have that government’s influence permeate the local culture.

There is no single magic solution to dealing with these potential biases. As with any interaction in another culture, sensitivity to the host point of view and to the time they invest in a study is extremely important. Thorough preparation, including extensive reading of the ethnographies written about the culture, results in an ability to see through the sucker bias. Researchers may want to strive for the goal of good lawyers: never ask a question of a witness unless the answer is already known. While impossible to achieve in all cases, the goal is a good one to keep in mind. Researchers should have a good
understanding of the potential range of answers to their questions or else they will be unable to make distinctions between honest attempts to answer questions and error of various kinds.

A specific technique for dealing with potential biases has been suggested by Schuman (1966). Called the “random probe,” researchers select a random sample of items from their instruments for a probing follow-up question. The sample of respondents which receives the probe for a given item is also randomly selected. The probe might be, “What do you mean?”, or “Could you tell me more about that?” The answer to the probe from any one person should make sense, especially when compared to the answers provided by other respondents. Further, the quality of answers can be judged according to whether or not answers make sense given the answers to other probing questions. Answers to all the probes should provide excellent contextualizing information which will make interpretation of the original measuring instrument more precise and will allow distinctions to be made between biased and valid answers.

NOTE

1. Some readers will undoubtedly argue that French and/or Spanish could be added here. Some of the guidelines to be presented are also applicable to materials prepared in these languages. Even proponents of research carried out from the beginning stages in indigenous languages, to study indigenous phenomena more precisely and to avoid imposed etics (e.g. Enriquez and Marcelino, 1984, in the Philippines) recognize the need for eventual publication in a language of world-wide use.