First edition published under the titles *Handbook of Cross-Cultural Psychology*, Volume 1, *Perspectives*, edited by Harry C. Triandis and William Wilson Lambert, and *Handbook of Cross-Cultural Psychology*, Volume 2, *Methodology*, edited by Harry C. Triandis and John W. Berry, copyright © 1980 by Allyn & Bacon.

Library of Congress Cataloging-in-Publication Data

Handbook of cross-cultural psychology. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Contents: v. 1. Theory and method / edited by John W. Berry, Ype H. Poortinga, Janak Pandey.
ISBN 0-205-16074-3 (v. 1)
1. Ethnopsychology. I. Berry, John W.
GN502.H36 1996
155.82—dc20 96-16261
CIP

Printed in the United States of America
10 9 8 7 6 5 4 3 2 00 99 98

CONTENTS

7

METHODS AND DATA ANALYSIS OF COMPARATIVE RESEARCH

FONS VAN DE VIJVER
Tilburg University
The Netherlands

KWOK LEUNG
Chinese University of Hong Kong
Hong Kong

Contents

Introduction

The major goal of this chapter is to provide a comprehensive overview of the methodological issues encountered in cross-cultural research. Since the reviews in the first edition of the *Handbook* on testing and assessment by Irvine and Carroll (1980) and on experimentation by Brown and Sechrest (1980), many developments have taken place. In our presentation, we focus on data sets that are comparative in nature. Most studies of this type involve data from at least two cultural groups, but some studies are monocultural. In such studies, previous work must provide data and results before meaningful cross-cultural comparisons to be made. Monocultural studies commonly conducted by ethnographers and anthropologists that do not touch upon cross-cultural comparison will fall outside of the scope of our review.

We see the process of conducting cross-cultural research as composed of three important steps. First, the research questions must be explicitly stated. Second, a method that is appropriate to the research questions raised should be selected. Method is defined here as the design, sampling, administration, and instrumentation involved in the collection of data. Finally, the appropriate data analysis should be chosen in light of the research questions raised and the method chosen. We consider these three steps as intertwined, and they should be considered simultaneously prior to data collection. This three-step framework is used in organizing the materials that follow.

The first section of the chapter describes specific issues of cross-cultural research, such as quasi-experimentation. The second section describes in more detail the methodological aspects of cross-cultural studies. The third section deals with the analysis of cross-cultural data. The fourth section reviews the main issues in the methodology and analysis of four common types of cross-cultural studies. Conclusions are drawn in the final section.

Specific Issues in Comparative Methodology and Data Analysis

Before the methods and analyses of cross-cultural studies can be discussed, the applicability of "true experiments" (Campbell & Stanley, 1966) and the associated statistical framework to these studies—the Neyman–Pearson theory—should be explored in order to highlight their special characteristics.

The classical Neyman–Pearson theory provides the most commonly applied statistical framework in testing intergroup differences in psychology. The framework is appropriate for analyzing data from experiments with experimental and control groups. The two groups are considered to be equal, except for the manipulation that is present in the experimental group and absent in the control group ("all other things being equal," as it is often called). The theoretical question the researcher wants to examine concerns the presence of a difference in the dependent measures between the experimental and control groups. This is tested by a t test or analysis of variance. The researcher chooses *a priori* a probability

that is considered appropriate, usually .05 or .01, for concluding whether or not there are differences between the experimental and control groups. The framework has been developed as a tool to analyze data collected in experimental settings and to reduce the risk of making false inferences.

The framework has turned out to work well mainly in so-called true experiments (Campbell & Stanley, 1966), in which subjects are randomly assigned to different experimental conditions. The "all other things being equal" argument in general does not apply to studies in which subjects are not assigned randomly to experimental treatments. Group membership, a major experimental treatment in cross-cultural studies, is predetermined and cannot be randomly assigned. When the cultural differences between the groups of subjects involved in a cross-cultural study are extensive, it does not make much sense to assume the validity of the "all other things being equal" argument and to compare the groups as if the data were collected in a true experiment. In cross-cultural studies, the application of the Neyman–Pearson framework can yield misleading results (Poortinga & Malpass, 1986). For instance, when cognitive tests are presented to Western literate and non-Western illiterate subjects, the educational and cultural differences between the two groups tend to be so massive that a test of the null hypothesis of no intergroup differences in performance is inadequate. Quite likely, every item will show a significant difference between the two groups.

Furthermore, the interpretation of such a test is equivocal. In the experimental paradigm the interpretation of the difference in the dependent measures is simple. The treatment, typically well defined, such as a drug that has been administered, has produced the score difference between the experimental and control groups. In a similar vein, the differences in the cognitive tests between the literates and illiterates can be attributed to the treatment "culture." However, the attribution does not convey much meaning and dodges the question of a proper interpretation of the score differences. Culture is too global a concept to be used as a meaningful independent variable in the interpretation. In comparison to the experimental branches of psychology, cross-cultural psychology should be much more sensitive to the interpretability of findings. Whereas the task often ends for the experimental psychologist with the observation of a significant difference because the observation will typically confirm or falsify a hypothesis, the task of the cross-cultural psychologist is certainly not complete with the observation of significant intergroup differences.

A crucial problem in quasi-experiments (in which there is no random assignment of subjects) is the ruling out of rival hypotheses. This issue has been extensively discussed, and the consensus is that culture must be "unpackaged" (e.g., Whiting, 1976; Poortinga, Van de Vijver, Joe, & Van de Koppel, 1989). The use of culture as an explanatory variable is not satisfactory, and culture must be decomposed into a set of psychologically meaningful constructs, which are then used to explain the cultural differences observed (e.g., Leung, 1989; Poortinga & Van de Vijver, 1987). When cultural differences on a dependent variable are documented, it is almost impossible to pin down which aspect of culture is responsible for the observed differences, in the absence of additional data. A feasible strategy is to identify the most likely variables that may account for the expected cultural differences and measure these variables in the study. A number of analytical procedures, which will be described later, can be employed to identify which aspect is indeed the most plausible explanation for the cultural differences observed. An adequate cross-cultural study must have built-in elements in its design to rule out plausible rival hypotheses (cf. Cook & Campbell, 1979).

Equivalence

Equivalence is a major concern in cross-cultural research; meaningful cross-cultural comparisons can only be made if the data from different cultures are comparable. Equivalence has been discussed extensively, and several types have been identified (e.g., Berry, 1969; Poortinga, 1971, 1989; Van de Vijver & Poortinga, 1982). Because terms used in the literature to describe equivalence are often unclear and confusing, we propose that three types of equivalence be distinguished: *structural, measurement unit,* and *scalar*. Cross-cultural researchers are often interested in *structural equivalence,* which refers to the similarity of psychometric properties of data sets from different cultures. Specifically, psychometric properties are often taken to refer to correlations of the items of an instrument (*instrument* is used in this chapter for any measurement device such as tests, questionnaires, and observational scales) or to correlations of an instrument with external measures. Multidimensional scaling, factor analysis, and the analysis of covariance structures (structural equations) are commonly employed to study structural equivalence. Thus, if equal factor structures are obtained in various cultural groups, it can be concluded that the psychological constructs underlying the instrument are identical. However, structural equivalence does not imply that both the origin and the measurement unit of the instrument are identical. Structural equivalence is primarily based on similarity in correlations across a variety of cultures and correlations are not affected by linear transformations of the variables. For example, if the scores of all persons in one cultural group are multiplied by a positive constant, the correlations remain unaffected and the factor loadings will also remain the same. Therefore, similar factor loadings can arise from scales with different origin and measurement units.

The second and third types of equivalence are concerned with measurement equivalence. When the scores of two cultural groups are compared, it is possible that the unit of measurement is identical, but that the scales do not have a common origin. This will be called *measurement unit equivalence.* Temperature scales in degrees of Celsius and Kelvin show this kind of equivalence. It has been argued that some intelligence tests can be validly applied within but not across cultural groups due to different origins of the scale in the cultural groups. In the case of measurement unit equivalence, differences between two scores (e.g., the scores between two classmates or the scores of an individual at two measurement occasions) can be compared both within and across cultures, while the scores themselves can only be compared within cultures.

If it can be ascertained that scores show not only an identical unit of measurement, but also a common origin, *scalar equivalence* or *full score comparability* is said to have been obtained. Scalar equivalence allows the comparison of the scores obtained, both within and across cultural groups. Examples are such variables as weight and height. For psychological measurements it is often difficult to establish scalar equivalence. In general it is easier to disprove than to prove scalar equivalence.

In the cross-cultural literature, the term *metric equivalence* has often been introduced to refer to the case when two or more data sets from different cultures exhibit similar psychometric properties (Berry, 1969). Within this framework, *subsystem validation* refers to the case when independent and dependent variables show the same relationship within cultures and across cultures (e.g., Roberts & Sutton-Smith, 1962). *Scalar equivalence* refers to the case in which scores from different cultures have a similar origin and unit of measurement (e.g., Poortinga, 1971).

We find some of this terminology imprecise. The term *metric* in metric equivalence denotes the unit of measurement in common usage in the psychometric literature, and does not denote structural equivalence nor a common origin of the scores, both of which are implied in the current usage of the term. Thus, we propose that this term be abandoned.

The first subtype of metric equivalence, subsystem validation, is actually a special case of structural equivalence, and can be subsumed under structural equivalence. The second subtype, scalar equivalence, is defined in the same way as in our scheme, and should be retained.

Methods

Sampling of Cultures

The selection of cultures in a cross-cultural study is often central to its scope for evaluating the hypotheses proposed. Three types of sampling procedures for the selection of cultures are commonly found in the literature. First, *convenience sampling* is often adopted in cross-cultural studies. Researchers select a culture simply because they may be from that culture, are acquainted with collaborators from that culture, or happen to be spending a sabbatical leave in that culture. The choice of culture is haphazard, driven by convenience, and not related to the theoretical questions raised. Very often, these studies adopt a "let's look and see" approach and do not develop any *a priori* predictions about cultural differences. When cultural differences are found, post hoc explanations are often developed to explain the differences.

The second approach is *systematic sampling*, in which cultures are selected in a systematic, theory-guided fashion. Usually, cultures are selected because they represent different values on a theoretical continuum. The classic study by Berry (1967) provides an excellent example of this approach. Two groups were studied, one agricultural and one hunting. It was hypothesized that agricultural societies impose stronger pressure on conformity, and hence will lead to field dependence. Hunting societies encourage their members to be autonomous and hence are conducive to field independence. These two groups were selected systematically to evaluate this hypothesis. Another example of this approach is provided by Leung, Au, Fernandez-Dols, and Iwawaki (1992). In their study, four cultures were selected, namely, Spain, Japan, Canada, and the Netherlands. Japan and Spain tend to collectivistic, whereas Canada and the Netherlands tend to be individualistic (Hofstede, 1980). The comparison of these two groups will reveal the impact of individualism–collectivism. On the other hand, Spain and the Netherlands tend to be feminine, whereas Japan and Canada tend to be masculine (Hofstede, 1980). The comparison of Spain and the Netherlands with Japan and Canada will reveal the effects of cultural masculinity and femininity. An interesting feature of this study is that in both types of comparison, each group is composed of a Western and an Eastern culture. If differences are found between the two groups, the possibility that the differences are due to East–West differences can be ruled out.

We believe that in the systematic approach, bicultural comparisons are adequate only if there is a compelling theoretical framework in which the results can be interpreted, as is the case in Berry's (1967) study. When a study is exploratory, or when the theoretical framework guiding the study is rudimentary, the number of cultures in a study should be preferably larger than two. Campbell (1986) argued that the number of rival explanations is greatly reduced when the number of cultures involved in evaluating a hypothesis increases (cf. Leung et al.'s, 1992, study mentioned above).

In order to maximize the effectiveness of the systematic approach, cultures that are far apart on the theoretical dimension upon which they vary should be selected. This approach will maximize the chance to detect cultural differences. However, if only two cultures are selected that are highly dissimilar, they are likely to vary in other dimensions as well, and numerous alternative interpretations have to be ruled out. The problem does not arise when more than two cultures are studied; the larger the number of cultures selected, the fewer the alternative interpretations will be possible.

The third approach is *random sampling*. In this approach, a large number of cultures are randomly sampled, usually for evaluating a universal structure or a pan-cultural theory. Truly random samples are basically nonexistent in the literature, as no one has the resources to select a large number of cultures on a random basis for a single study. However, several studies have tried to follow this approach, and their sample may eventually begin to approximate a random sample (usually not of all groups but of all literate groups). For instance, Schwartz (1992, 1994) has sampled 36 cultures to evaluate the structure of human values. He basically included any cultural group in which he could find a collaborator to participate in the project. Buss et al. (1990) also followed a similar approach in sampling 37 cultures in their study of mate selection. Peterson et al. (1995) have surveyed managers from more than 20 countries on event management issues.

Sampling of Subjects

In order to make valid cross-cultural comparisons, the subjects from different cultural groups must be similar in terms of relevant background characteristics. Otherwise, it is hard to conclude whether the cultural differences observed are due to cultural differences or sample-specific differences. If we compare a group of illiterate subjects from one culture to a group of highly educated subjects from another culture, the differences observed are likely to be explainable in terms of educational differences rather than differences in some other aspect of their cultures. One approach to overcome this problem is to match the samples in terms of demographic characteristics so that sample differences can be ruled out as alternative explanations for observed cultural differences. For instance, college students from different cultures are often compared, and it is usually assumed that college students from different cultures are similar in their demographical characteristics. In a similar vein, Hofstede (1980, 1983) reduced the influence of unwanted intergroup differences by studying subjects from a single multinational organization from 53 countries. Schwartz (1992, 1994) sampled secondary school teachers from various countries to maximize the comparability of his subjects.

It is sometimes impossible to match samples from different cultures because of practical reasons, or because there are sharp cross-cultural differences in the demographic background of subjects. An adequate approach is then to measure the major demographic variables and treat them as covariates in the subsequent data analysis. For instance, in a study comparing the delinquent behaviors of adolescents in the United States, Australia, and Hong Kong, it was found that there were substantial differences in the father's educational standing in the three cultures (Feldman, Rosenthal, Mont-Reynaud, Leung, & Lau, 1991). The educational standing of the fathers of the Hong Kong subjects was significantly lower than that of the fathers of the Australian and American subjects. To overcome this problem, an analysis of covariance was used to compare cultural means partialling out the influence of father's educational standing.

It is unfortunate that many cross-cultural studies tend to ignore sample differences and fail to assess the impact of such differences. As the results are confounded by sample differences, it is difficult to provide an unambiguous interpretation.

Procedure

In this section we will review issues related to the procedural aspects of a cross-cultural study: the selection and evaluation of the adequacy of a measurement instrument, its translation, and its administration.

In an early stage of a project the question has to be raised whether the same instrument can be applied in all cultural groups. In the case of an already existing measurement instrument, its appropriateness in an intercultural context has to be judged. This amounts to answering the question whether the operationalizations

chosen in the instrument will be adequate in all cultural groups studied. Are the measurement operations specified in the instrument an adequate representation of the psychological domain that is to be covered? Embretson (1983) has introduced the concept of construct representation. The concept refers to the coverage of the psychological domain. Do the measurement operations specified in the instrument represent an adequate and sufficient sample of the behavioral manifestations of the psychological construct that is measured by the instrument? Any answer to this question requires knowledge of the cultural context in which the instrument will be applied.

The outcome of the decision process can take three forms: to *apply* the instrument, to *adapt* it, or to *assemble* a new version. In the first alternative the instrument or a translated version will be used without any modification. If the construct is not fully covered in the new group, the instrument can be adapted by rephrasing, adding, or replacing items that measure the missing aspects. If the researcher finds the original instrument entirely inadequate, a new instrument has to be assembled.

The decision whether to apply or adapt an existing instrument or to assemble a new one has both theoretical and practical implications. We propose to make the application of the same instrument the default choice. The advantages of this choice are (1.) the possibility to compare research results with other results reported in the literature, (2.) the possibility to maintain scalar equivalence (which is not achievable if results of newly assembled instruments are compared), and (3.) the small amount of money and effort that is required to administer an existing instrument as compared to the development and establishment of the psychometric properties of a new or adapted instrument. However, the direct application of an existing instrument may not always be the best choice. If an instrument does not cover important aspects of the psychological construct under study or if it shows a clear ethnocentric bias, adaptation or the assemblage of a new instrument would be a better choice. The decision may be seen as involving a cost–benefit analysis, with time and money as the costs and construct representation as the benefit.

There are numerous examples of *application* in the literature. For instance, Hofstede's (1980, 1983) classic study involves a value questionnaire that was administered in over 10 languages in 53 countries. The use of the Minnesota Multiphasic Personality Inventory (MMPI) in China provides a good example to illustrate the process of *adaptation.* When the items of the MMPI were tested in China, it was found that some items were meaningless in the Chinese context, and these items had to be modified (Cheung, 1989). However, most of the original items in the MMPI were retained, and it was actually possible to interpret the Chinese results in light of the American norms. The case of *assembling* a new instrument is rare in the literature, but two examples can be cited. Church (1987) argued that Western personality instruments are unable to capture many of the indigenous personality constructs of the Filipino culture. In light of these difficulties, he proposed a number of directions for the construction of a new personality instrument for the Filipino culture. In a similar vein, Cheung et al. (1996) have

argued that adaptation of Western personality instruments is inadequate in capturing all the major dimensions of personality in the Chinese culture. They started from scratch and created a personality instrument, called the Chinese Personality Assessment Inventory (CPAI), for the Chinese people. This instrument contains several indigenous personality dimensions, such as "face" and "harmony," as well as many items that are particularly meaningful in the Chinese context.

Instrument Translation

In the case of the *application* and the *adaptation* the instrument has to be translated. The translation–backtranslation method is probably the best known method for instrument translations (e.g., Brislin, 1980; Hambleton, 1993, 1994). An instrument is translated from one language to another and then backtranslated to the original language by an independent translator. This method often provides adequate results, but sometimes it produces a stilted language that reproduces the original language version well, but is not easily readable and comprehensible. This is particularly the case when test items contain local idioms that, almost by definition, are difficult to translate. Backtranslations can provide researchers who lack proficiency in the target language control of the adequacy of the translation. However, it is noteworthy that in the field of professional translations the procedure is almost never utilized (Wilss, 1982). Professional translations are commonly produced and checked by teams of competent bilinguals; hence, instead of relying on backtranslations, these teams utilize judgmental methods to assess the accuracy of the translation.

Werner and Campbell (1970) have proposed to decenter instruments that are used in a cross-cultural context—to adjust both the original and the translated versions simultaneously. The aim in decentering is not the verbatim reproduction of the original text but the enhancement of the naturalness and readability of the original and translated version.

Brislin, Lonner, and Thorndike (1973) have generated a useful set of guidelines to ensure good translatability (cf. Brislin, 1980, p. 432):

1. Use short, simple sentences in order to minimize the cognitive load of the instrument; a simple item-per-item check whether the phrasing can be simplified can lead to considerable improvement in translatability.
2. Employ the active rather than the passive voice.
3. Repeat nouns instead of using pronouns (which in some languages may be difficult to translate).
4. Do not use metaphors and colloquialisms, which are usually not well translatable.
5. Avoid the subjunctive mood (e.g., verb forms with "could" and "would").
6. Add sentences when key concepts are communicated. Reword these phrases to provide redundancy.
7. Avoid adverbs and prepositions telling "where" and "when," such as beyond and upper.

8. Avoid possessive forms where possible.
9. Use specific words, such as chickens and pigs, rather than general terms, such as livestock.
10. Avoid words indicating vagueness, such as probably and frequently.
11. Use wording familiar to translators where possible.
12. Avoid sentences with two different verbs that suggest different actions.

Various techniques have been proposed to check the accuracy of translations. An overview has been presented by Hambleton (1993, 1994). A distinction can be made between judgmental and empirical methods. Judgmental evidence of translation equivalence usually amounts to the application of a translation–backtranslation design. An assessment of the accuracy of the translation by a set of competent bilinguals is an alternative way to assess accuracy. Hambleton proposes three designs to study the accuracy of translations: (1.) bilinguals take the source and target versions of the test; (2.) source language monolinguals take the original and backtranslated versions, and (3.) monolinguals in both languages take the test. The latter is by far the most frequently applied design. Various psychometric techniques are available to evaluate the equivalence of the items in the source and target languages. These are known as item bias or *differential item functioning* techniques and will be discussed later.

Administration

Four areas will be distinguished in the following overview of issues related to a proper administration of instruments in a cross-cultural study (cf. Van de Vijver & Poortinga, 1991, 1992): the personal characteristics of the tester (or interviewer), interactions between the tester and the examinees, response procedures, and the stimuli of the instrument. In general, it will be difficult or even impossible to generate an exhaustive list of the problems that may arise in the administrative aspects of cross-cultural research. However, an overview of the common problems may sensitize the reader to the kinds of problems that can be encountered.

The presence of a tester, experimenter, or interviewer can be a threat to the validity of the results, particularly when this person has a different cultural background from the subjects in the sample. The potential influence has been recognized in observational studies of mother–child interactions (Super, 1981). In intelligence testing, the influence of racial differences between the tester and the examinee has been studied systematically (Jensen, 1980). Overall, the influence tends to be small, though the results are not consistent. In many cross-cultural studies the cultural distance between the tester or interviewer and the subjects will be considerably larger than in the American studies reviewed by Jensen. No systematic study has been undertaken of tester effects in settings more representative of cross-cultural settings.

A second area to be considered is the interaction between the tester and the respondent. In many research designs there is verbal communication between the two, and various problems may occur as a result of such communication. In

some cases the choice of the language used may be problematic. For instance, when Reuning and Wortley (1973) administered a variety of cognitive tests to the Bushmen, Kalahari desert dwellers, they faced the problem that their subjects had a highly heterogeneous linguistic background. Because it would have been difficult to hire and train an interpreter for each vernacular, they chose to minimize the verbal exchange in the testing procedure.

The reduction of verbal communication is not always possible because verbal exchange is essential in surveys and psychological testing. If the researcher decides to administer the instruments with the help of one or more interpreters, the potential influence of the interpreters should be evaluated, even when they are carefully trained. An assessment of the interpreter's influence usually requires that a group of respondents be interviewed by two interpreters. The results obtained by these interviewers are then compared with the help of an index of agreement. The choice of this index depends, among other things, on the nature of the data gathered. Cohen's kappa or its weighted version can be used in the case of nominal or ordinal data (Cicchetti, Showalter, & McCarthy, 1990; Cohen, 1960), and an intraclass correlation (Shrout & Fleiss, 1979) or Cronbach's alpha (e.g., Winer, 1971) in the case of interval data.

The third area involves response procedures. Subjects may be unfamiliar with a certain response procedure. For instance, the Porteus' Maze Test, a paper-and-pencil test, has been administered to groups of subjects who had never used a pencil before. Not surprisingly, their scores were very low (cf. Van de Vijver & Poortinga, 1991). If subjects are unfamiliar with a response procedure, it is important to reserve time for familiarizing the subjects with the procedure as part of the test introduction. In the area of personality and social psychology, Likert scales are often applied. Particularly among groups having little experience with this response format, the use of verbal descriptions of the response alternatives instead of numbers might be preferred.

A good example of the impact of response procedures can be found in the work of Serpell (1979). He administered a pattern-copying task to children in the United Kingdom and Zambia. The children's copying skills were assessed using two response media: pencil-drawing and iron-wire modelling, a popular pastime among Zambian boys. It was found that the British children scored higher than the Zambian children on the pencil-drawing task while the Zambian children reached higher scores on the iron-wire modelling task.

In some cases no empirical evidence may be available to judge the accuracy of a response procedure. A pilot study could then be carried out in which potentially useful response procedures are compared in a monotrait–multimethod matrix, in which several response procedures for measuring the same construct are examined. The correspondence of the results across the response procedures indicates the validity of the procedures.

Stimulus-related aspects are by far the most extensively studied area of procedural problems in cross-cultural research. Stimulus familiarity is the most often mentioned source of invalid intergroup score differences in the literature (e.g., Irvine & Carroll, 1980). A study by Deregowski and Serpell (1971) illustrates the importance of stimulus familiarity. Scottish and Zambian children were asked to sort miniature models of animals and motor vehicles in one experimental condition and their photographs in another one. No intergroup differences were found for the actual models whereas in the sorting of photographs, the Scottish children obtained higher scores than the Zambian children.

In the past, various attempts have been made to adapt the stimuli of cognitive tests in such a way that intergroup differences caused by stimulus familiarity would be eliminated. Both the culture-free and culture-fair test movements were intended to serve this purpose. Even though the original ideas of the movements have been long abandoned and it is widely acknowledged that such tests cannot be constructed (Frijda & Jahoda, 1966), the concern for stimulus familiarity is still widely shared. Stimuli differ in terms of their cultural entrenchment. Simple geometrical stimuli such as squares, circles, and triangles are often used as stimuli in cognitive tests because their cultural loading is assumed to be limited though certainly not absent.

In the area of personality and social psychology, stimulus familiarity also plays an important role. Items of personality scales frequently use complex words or expressions. Effort should be made to use simple, unambiguous stimuli and to avoid the undesirable introduction of verbal abilities, such as vocabulary and text comprehension skills, as sources of individual differences.

Design

A distinction will be made between the design of structure-oriented and level-oriented studies in cross-cultural psychology. Structure-oriented studies examine relationships among variables and attempt to identity similarities and differences in these relationships across cultures. For example, is the structure of intelligence universal? Level-oriented studies, on the other hand, focus on differences in the magnitude of variables across cultures. For example, are members of culture A more individualistic than members of culture B?

The design of structure-oriented studies is often straightforward: it replicates the design of the original study. The design of level-oriented studies tends to be more complicated, and an adequate choice of research variables and design is needed to enhance the interpretability of the findings obtained. There is at least one important issue common to all level studies: Which covariates should be included? It was argued before that the Neyman–Pearson framework assumes a random assignment of individuals to treatments and that cross-cultural studies can never adopt a truly experimental design. Cultural groups differ in many respects, only some of which are of interest in a particular study. All these group differences can in principle explain observed score differences. An important aid in the reduction of the number of rival explanations are covariates. Covariates can be helpful in the interpretation of cross-cultural score differences in two ways. First, they can be used to validate the interpretation of the cross-cultural differences as hypothesized by the experimenter. For instance, if individualism–collectivism is assumed to be related to a psychological phenomenon, say inter-

group hostility, individuals from individualistic and collectivistic countries could be included in the study. In addition to an intergroup hostility measure, a test of individualism–collectivism should be administered to all individuals. These scores could then be used in an analysis of covariance, in which cultural groups are the independent variable, the hostility measure the dependent measure, and the individualism–collectivism score the covariate. The covariate is used to validate the cross-cultural differences postulated by the theory. Earley (1989) has evaluated the effect of individualism–collectivism on social loafing with this approach.

Second, covariates can also be used to check the effects of nuisance variables. The inclusion of such covariates will control for cultural differences that influence the behavior in question, but that are not specified by the theory. For instance, if men and women differ in the level of hostility and if the student groups in the two cultures in the previous example have a different male–female ratio, gender could be used as a covariate, because the observed cross-cultural differences could be due to the difference of gender composition of the two groups as well as to cross-cultural differences in intergroup hostility. The covariate is not meant here to provide an explanation of the cross-cultural differences, but to control for nuisance variables. Covariance analysis as discussed in textbooks is almost always exclusively concerned with the elimination of the impact of nuisance variables. The conclusions of an analysis of covariance can be misleading if the assumption of parallel regression lines within each cultural group is violated (cf. Lord, 1967). A simple statistical test of the equality of regression coefficients in two cultural groups is described in Cohen and Cohen (1983: chapters 10 and 12) and Pedhazur (1982, chapter 12).

Covariates can be based on aggregate rather than individual measures as the previous examples could suggest. In a study of intergroup differences in some cognitive test, educational quality could be assessed. Such a measure located at the class or even cultural level can be used as a covariate at the individual level, meaning that all subjects of a class or school will get the same score on the variable.

We strongly encourage the use of covariates because they provide an effective way to confirm a particular interpretation of intergroup differences and to falsify alternative interpretations. Yet, the limitations of methodological and statistical procedures should be acknowledged. Statistical techniques can help to evaluate the impact of contextual variables, but will not provide information on which covariates to choose. For example, intergroup differences in cognitive test performance might be assumed to be related to educational quality or to Westernization, to mention a few possibilities. Methodological and statistical considerations cannot dictate the choice. All that can be asked from methodology and statistics is a set of tools to enable the evaluation of the accuracy of the choice, or, in case both sets of variables have been measured, the evaluation of their relative importance.

Leung and Zhang (1995) have concluded that many studies have been exported from the West to non-Western countries, and some of the issues examined in these studies are of little relevance to the local culture. It is entirely possible that results obtained in many of these studies are shaped by the cultural background of the researchers, and that different results may be obtained if a different cultural vantage point is taken in the design of these studies. Two approaches may be adopted to design a culturally balanced study, in which no single culture will dominate the research questions explored and bias the results obtained. First, a *decentered* approach can be adopted, in which a culturally diverse perspective is taken in the conceptualization and design of a study. For instance, when Schwartz (1992) tested his pan-cultural model of value structure, he encouraged researchers from different cultures to add culture-specific value items to his pan-cultural set. Smith and Peterson (1988) have taken into account the influence of culture in their formulation of a theory of leadership behavior and their empirical test of the theory (Peterson et al., 1995).

The second approach is the *convergence* approach. The basic idea is to design a study that is as culturally distant as possible from existing studies and to see if the results obtained overlap with existing results. If the new results overlap with existing results, it can be concluded that the cultural origin of existing studies have not biased the results obtained. If different results are obtained, however, the possibility that the cultural origin of existing studies has biased the results must be further investigated. The best examples to illustrate this approach are provided by Bond and his colleagues. The Chinese Culture Connection (1987) designed a value survey based entirely on Chinese values and administered it in 22 countries. It was found that three factors showed overlap with factors identified by Hofstede (1980), whose results were based on a Western instrument. A new factor emerged, termed Confucian work dynamism, which correlated highly with economic growth. In the realm of person perception, Yang and Bond (1990) administered a set of emic Chinese descriptors together with a set of imported American descriptors to a group of Taiwanese subjects. Of the five Chinese factors identified, only four were adequately explained by the American factors, and one factor was uniquely Chinese.

Data Analysis

In this section we will first describe bias, followed by a description of psychometric techniques to detect differential item functioning as a special case of bias. In the last part of the section we will describe the most common statistical techniques for analyzing cross-cultural data sets.

Preliminary Analyses

Prior to the data analysis that addresses the central research question or hypothesis, preliminary analyses will often be required. If a psychological instrument is used, its psychometric properties should be established, in particular its reliability. In most cross-cultural studies this seems to be routine practice. It is surprising that tests of intergroup differences in reliability are almost never carried out even though the observation of dissimilar reliability coefficients can provide valuable

clues about measurement accuracy and hence, the appropriateness of an instrument for cross-cultural comparison. Procedures to test the equality of independent alpha coefficients have been described by Kraemer (1981) and Hakstian and Whalen (1976).

The interpretation of intergroup differences can be seen as an attribution process. Two kinds of attributions can be envisaged. Observed intergroup differences may be valid, and members of group A have on average more of a particular propensity such as anxiety, intelligence, or collectivism than members of group B. The observed differences may also be due to bias (measurement problems). For instance, the items used may be affected by intergroup differences in stimulus familiarity or social desirability, which have produced the cultural differences observed.

A distinction can be made between three types of bias. The first is called *construct bias*. This kind of bias occurs when the psychological construct is not identical across cultural groups. Construct bias implies that the theoretical construct is not or is inadequately represented in the instrument. In Embretson's (1983) terms, construct bias refers to a poor construct representation. An example can be found in the area of intelligence. Everyday conceptions of intelligence, mainly in non-Western cultures, have been found to differ from the conception underlying intelligence tests (Serpell, 1993; Sternberg, 1985; Super, 1983). Everyday conceptions of intelligence tend to be broader than scientific theories. In addition to reasoning and factual knowledge that are shared in both conceptions, "social intelligence" is also included in everyday conceptions. "Social intelligence" involves social skills, obedience, and knowing one's role in the family, class, and peer group. A Western intelligence test will therefore show construct bias in many non-Western contexts. Culture-bound syndromes, such as amok, that are studied in ethnopsychiatry provide another example (Draguns, 1989; Harkness & Super, 1990). In the area of personality the Chinese concept of "filial piety" can be mentioned; filial piety refers to taking care of one's parents, conforming to their requests, and treating them well. The Chinese concept is much broader than the Western concept of being a good son or daughter (Ho, in press). A direct comparison of these two will result in construct bias.

It was argued before that a cross-cultural researcher may choose to apply or adapt an existing instrument, or assemble a new one. In the terminology of this section, the decision should be based on whether construct bias is present in the instrument. The assessment of construct bias should be based on knowledge about the cultural groups. If an instrument has been applied in several cultural groups with the same instrument and no additional data are available, statistical tests alone will not lead to a full understanding of the nature of the construct bias present. A proper assessment of construct bias should be based on research conducted in each cultural group, exploring whether the implicit definitions of the concept of the test are consistent across the cultural groups. Examples of this approach can be found in the work of Serpell (1993), Sternberg (1985), and Super (1983).

The second kind of bias is called *method bias*. If method bias occurs, the psychological construct is well represented by the instrument but the assessment procedure introduces unwanted intergroup differences. Empirical studies that reveal method bias are Deregowski and Serpell's (1971) sorting task of miniature models and pictures of animals and motor vehicles, described earlier and Serpell's (1979) study of pattern copying using a paper-and-pencil format and iron-wire models.

Method bias can be examined by monotrait–multimethod matrices or triangulation. In this approach, a psychological construct is investigated using a systematic variation of methods. If the cross-cultural differences observed are similar across methods, method bias is unlikely. Method bias is said to occur if the intergroup differences vary across the methods. An analysis of covariance structures is often used in this situation, as will be illustrated later on.

A specific way to study method bias involves the repeated administration of the same instrument. Test–retest studies of cognitive tests have often shown score increases that are larger in non-Western groups than in Western groups (Kendall, Verster, & Von Mollendorf, 1988; Van de Vijver, Daal, & Van Zonneveld, 1986). A significant improvement in one group at the second occasion, or a gain pattern that is differential across groups, undermines the validity of the first test administration.

The third kind of bias is the most investigated. It was originally called *item bias* and is now better known as *differential item functioning*. Whereas construct bias and method bias involve the appropriateness of the whole instrument, differential item functioning occurs at the item level. Item bias refers to anomalies in the instrument at the item level caused by poor translation or inappropriate items in a particular context. A widely accepted definition of differential functioning has been proposed in the area of ability testing. An item is said to show item bias if persons from different cultural groups with an equal ability do not have the same probability of giving a correct answer. Individuals with an equal ability or attitude from different cultural groups should, apart from chance fluctuations, show the same average score for items of an unbiased instrument. From a psychometric point of view, the assessment of this kind of bias is best developed. A multitude of psychometric techniques have been proposed to test the presence of item bias. We will not describe them in detail. Rather, we shall briefly describe and illustrate two of them, followed by the presentation of a taxonomy of the techniques.

Historically speaking, analysis of variance was probably the first technique that has been applied to study differential item functioning (Cleary & Hilton, 1968). We shall discuss here a slightly modified procedure. Suppose that a test for authoritarianism of 30 five-point Likert-scale items has been administered in two cultural groups of 200 persons each. If we are interested in the presence of differential item functioning, the first step is to divide the subjects into score level groups. Individuals with an equal score are assumed to have an equal level of authoritarianism, and subjects with the same score are grouped together. Because the scores on the Likert scale range from one to five, the total score can vary from a minimum of 30 to a maximum of 150. The split of the score distribution into score levels should be based on the score of all cultural groups together; the same

cutoff scores should be applied to all cultural groups. Theoretically speaking, there can be 121 score level groups in this case (from 30 to 150, including both ends). In practice, a much smaller number will be used as the number of subjects will be unevenly distributed across the score levels (Clauser, Mazor, & Hambleton, 1994). Quite often, an attempt is made to choose the cutoff scores in such a way that the number of subjects in each group is approximately the same. Score level will be one of the independent variables in our data analysis; the other one will be the cultural group. Differential item functioning is tested in a set of analyses of variance, one per item, with culture and score level as independent variables and the item score as dependent variable.

Following Mellenbergh (1982), we shall make a distinction between two types of item bias: uniform and nonuniform. Figure 7–1 presents the curves which depict the average score of two groups on a particular item, technically called empirical item characteristic curves (Allen & Yen, 1979). When the curves more or less coincide, there is no bias (Figure 7–1a). When the curves are more or less parallel without coinciding, there is uniform bias (Figure 7–1b). When the curves are not parallel, the items are said to show a nonuniform bias (Figure 7–1c). In this case, the difference in the average test score will depend on the score level. For instance, for low authoritarian subjects, the item is endorsed more strongly in one culture, while for high authoritarian subjects, the item is endorsed more strongly in the other culture. A combination of both types of bias is presented in Figure 7–1d. In terms of the analysis of variance, an item is said to be uniformly biased when the main effect of culture is significant. In this case subjects from one cultural group have a consistently higher score than individuals with the same underlying propensity from another cultural group. A significant interaction of level and culture indicates the presence of nonuniform bias.

Item bias analyses can be carried out in an iterative or a noniterative way. In the latter case the analyses of variance are carried out for all items and the presumably biased items (i.e., all items with a significant main effect for culture and/or a significant interaction between culture and level) are removed simultaneously. Intergroup score comparisons are carried out on the reduced item set. In an iterative procedure the elimination proceeds on an item-by-item basis. In the first step, all items are considered. The item with the largest bias component (i.e., the smallest probability in the computer output) is then removed if the component is significant. The whole procedure is then repeated for the reduced set of items until no more bias components are significant. An attractive feature of iterative procedures is that the total score is updated in each iterative step, which allows for a finer detection of bias. It might well be that after the removal of a few items the meaning of the total score changes somewhat and this change can result in the removal of different items than in the case of a noniterative procedure. However, iterative procedures are cumbersome because after the removal of an item new cutoff scores for the score levels have to be calculated.

The removal of biased items does not inevitably lead to the elimination of intergroup differences in the average scores (Poortinga & Van der Flier, 1988). Items can be biased or unbiased, irrespective of the presence (or absence) of inter-

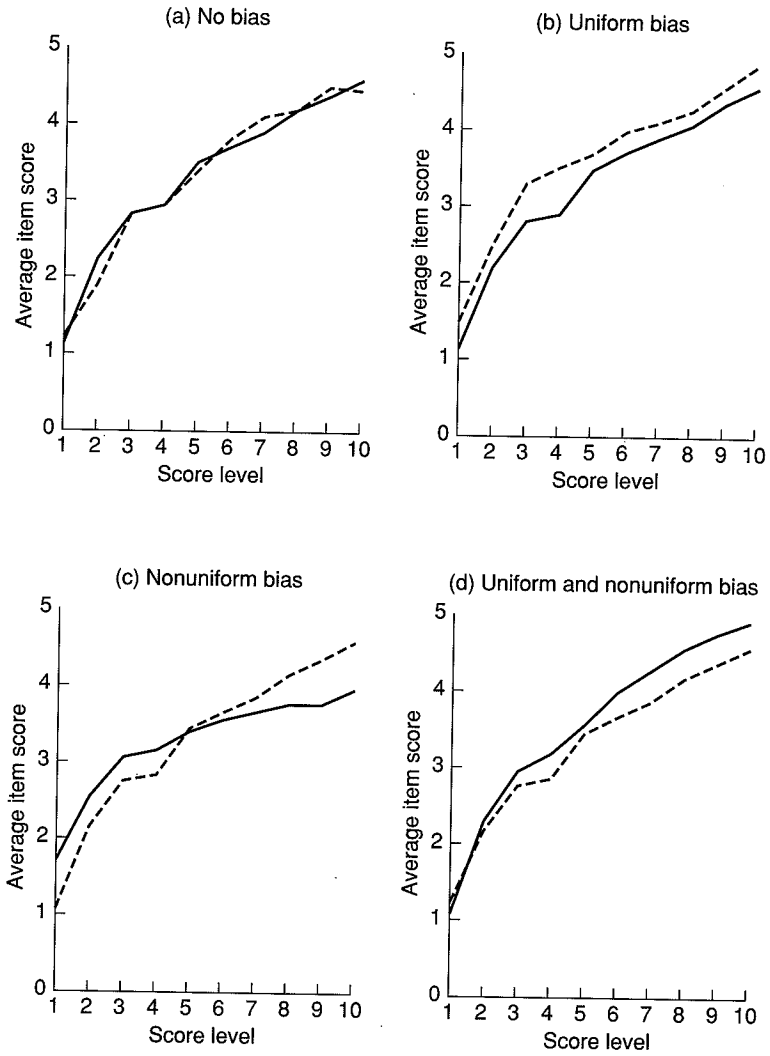

FIGURE 7–1 The average performance of the two cultural groups on an item that shows (a) no bias (b) uniform bias (c) non-uniform bias, and (d) both uniform and nonuniform bias (hypothetical example).

group differences. After all, item bias analysis does not test whether there are overall intergroup differences in total score, that is, whether individuals of group A would have a higher propensity on X than individuals of group B. Rather, item bias analyses test whether there are intergroup differences per score level, that is, whether individuals from group A with a particular attitude level have the same average score on a particular item as people from group B with the same attitude level. The item bias analysis uses an analysis of variance with level and cultural group as independent variables and a particular item score as dependent variable. In contrast, an analysis of variance testing the presence of overall intergroup differences treats culture as the independent variable and the item or total test score as the dependent variable.

The most popular technique to test differential item functioning today is the Mantel–Haenszel statistic (Holland & Thayer, 1988). The statistic is closely related to item response theory (e.g., Hambleton & Swaminathan, 1985). More specifically, the Mantel–Haenszel statistic tests whether a single Rasch model, a model from item response theory, fits the data in each group. The rationale behind the Mantel–Haenszel statistic and the analysis of variance approach explained earlier is similar. The major difference is that the Mantel–Haenszel procedure works with dichotomous data whereas the analysis of variance is based on interval data.

Item response theory represents a more general approach for assessing differential item functioning (e.g., Hambleton & Swaminathan, 1985; Hulin, 1987). This model assumes that an unbiased item evokes a similar response from respondents that are similar in their standing on a latent trait regardless of their cultural backgrounds. In the general form, this model links item responses to latent traits by means of a logistic curve specified by three parameters. The first parameter is concerned with the discrimination capability of the item; the second parameter is concerned with the difficulty level of the item; and the third is concerned with the extent to which guessing is involved in responding to the item. In specific applications, two-parameter models, which exclude the guessing parameter, are often employed for modelling attitudinal data. To detect biased items, item characteristic curves, which relate the probability of making a certain response to standing on a latent trait, are examined. Items are equivalent across two cultures if their item characteristic curves are similar across these cultures. Differential item functioning is present when parameters differ significantly across cultural groups. Item response theory has been applied in cross-cultural research on self-concept (Leung & Drasgow, 1985), job satisfaction (Candell & Hulin, 1987), intelligence (Ellis, 1989; Van de Vijver, 1988), and attitudes toward mental health (Ellis & Kimmel, 1992).

The standard procedure for the application of item response theory is as follows:

1. Item response theory assumes that a scale is unidimensional, and the unidimensionality of the scale must be checked. If the scale is multidimensional, each unidimensional subscale must be examined separately.
2. An item response theory model with the appropriate number of parameters is selected to fit the data in each culture.

3. The parameters identified for each cultural group are equated on the same metric through an iterative linking procedure.
4. Biased items are detected and eliminated with the aid of item characteristic curves and a chi-square test. The parameters are equated again with the linking procedure with unbiased items only, and this procedure stops when no biased items are detected.
5. The biased items identified are eliminated from the scale before cross-cultural comparisons are made.

Item response theory has characteristics that make it appropriate for cross-cultural applications. First, the estimates of item parameters do not depend on the propensity level of the group studied. This is not the case in classical test theory in which the difficulty of an item, operationalized as item average, depends on the average ability level of the group. Similarly, the estimates of person parameters in item response theory are independent of the items of the instrument. Second, most models in item response theory allow for a fit test. The extent to which the empirical data can be taken to obey the theoretical model can be examined (e.g., Hambleton & Swaminathan, 1985; Lord, 1980; Van den Wollenberg, 1988).

The most important limitations of item response theory are twofold. The applicability of item response models may be reduced by the strict assumptions that have to be met, particularly in the Rasch model. Furthermore, large sample sizes are required to obtain stable estimates, particularly in the three-parameter model.

TABLE 7–1 Schematic overview of differential item functioning techniques (after Van de Vijver, 1994)

	Model equation	
Sampling distribution	Linear	Nonlinear
Unconditional procedures		
Unknown	Partial correlation index (Stricker, 1982)	Delta plots (Angoff, 1982)
Known	Analysis of variance (Cleary & Hilton, 1968)	
Conditional procedures		
Unknown	Standardized *p*-difference (Dorans & Kulick, 1986)	Item response theory (McCauley & Mendoza, 1985)
Known	Analysis of variance with score level as one of the variables	Mantel-Haenszel procedure (Holland & Thayer, 1988)

Three questions are relevant in the choice of a particular item bias statistic, First, what kind of measurement model should be used? Some techniques are based on a linear model such as an analysis of variance, while others, such as the Mantel–Haenszel statistic, are based on a nonlinear model. In general, interval-level data tend to be analyzed using linear models, while dichotomous data are often analyzed by item response theory, a nonlinear model. Second, is the technique conditional or unconditional? Most modern techniques are so-called conditional procedures. These techniques compare the scores of individuals across cultural groups per score level. Both of the previous examples are conditional. Until the eighties, unconditional procedures were more common, such as the comparison of item averages. It has been shown several times (e.g., Lord, 1977, 1980) that unconditional procedures can underestimate the number of biased items. Therefore, conditional procedures are to be preferred. The third question refers to the sampling distribution of the item bias statistic. In both our examples, the sampling distributions are known. This allows for a statistically rigorous test of the null hypothesis of no bias. Yet, various bias statistics that have been proposed have unknown sampling distributions, which makes a statistical evaluation of item bias questionable, whatever the intuitive appeal of the statistic (e.g., Stricker's, 1982, partial correlation index). A taxonomy of basis statistics on the basis of these three questions is presented in Table 7–1.

A perusal of the cross-cultural literature shows that differential item functioning techniques are infrequently applied in cross-cultural psychology. We find this disappointing; in many cases it should be standard practice to carry out an item bias analysis prior to the actual data analysis. Item bias techniques have been mostly applied to cognitive test scores, and much less so in the area of personality and social psychology. There is no good reason for the uneven distribution of the application of item bias techniques, unless one would want to maintain that items in personality questionnaires are of a much higher quality and less open to bias than are cognitive test items.

Two general findings emerge from the application of differential item functioning techniques in cross-cultural psychology. First, item bias may be psychometrically well defined and operationalized, but it may be difficult to grasp its psychological meaning. In current applications, it is not uncommon to find that item bias is reported but no sensible explanation can be provided for the bias (Scheuneman, 1987; Van de Vijver, 1994). Furthermore, item bias indices are not stable in cross-validation studies. Retests with the same instrument may show other items to be biased. The common difficulties encountered in empirical applications of item bias techniques, such as inadequate stability and interpretability, may reduce the attractiveness of these procedures. Still, if we start to routinely apply item bias techniques to cross-cultural data, we may build up a body of knowledge about item quality from a cross-cultural perspective.

Second, some item bias studies have shown a substantial proportion of items to be biased, sometimes more than half of the items. In such a case the item bias analysis seems to point to a serious lack of validity of the instrument. A prudent approach would then be to refrain from intergroup comparisons.

Establishing Scalar Equivalence

Techniques based on correlations such as factor analysis have been proposed and used to test scalar equivalence. For instance, Eysenck and his coworkers concluded that scalar equivalence can be assumed when the factor structures obtained with a measurement instrument in various cultural groups are similar. A similar argument has been put forward by Berry (1980). However, as argued before, similarity of correlations matrices or factor structures across cultural groups can only demonstrate structural equivalence, and does not speak to scalar equivalence. Structural equivalence imposes fewer restrictions on the data than scalar equivalence.

At least three approaches have been proposed to establish full score comparability in the literature. First, various authors assume but do not test full score comparability. If a test is administered in two cultural groups and the test scores are compared without any concern for comparability, full score comparability is implicitly assumed. An example comes from the literature on culture-free and culture-fair intelligence testing. Reports involving these somewhat obsolete instruments hardly involve statistical tests of full score comparability (e.g., Anastasi, 1976; Cattell, 1940; Cattell & Cattell, 1963). In our view, researchers should attempt to provide evidence for full score comparability of their instruments.

The second and third approaches are internal validation procedures. The procedures are called internal because the data used to validate equivalence are derived from the instrument itself. The second approach involves intra-cultural techniques in which empirical data are compared to theoretical expectations for each culture. It is possible to formulate hypotheses about the order of difficulty or endorsement rate of items in some instruments. For instance, items of tests of arithmetic abilities can often be ordered by the complexity of the arithmetical operation required. Operations requiring the manipulation of one-digit numbers will be easier than operations requiring two-digit numbers; additions and subtractions will be easier than multiplications and divisions. Strong evidence against scalar equivalence is obtained if theoretical expectations are not borne out. As a second example, the use of fit tests in applications of item response theory can be mentioned (e.g., Hambleton & Swaminathan, 1985; Lord, 1980; Van den Wollenberg, 1988). A good fit within each group provides initial evidence for scalar equivalence. Intracultural validation techniques provide necessary though insufficient evidence for the presence of scalar equivalence.

The third approach can be called *cross-cultural validation.* The best known example is the work on item bias, or differential item functioning (Berk, 1982; Holland & Wainer, 1993). Various psychometric techniques have been developed which scrutinize consequences of the lack of bias at the item level (cf. the description of item bias before).

A special case of the monotrait–multimethod approach, described earlier for the examination of method bias, is the use of multiple measures to capture the same construct. Triangulation, as this procedure is often called, can provide some insight in scalar equivalence, especially when the statistical techniques described in the previous section do not apply, such as in the case of single- item measures

(e.g., measures in Piagetian psychology, social behavior). Triangulation amounts to utilizing multiple measures, as diverse as possible, to measure the construct. If convergent results are obtained with different measures, bias is not likely to have produced the results. For instance, Hess, Chang, and McDevitt (1987) found that in comparison with American mothers, Chinese mothers were more likely to attribute the academic performance of their children to effort. Consistent with this result, Chinese children were also more likely to attribute their academic performance to their own effort than were American children. The convergence between the children and mothers has strengthened the validity of the cultural difference observed. In contrast, Serpell's (1979) study of Zambian and Scottish children's copying skills using iron-wire models and pencil-drawing is an example of nonconvergent operations. It should be pointed out that although multiple measures can assess the confounding influence of bias, it does not guarantee scalar equivalence even when convergence is obtained. The equality of the origin and the unit of the measurement scale is not directly assessed in triangulation.

Statistical Tests of Cross-Cultural Differences: Introduction

The statistical techniques described in the previous section examine the cross-cultural applicability of research instruments and the validity of the use of these instruments in cross-cultural comparisons. In this section, we will describe statistical tests that are applied after the adequacy of the psychometric characteristics and the absence of bias have been established. A distinction between structure- and level-oriented techniques will be made in our presentation. Because of space limitation, we will only provide a brief overview of the statistical techniques.

Prior to any statistical analyses, it should be decided whether the data need to be standardized, and if so, which standardization procedure is to be used (e.g., Hofstede, 1980; Leung & Bond, 1989). Culture-level analyses can yield strikingly dissimilar results for standardized and nonstandardized data sets. Standardization is usually defined as the computation of z scores ($z = (X - M)/S$, in which X is the score to be standardized, M is the mean and S is the standard deviation). Standardization is defined here more generally and refers both to z scores and to transformations to other deviance scores such as X/S and $X - M$. The aim of standardization is the reduction or elimination of unwanted intergroup differences such as those due to response sets. If scores are standardized per cultural group, intergroup differences in means, standard deviations, or both are eliminated. Such a procedure requires justification, because intergroup differences in average scores may not be exclusively due to response sets or other unwanted sources but may reflect valid differences. The justification is usually based on the presumed equality of averages across cultures. For instance, Schwartz (1992), who has transformed raw scores to deviations from the mean in his value survey, argues that the average importance score that people give to all the value items in his inventory should be similar across individuals, because his instrument represents a comprehensive set of human values. If such a reasoning cannot be justified, analyses based on the original as well as the standardized data should be conducted and the results obtained compared.

Statistical Tests of Cross-Cultural Differences: Level-Oriented Techniques

The most frequently reported statistical tests of *level* differences are the *t* test and analysis of variance (e.g., Glass & Hopkins, 1984; Hays, 1994). The most commonly tested null hypothesis specifies that there are no intergroup differences. In a *t* test, the cultural group is the independent variable and the score on a psychological instrument is the dependent variable. The popularity of the *t* test, in cross-cultural psychology as well as elsewhere, is undoubtedly attributed to its simplicity, availability (in computer packages), and robustness against violations of assumptions. The same holds for the analysis of variance, which is carried out when data of more than two cultural groups are studied. The major interest tends to be in the main effect for culture, which, assuming that the effect is significant, indicates that at least one culture has an average on the dependent variable different from the other cultures. More complex designs, so-called factorial designs, are often reported in cross-cultural research. These are designs in which, in addition to culture, one or more independent variables, such as gender or age, are included. The inclusion of such additional variables, say gender, is particularly relevant when the reaction patterns of men and women are expected to differ across the cultures studied (e.g., the male–female differences on the dependent variable are more pronounced in one culture). These differences in reaction patterns will come out in an analysis of variance as a significant interaction of culture and gender.

Regression analysis is often used in level-oriented studies. Regression analysis evaluates the influence of one or more independent variables on a dependent variable in terms of the amount of variance in the dependent variable that the independent variables can explain. Regression coefficients express the degree of relationship between the independent and the dependent variables. The squared multiple correlation, another relevant statistic of the regression analysis, is the amount of variance explained by the independent variables, which gives an overall evaluation of the success of the independent variables in predicting variation in the dependent variable. In cross-cultural studies, level-oriented hypotheses involve a test of whether the intercept of a regression equation is similar across different cultural groups (Cohen & Cohen, 1983; Pedhazur, 1982).

Regression analysis can be carried out on raw or standardized scores (mean of zero and unit variance). Standardization affects the size of the coefficients, but leaves the significance level unaffected. In practice it has become more common to report standardized regression coefficients because they are independent of the measurement units of the independent variables.

The choice between an analysis of variance (or *t* test or *z* test) and a regression analysis mainly depends on the measurement level of the independent variables. Nominal and ordinal data are often analyzed in an analysis of variance