Testing across cultures

van de Vijver, F.J.R.; Poortinga, Y.H.

Published in:
Advances in educational and psychological testing

Publication date:
1991

Advances in Educational and Psychological Testing: Theory and Applications

edited by
Ronald K. Hambleton
Jac N. Zaal

1991

Kluwer Academic Publishers
Boston/London/Dordrecht

10 TESTING ACROSS CULTURES

Fons J. R. van de Vijver
Ype H. Poortinga

There has been a longstanding, scientific interest in the comparison of people belonging to different cultural groups. In the course of the history of Western science, practitioners of different disciplines have been involved. During the Renaissance the equality of races was an issue for theologians. In 1550 a number of them convened at the court of Charles V in Spain to solve the question of how the American Indians could be colonized "in a Christian fashion." According to the chronicles, the debate focused on the question of whether the Indians formed an inferior race in comparison with their Spanish colonizers. The issue was never settled, even though "some of the most learned and powerful men of the age" participated (Boorstin, 1985, p. 633). During the nineteenth century, racial differences had become the domain of social philosophers, who, in turn, "passed the buck" (the use of the expression in this context coming from Mann, 1940) to psychologists.

Each scientific discipline formulated somewhat different questions: the theologians were concerned with moral equality and inequality, social philosophers studied cultural evolution, and psychologists concentrated on individual behavior. The lack of agreement already present among the sixteenth century theologians continues to exist today, in psychology notably with respect to cognitive abilities. Authors such as Jensen (1980)

and Eysenck (1984) are proponents of the view that marked differences exist in cognitive abilities among individuals of various cultural groups, while others like Mercer (e.g., 1984) defend the opposite opinion. By far the most cross-cultural studies are in line with the latter position.

In these studies, ecological variables, such as climate or sociocultural variables, are postulated as the determinants of observed differences. The plasticity of human behavior is emphasized, and it is often assumed that, through formal education and technological development, intergroup differences in cognitive abilities will gradually disappear. However, it should be emphasized that most of the opinions on the nature of cultural antecedents of observed test score differences are, at least to some extent, speculative in view of the serious methodological difficulties which often arise if we want to identify the specific determinants of an observed intergroup difference. On the other hand, the occurrence of often ill-founded speculations is not restricted to "environmentalists" such as Mercer; much work which is more in line with a "geneticist" position suffers from the same problem.

An early attempt at a systematic investigation of the cognitive abilities of individuals in non-Western cultures can be found in the work of Porteus (1917), who composed the so-called Maze Test, an instrument similar to the Mazes subtest in the Wechsler Adult Intelligence Scale and Wechsler Intelligence Scale for Children batteries of today. Porteus' Maze Test has been used extensively in cross-cultural research (for a review see Porteus, 1965). David (1974) cites a number of features of this test which are meant to optimize its suitability in a cross-cultural context, namely "a high intrinsic interest for most persons, simple instructions, easy to administer and objectivity of scoring" (p. 11). Numerous studies with this test have revealed large cross-cultural differences in mean score levels. However, the interpretation of these differences is far less clear than their replicability might suggest. All kinds of factors can threaten a straightforward interpretation of observed cross-cultural score differences. Among other things, they can be caused by differences in the familiarity of subjects with testing situations, the nature of the stimulus materials, or differences in motivation. This is true not only for Porteus' test but for all assessment instruments. The question of how to arrive at more valid explanations of observed intergroup differences is the central problem of this chapter. In the final section a procedure will be outlined which is aimed at reducing the number of rival hypotheses that can explain observed intergroup differences. In this procedure it is crucial that potential determinants of intergroup differences are recognized beforehand and that variables to assess these determinants are included in the design of a study.

A second problem of cross-cultural testing has to do with the

administration of tests. In this context it is illuminating to look at the difficulties that have emerged in the application of Porteus' Maze Test. Porteus himself (1965), for instance, found it difficult to persuade Australian aboriginal subjects to solve the items by their own effort rather than in cooperation with the tester. As another example, it can be mentioned that the Maze Test, which is a paper-and-pencil test, has been applied among groups from which the members had never touched a pencil before (cf. Porteus, 1965). In the case of some cultural groups it is even debatable whether mazes are suitable as stimulus material. In a discussion on the use of the Maze Test among Bushmen, Reuning and Wortley (1973) argue that "the idea of a maze is not likely to occur to a Kalahari-dweller (like the Bushman) and must be utterly foreign to him" (p. 61). Their argument is based on the consideration that in a savannah, the natural ecology of the Bushmen, a person can invariably go along a more or less straight line from one point to another. The obvious conclusion from these examples is that the validity of the results obtained with a test will be questionable in all instances where the administration raises the kind of difficulties referred to.

The third problem of cross-cultural testing to be discussed here is that numerically identical test scores can have a psychologically different meaning. If scores are numerically comparable across cultures, they will be called *score equivalent.* However, such equivalence should be established instead of assumed. Test scores obtained in different cultural groups can have a quite distinct psychological meaning. Porteus' (1965) observation that Australian Aborigines perform significantly better than Kalahari Bushmen does not tell much about differences in planning ability between these groups. Rather, the low scores are likely to reflect the use of materials with highly unequal ecological validities across the groups and misunderstandings in the administration through the use of an interpreter, as was done among the Bushmen.

The three problems of cross-cultural testing mentioned here—the explanation of observed intergroup differences, proper test administration, and score equivalence—are interrelated. An adequate test administration procedure is a necessary, though insufficient, condition for score equivalence across groups. Score equivalence, in turn, is a necessary condition for an adequate explanation of intergroup differences.

The difficulties of cross-cultural testing have been emphasized here because their impact is greatly underrated, in our opinion. The applicability of tests in settings which are culturally widely discrepant from the usually Western context in which they are constructed is too often taken for granted. All kinds of factors may render intergroup differences invalid. Differences in formal education, unfamiliarity with tests, or even a poor

nutritional state and poor general health, to mention only a few relevant factors, can form a threat to the equivalence of scores.

The underrating of testing problems in cross-cultural research is a reason for the wide gap sometimes found between psychological test data and daily observations of related phenomena. Work on (Piagetian) formal operational thinking can illustrate this point. In a review of research in this area, Neimark (1975) states that there is "clear evidence of retardation of development and even failure of attainment in most non-Western groups" (p. 578). This means that the psychological data seem to imply that many individuals in non-Western groups are incapable of abstract reasoning. It is obvious that this statement refers to the testing situation rather than to daily life, in which the same people are definitely not incapable of abstract reasoning (cf. Biesheuvel, 1949; Hutchins, 1980). It appears that for some groups the assessment procedures reviewed by Neimark have a very low generalizability to daily life.

Problems in Test Use and Administration

The proper use of tests starts with administrative procedures that are suitable to represent the psychological phenomena under study. In the introduction a few examples have been given of what can go wrong when tests are applied in a different cultural setting. In this section a broader overview is presented of the possible sources of error in test administration procedures, followed by some precautions which can be taken to reduce the effects of these errors.

Five areas are distinguished: problems related to the tester, the examinees, the interaction between tester and examinee, the response procedure, and the stimulus material.

Tester

The (obtrusive) presence of the tester during the data gathering can be a threat to the validity of the results. It is recognized that in observational studies of mother-child interactions the mere presence of the tester may provoke or inhibit particular behavior of the mother and the child (Super, 1981). The potential effect of the race of the tester on the performance of the examinee has been extensively studied in the United States, with black and white testers for both white and black subjects. The results are not very consistent, but the effects tend to be small (Jensen, 1980; Vernon,

1979). However, this conclusion cannot be generalized to other cultural settings without additional evidence.

Examinees

The second problem area involves the choice of examinees. It is a major difficulty in cross-cultural psychology to select corresponding samples of subjects across cultures (Pick, 1981; Malpass & Poortinga, 1986). Cultures differ in many ways, and hence samples recruited from these cultures will also differ in many respects. This violates the condition that samples of subjects should only differ on one variable, namely the independent variable. This condition, which is the cornerstone of experimental psychology, does not hold in cross-cultural psychology where intact groups are compared. Usually, particular cultures are selected because they are assumed to vary in terms of some background characteristic which is relevant for the construct under study. However, the researcher should be alert to the existence of other background variables—unintentionally varied through the particular choice of cultures—which can also legitimately explain intergroup differences in performance. The most clear-cut examples are studies in which the test scores of illiterates and literates are compared. Two such groups differ not only in ability to read and write but in a host of variables related to formal schooling. Comparisons between literates and illiterates are almost by definition comparisons between schoolgoing and non-schoolgoing populations. A noteworthy exception is the work of Scribner and Cole (1981) with the Vai in Liberia. Among the Vai different forms of literacy are found. Some of these people are literate in their indigeneous syllabic script which is learned in an informal setting, labeled "unschooled literacy" by the authors. By a careful choice of subjects Scribner and Cole were able to disentangle the traditionally confounded effects of schooling and literacy.

Tester-Examinee Interaction

The third problem area has to do with the interaction between tester and examinee. Establishing ways of adequate, unambiguous communication between tester and examinee is an essential condition for meaningful test use. When Reuning and Wortley (1973) planned to administer a variety of tests to the Bushmen, they were confronted with the problem of many locally different vernaculars and with the inherent difficulty of recruiting

competent interpreters for each new linguistic group to be tested. Therefore, in their choice and composition of tests they tried to minimize the dependence on verbal exchange both in the instructions and in the examinees' responses. According to these authors, instructions should be understandable without any verbal explanation. Items should invite the intended action, they should have what in German is called *Aufforderungscharakter*, i.e., incite the subject to do what is required. Also, responses should be concrete actions rather than verbal explanations (Reuning & Wortley, 1973, p. 12).

Minimal reliance on verbal communication circumvents only some of the difficulties. It is no solution for the absence in indigenous languages of particular words which are essential for a good understanding of a task. When reading Lancy's (1983) classification of indigenous counting systems among the Papuas of New Guinea,which vary considerably in their degree of complexity, it is easy to see that tests in which arithmetic reasoning plays a role may be hard to understand for certain groups, because the necessary number concepts are lacking in their language.

Sometimes it may seem possible to circumvent language problems by using the official, national language (e.g., French in Zaire) which often is also the official medium of instruction at school. However, for many subjects this national language will be their second or third language, and it is unrealistic to expect an equal proficiency in the national language as in the native tongue.

All illustration of a subtle but important communication failure in cross-cultural testing is offered in two studies among the Wolof on the Piagetian principle of conservation. Greenfield (1966, 1979) administered conservation tasks to Wolof subjects in their native language using the clinical interview method commonly found in the Piagetian tradition. In such a test, two identical short, broad beakers containing an equal amount of water are placed on a table in front of the subject. One of these beakers is poured into a tall, thin beaker. The subject is then asked which beaker contains more water. Children of preschool age—most of them "non-conservers"—typically will say that the tall glass contains more, while older children—frequently "conservers"—will give the correct answer. Greenfield's results indicated that among unschooled Wolof, nonconservation responses were found frequently, even at 12 years of age, especially with a task in which the water was distributed over more than two beakers. In an interesting replication Irvine (1978) argued that the question of which beaker held more water appeared to be ambiguous in the language of the Wolof, as "more" could refer both to the quantity and the level of the water. With this in mind, Irvine found that all subjects she tested under-

stood the principle of conservation, although admittedly, her case is weakened by the fact that her sample included only five subjects.

Response Procedures

Response procedures are the fourth topic of discussion. We have mentioned already the use of paper-and-pencil tests among groups who have never touched a pencil before. Another example is the use of a multiple-choice format which presupposes a balanced strategy between solving a problem until one is perfectly sure and a liberal amount of guessing the correct alternative.

A further example can be found in the work of Serpell (1979). He administered a pattern design copying task to children in the United Kingdom and Zambia. Two different media were used to assess the child's skill in copying, namely pencil-drawing and iron-wire modeling, a popular pastime among boys in Zambia. The British children outperformed their Zambian counterparts in pencil-drawing, while the reverse was found for the wire-modeling task. It appears that the response medium can affect the scores to a substantial extent. It is highly unlikely that groups unfamiliar with a particular response procedure, be it iron-wire modeling, multiple-choice format, or whatever, will attain the highest level of performance with that medium.

Stimuli

The final topic to be treated, problems connected with the stimulus material, is the best documented. A factor mentioned over and over in the literature as an important determinant of intergroup differences is the differential familiarity of subjects with certain stimulus materials (e.g., Biesheuvel, 1949; Irvine & Carroll, 1980; Ord, 1970; Pick, 1981; Schwarz, 1961). An elegant demonstration of the effect of stimulus unfamiliarity is offered by Deregowski and Serpell (1971). Scottish and Zambian children were asked to sort miniature models of animals and motor vehicles in one experimental condition and photographs of these items in another condition. With the actual models no intergroup differences were found, whereas in the sorting of the photographs Scottish children obtained higher scores than Zambian children. This can be explained in terms of a lower familiarity of the Zambian children with photographs. Similarly, Price-Williams (1962) found that children in a rural Nigerian community

displayed a higher ability in sorting indigenous leaves than in sorting toy models of animals.

Skill Reduction

Rather than pursuing any complete coverage of the extensive literature, we shall focus on some principles of cross-cultural test use, meant to minimize the impact of the previously mentioned problems. Van der Flier (1972, 1980) has formulated a so-called "skill reduction" approach. In this approach it is assumed that the completion of a test requires a number of "skills" from the subject. These skills can be defined as the set of abilities which are needed to perform well on the test, in addition to the construct the test is supposed to measure. The skill to recognize pictures in Deregowski and Serpell's experiment on the sorting of photographs of toys is an example. According to Van der Flier's rationale, score differences in cross-cultural research are caused not only by genuine ability differences but also by skill differences.

Van der Flier has distinguished three ways to reduce unwanted effects of skills. First, he suggests to restrict the comparison of scores to those parts of a population where the skills needed are readily available. As a check on the proper understanding of a multiple-choice response format, a researcher can administer a few extremely simple multiple-choice items, preceding the actual test items. (These simple items may even be unrelated in content to the test.) In the data analysis only those subjects with correct answers on the first simple items will be considered. Second, the researcher can try to eliminate the need for certain skills, for example, by an appropriate choice of stimulus materials. The abovementioned experiment of Price-Williams (1962), in which Nigerian subjects were asked to sort leaves found in the natural environment, is a good example of this strategy to reduce the effects of skills. Third, skill differences can be reduced by extending the test instruction and by administering training items. A classic example is the repeated administration of Raven's Matrices in Congo by Ombrédane and associates (1956), showing that the validity increased from the first to the third time. Van de Vijver (1984, 1988) included a lengthy instruction procedure with a sample item for each of the problem-solving rules which the subject needed for an inductive reasoning test. In this way the domain of responses required by the test was explicitly defined.

Similarly, Van de Vijver and colleagues (1986) gave training to Dutch, Surinam and Zambian youngsters on a test of inductive reasoning. In the test the subject had to mark the group of letters which did not fit in a

group of five: for example, DDDGFH NHDTTT KLMMMB WWSXZA HHHRDS. When compared with the Dutch and Surinam groups, a remarkable score increase was found in the Zambian group at a retest after the training. Interestingly, substantial score increments were also observed in the Zambian control group. (It was not found for the control groups in the other cultures.) Since the experiences of the control groups with the tests were restricted to the first test administration, it was argued that the score increases of the Zambian subjects were caused by improved test-taking skills learned the first time.

Research such as that by Van der Flier shows that systematic approaches to cross-cultural testing can help to improve the validity of test scores. At the same time it should be clear from this section that it is impossible to offer exhaustive rules about how to design assessment instruments which are universally applicable. Rather, the use of tests requires a thorough knowledge of the local circumstances of the subjects to which the test will be administered. Serious anomalies will result unless the researcher avoids making (implicit) assumptions about testing, which can be valid in his or her own culture, but may not apply in other cultural contexts. Examples of such assumptions are that subjects can cope with multiple-choice formats, will work fast on speeded tests, will try to achieve a high score (rather than maintain good interpersonal relations with the tester), and will easily grasp the meaning of pictorial stimuli.

A Conceptual Framework for the Analysis of Cross-Cultural Score Equivalence

During the last two decades much effort has been put into developing and refining procedures to analyze score equivalence in intergroup comparisons. In the literature these are referred to as item bias studies. (The terms *unbiased* and *score equivalent* are used interchangeably here.) Although there is no agreement in the literature about the definition of item bias (Rudner et al., 1980), most differences involve the statistical analyses and the psychometric models used rather than the underlying ideas. Many definitions share the notion that an item is biased when the psychological construct represented by that item is not the same in each cultural group under study. We would like to propose a definition that is in line with this idea. Item bias is defined here as any difference in an observed score for which there is no corresponding difference in the psychological domain to which the scores are generalized (Poortinga & Malpass, 1986). Suppose that an arithmetic test contains an item, asking how many pencils will go

into six dozen pencils. When "dozen" is a concept that is only known to the examinees in a subset of the cultures involved, a strange response pattern will emerge, since the item does not measure arithmetic reasoning in all groups. The psychological domain of the item in a culture in which the concept of dozen is absent is different from arithmetic reasoning, the domain of the other items of the test. Item bias analyses are implemented to detect those items in a test which do not have the same psychological meaning across cultural groups.

Bias is defined here in terms of the domain of generalization. The latter is not an intrinsic property of an instrument but depends upon the context in which the tests is used. In our opinion, the same is true for item bias. An item can be unbiased in respect to one domain but biased in respect to some (usually larger) domain. For example, suppose that an arithmetic test has been administered to a group of schooled and a group of unschooled children of the same age. If the domain of generalization would be arithmetic *achievement*, the test may well yield unbiased results. However, when the test is taken to measure arithmetic *aptitude* or, even more generally, *intelligence*, any intergroup comparison may be precluded by the presence of item bias.

An important kind of generalization domain is formed by performance criteria in organizational or educational selection. The question of bias in this context has received considerable attention in the literature on fair employment.

A Classification of Item Bias Detection Procedures

An attempted coverage of available item bias detection techniques would go well beyond the scope of the present chapter. We shall restrict ourselves here to a schematic overview of the most important approaches. For a more extensive discussion of various techniques, the reader is referred to Berk (1982) and, more recently, Cole and Moss (1989) and Mellenbergh (1989).

Our scheme is based on three criteria to distinguish bias detection techniques. First, some techniques start from the assumption that a test constitutes a common scale on which scores can be compared. Either for raw scores or for derived scores (e.g., the ability or item difficulty scale in item response theory; see below) corollaries of the assumption are tested at item level. If an item does not meet the requirements postulated, it is removed. The items remaining after item bias analysis are taken to satisfy the requirements for a common scale. In other approaches it is the objective of the analysis to establish whether such a common scale does exist. An example is exploratory factor analysis. The existence of a

common scale is made plausible through the analysis rather than being assumed beforehand.

Second, techniques differ in the kind of data used in the analysis. Some methods are based on the items-by-persons matrix for each culture. In other methods the data matrix is restructured prior to the actual computations. This usually implies some aggregation of the data in the form of averages, inter-item correlations, or contingency tables. These aggregates contain all the information needed for the computation of some item bias statistic. In other words, some techniques use the information available in the full data matrix, while in other instances the relevant information on item bias is assumed to be present in the statistics.

Our third dimension reflects a distinction between so-called conditional and unconditional procedures (Mellenbergh, 1982; Van der Flier et al., 1984). In conditional methods, bias is investigated per ability level, or conditional on the ability level, hence the name. The idea behind conditional approaches is that item bias may not be invariant across the whole range of test scores; the bias effects may be larger for a particular ability level, e.g., for examinees who have a low level of ability. In unconditional procedures the data matrices are compared without any concern for possible group differences in ability levels. Mellenbergh (1982) has argued that conditional methods should be preferred over unconditional methods, because the latter yield more detailed information.

Both conditional and unconditional methods assume the existence of a common scale; hence, the first and third dimensions of our distinction of item bias techniques are not independent. Taking this into account, the three dimensions lead to a scheme as presented in table 10–1.

Table 10–1. Schematic Overview of Item Bias Techniques

Scale	*Input Data for Analysis*	
	Raw Data	*Aggregated Data*
Common scale not assumed	None	Factor analysis, comparison of correlation matrices
Common scale assumed (unconditional)	Analysis of variance	Analysis of p-values, transformed item difficulties, linear structural models
Common scale assumed (conditional)	Item response models	χ^2-approaches

Correlational Techniques. No common scale is assumed in a test on the equality of correlation matrices obtained in different cultures (Browne, 1978) or in exploratory factor analysis (e.g., Irvine, 1979). When using factor analytic techniques, separate analyses are carried out for each culture and the matrices of factor loadings are combined, either by rotating them to a matrix which is closest to the separate matrices (Kaiser et al., 1971) or by rotating all matrices to one target matrix—for example, the matrix obtained in one cultural group (e.g., Van der Flier, 1980). An illustration of the use of factor analysis to establish dimensional identity can be found in the work by Eysenck and colleagues (e.g., Eysenck & Eysenck, 1983). In these studies a test, usually Eysenck's Personality Questionnaire, is administered to a number of subjects in a particular culture. A factor analysis is carried out, followed by a comparison of the factors with those derived from the sample in the United Kingdom on which the original norms of the questionnaire were established. In this procedure the two matrices are rotated to one target.

In the past there has been some debate around the presumed lack of discriminatory power of these techniques (e.g., Horn, 1967; Horn & Knapp, 1973; Humphreys et al., 1969; Ten Berge, 1977). The main objection against target rotations is their "extreme kindness for the data," i.e., it is too easy to get a reasonable fit between hardly related input matrices. Only large differences will be discovered in this way. A demonstration of the extreme flexibility of target rotations as used in the Eysenck tradition has been given by Bijnen and associates (1986).

Unconditional Methods. In most bias detection techniques the existence of a common scale is assumed rather than demonstrated. This is the case for the unconditional methods as well as the conditional methods. In the former, raw item scores or statistics derived from item scores are compared across groups. Examples of the latter approach are the comparison of p-values (Poortinga & Foden, 1975) or their normal deviates (Angoff & Ford, 1973). In these methods a scatter plot of the p-values for a set of items in two groups is prepared. The points representing unbiased items will fall in a fairly narrow region. An item that clearly falls outside that region is considered to be biased. Other examples of unconditional methods are Cleary and Hilton's (1968) use of analysis of variance and Jöreskog's linear structural models. (Applications can be found in Rock and associates, 1982, and Benson, 1987.)

At this point it should be noted that the classification of particular techniques as unconditional methods is mainly determined by their empirical use. The methods mentioned can also be applied as conditional

methods, namely by including level of ability as an additional factor in the analysis. Suppose a researcher wants to compare p-values obtained in various cultural groups. An unconditional analysis entails a direct comparison of the item statistics while, in a conditional analysis, the samples of subjects will be divided according to the level of their raw score and analyzed per level. Conversely, the conditional methods that will be discussed can also be used in an unconditioned way by eliminating ability as a separate factor during the analysis.

Conditional Methods. In the conditional methods of item bias analysis, one particular corollary of the assumption of a common scale is crucial, i.e., that subjects with equal abilities have equal probabilities of correctly answering the test items, irrespective of their group membership. Two kinds of conditional methods can be distinguished: namely, those based on item response theory and χ^2-approaches.

Within the two- and three-parameter models of item response theory, various indices of item bias have been defined (Cole & Moss, 1989; Lord, 1980; Mellenbergh, 1989; Rudner et al., 1980; Shepard et al., 1981, 1984). Some of these indices have a strong intuitive appeal, but it has to be noted that their sampling distribution is usually unknown. This means that the distinction between biased and unbiased items lacks a sound statistical basis and, hence, is arbitrary to some extent.

Within the Rasch model, the one-parameter model of item response theory, there are also various fit statistics, which can be used as bias indices. These statistics, with known sampling distributions, vary from omnibus tests in which all items are evaluated simultaneously (e.g., Andersen, 1973) to highly specific tests in which the contribution of each separate item to the overall fit can be evaluated (e.g., Van den Wollenberg, 1982).

In the second kind of conditional techniques, the χ^2-approaches, contingency tables are analyzed (Marascuilo & Slaughter, 1981; Mellenbergh, 1982). The most frequently used table has three factors, score level, culture, and response (right/wrong). The observed frequencies are entered in the cells. The table is analyzed for each test item separately. The fit of an item is evaluated by means of a χ^2-statistic, hence the name. An application can be found in Van der Flier and associates (1984).

A Procedure Based on Generalizability Theory

After this general overview of different item bias techniques, an example of a procedure and the rationale behind it will be presented in some detail;

it has previously been described by Van de Vijver and Poortinga (1982).

Our framework for the investigation of bias (or score equivalence) is based on generalizability theory (Cronbach et al., 1972). In the most simple study a test is administered to members of two culturally different groups. This design has been labeled as "Design V-B" by Cronbach and associates (1972, p. 38). It can be described as a crossing of the factors Stimulus and Persons, with the latter nested in the factor Culture, designated as Stimulus x Persons (Culture). A single item score, denoted by $X_{sp(c)}$, is assumed to consist of the following linear, additive components (Van de Vijver & Poortinga, 1982)[1]:

$$X_{sp(c)} = \mu + S_s + P_{p'}PC_{pc} + C_c + SC_{sc} + SP_{sp'}SPC_{spc'}E_{spc} \qquad (1)$$

where
μ is the overall mean;
S_s $(s = 1, \ldots, n_s)$ is the main effect for Stimulus (items);
$P_{p'}$ PC_{pc} $(p = 1, \ldots, n_p)$ is the confounded effect for the main effect Persons (P) and the Person by Culture interaction (PC);
C_c $(c = 1, \ldots, n_c)$ is the main effect for Culture;
SC_{sc} is the interaction between Stimulus and Culture;
$SP_{sp'}$ $SPC_{spc'}$ E_{spc} is the confounded effect of the Stimulus by Person interaction (SP), the Stimulus by Person by Culture interaction (SPC) and the error term (E).

The analysis of score equivalence starts with an analysis of variance. The sources of variance are schematically drawn in figure 10–1. On the basis of the observed mean squares, variance components can be estimated. The computational formulas are presented in table 10–2. Most current

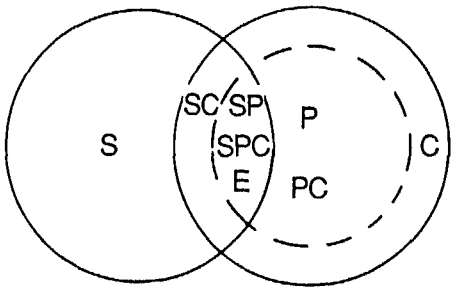

Figure 10–1. *Schematic Representation of Variance Components in a Stimulus (S) by Culture (C) Design with Persons (P) Nested in Cultures*
Source: From van de Vijver and Poortinga (1982). Reprinted by permission from Sage Publications.

Table 10–2. The Computation of the Estimated Variance Components

Estimated Variance Component	*Computational Formula*
$\sigma^2(SP, SPC, E)$	$= MS(SP, SPC, E)$
$\sigma^2(SC)$	$= (MS(SC) - MS(SP, SPC, E))/n_p$
$\sigma^2(P, PC)$	$= (MS(P, PC) - MS(SP, SPC, E))/n_s$
$\sigma^2(S)$	$= (MS(S) - MS(SC))/n_p n_c$
$\sigma^2(C)$	$= (MS(C) - MS(P, PC) - MS(SC) + MS(SP, SPC, E))/n_s n_p$

statistical computer packages contain a program for the estimation of variance components (e.g., the program P8V in BMDP; Dixon, 1981).

Generalizability Coefficients. The components of variance are used to calculate coefficients of generalizability, which reflect the impact of particular sources of variance in the dependent variable (cf. Cronbach et al., 1972). For example, when a researcher is interested in the contribution of cross-cultural score differences to the overall score variance, an estimated generalizability coefficient for the main effect culture can be computed. This coefficient indicates what proportion of the score variance is accounted for by cross-cultural differences in mean scores. Generalizability coefficients are closely related to traditional reliability coefficients; both have a lower limit of 0.00 and an upper limit of 1.00. A high value of a generalizability coefficient for a particular source indicates a high contribution of this source to the total score variance.

Generalizability coefficients also have the same disadvantage as reliability coefficients, namely that their size depends on the number of items on which they are based. In classical test theory the Spearman-Brown formula is used to estimate the reliability of a test at various lengths under the assumption of parallelism of the items (e.g., Allen & Yen, 1979, formula 4.7). The same formula applies to generalizability coefficients. Thus, given a particular number of levels for a factor in a study, the generalizability coefficient of that factor can be estimated for any other number of levels by means of the Spearman-Brown formula (Golding, 1975). This provides us with a method to overcome the disadvantage mentioned. A convenient way to get mutually comparable coefficients is to compute these at unit level, equivalent to the computation of the reliability of a one-item test in classical test theory. These unit length coefficients of generalizability are expressed on an identical scale, irrespective of the kind of factors or the number of levels in a factor.

In the present context two generalizability coefficients are of major interest. (This choice will be motivated later.) The first one, denoted by ρ^2_{sc}, evaluates the importance of the stimulus by culture interaction, the traditional item bias statistic (e.g., Cleary & Hilton, 1968; Poortinga, 1971). For the second coefficient, denoted by ρ^2_{c+sc}, both the main effect culture and the stimulus by culture interaction are of interest.

The computational formulas for these coefficients are:

$$\rho^2_{sc} = \frac{\sigma^2_{sc}}{\sigma^2_{sc} + \sigma^2_{sp,spc,e}/n'_p} \tag{2}$$

$$\rho^2_{c+sc} = \frac{\sigma^2_c + \sigma^2_{sc}}{\sigma^2_c + \sigma^2_{sc} + \sigma^2_{p,pc}/n'_p + \sigma^2_{sp,spc,e}/n'_p} \tag{3}$$

in which $n'_p = n_p$ for full length coefficients and $n'_p = 1$ for unit length coefficients.

In generalizability theory the statistical significance of a generalizability coefficient is considered relatively unimportant. In fact, a generalizability coefficient is an estimate of effect size rather than significance level. Although we concur with this position, it may be noted that the sampling distribution of ρ^2_{sc} can be derived quite easily. Only the distribution of ρ^2_{c+sc} is unknown. For a test of the hypothesis that $\rho^2_{sc} = 0$, assuming full length estimates, the following holds (Kristof, 1963; Kraemer, 1981):

$$\frac{1}{1 - \rho^2_{sc}} = \frac{1}{1 - \sigma^2_{sc}/(\sigma^2_{sc} + \sigma^2_{sp,spc,e}/n_p)} = \frac{\sigma^2_{sc} + \sigma^2_{sp,spc,e}/n_p}{\sigma^2_{sp,spc,e}/n_p} \tag{4}$$

When both nominator and denominator are multiplied by n_p, the latter coefficient is the F-ratio for the SC-interaction with $(n_s - 1)\,(n_c - 1)$ and $n_c(n_s - 1)\,(n_p - 1)$ degrees of freedom; thus, it appears that ρ^2_{sc} differs significantly from zero whenever the F-ratio for the SC-component in the analysis of variance is significant.

There are fewer conventions about effect size than about significance in the literature. What minimum value a coefficient of generalizability should attain before it can be considered to be meaningfully contributing to the score variance is a matter of debate. As a rule of thumb, a value of .05 for the unit length coefficient seems to work quite well.

ρ^2_{sc}. The size of ρ^2_{sc} is particularly important when a researcher has good reasons to believe that bias will manifest itself primarily at the level of separate items. A value larger than .05 indicates the presence of item bias. When the value of ρ^2_{sc} is substantial, an inspection of the residuals in each cell of the data matrix after removal of the main effects for Stimulus

and Culture will indicate which items induce bias. After these have been eliminated, a new analysis of variance is carried out for the reduced data matrix. This iterative procedure can be repeated until ρ^2_{sc} becomes acceptably low—say, less than .05.

This kind of procedure leans heavily on the particulars of a data set, thereby implicitly threatening the replicability of the results. To control for this, a researcher can split each sample randomly in two. Separate bias analyses are carried out for the two data sets. Afterwards the results are combined again. A conservative strategy to deal with bias is to discard all items that turn out to be biased in at least one of the analyses. A more lenient strategy is to exclude only those items that show evidence of bias in both data sets.

So far, the present approach does not differ from many methods of item bias analysis described in the literature. When ρ^2_{sc} is acceptably low, the bias analysis ends and possibly remaining intergroup score differences (i.e., $\rho^2_c > 0$ in terms of generalizability theory) are interpreted as reflections of valid cross-cultural differences.

ρ^2_{c+sc}. In our procedure, the computation of ρ^2_{c+sc} is included in the bias analysis, as this coefficient can also reflect bias. If both ρ^2_{sc} and ρ^2_{c+sc} have low values, it can be concluded that the scores are equivalent across the cultural groups, but this will only be the case if there are no cross-cultural differences in the test score levels. In cross-cultural studies equal averages are the exception rather than the rule. Consequently, more often than not, ρ^2_{c+sc} has a substantial value. The researcher is then confronted with the far from trivial problem of how to interpret the coefficient. The iterative item bias procedure just described (and most related bias detection techniques) provide adequate information only if the factors causing the bias leave a substantial proportion of the items unaffected. This presupposition, almost invariably used in item bias studies, is not self-evident and is even unlikely to be realistic when groups with a large cultural distance are compared. It is more likely that a source of bias has an effect on all items and, consequently, exerts a strong influence on the overall test score. In previous sections a number of examples have been given: The notions on which the items of a test are based may be foreign to a cultural setting (Reuning & Wortley's 1973 example of the application of Porteus' Maze Test among the Bushmen), the response medium can induce bias (Serpell's 1979 study on iron wire-modeling versus drawing), particular aspects of the test administration can cause problems (Greenfield's 1966 and Irvine's 1978 studies on conservation among the Wolof), and so on. In these cases bias leads to massive cross-cultural differences in performance, which in

an analysis of variance will come out in the main effect for Culture and probably to a much lesser extent in the Stimulus by Culture interaction.

The Analysis of Item Bias Reconsidered

The shift of a bias effect from the *SC*-component to the *C*-component in an analysis of variance can easily be demonstrated in a Monte Carlo study (Poortinga & Van de Vijver, 1987). In figure 10–2 some results are presented for a low and a high level of bias. The graphs show that $\hat{\rho}^2_{sc}$ initially increases with the number of biased items, as expected. However,

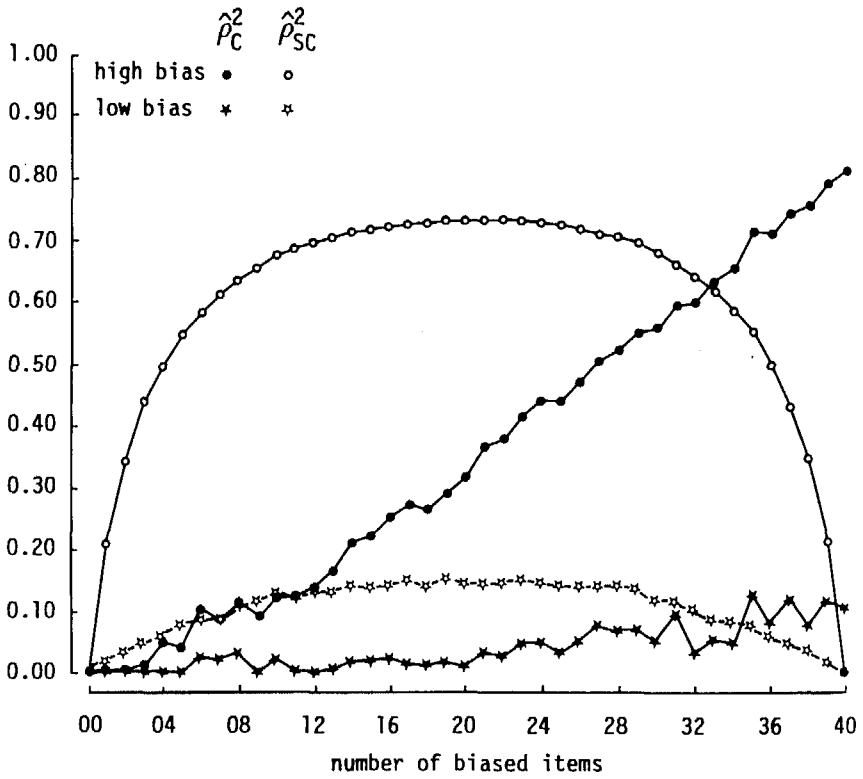

Figure 10–2. Estimated Size ($\hat{\rho}^2$) of the SC-Interaction and the Main Effect for Culture as a Function of the Number of Biased Items for Low Bias and High Bias
Source: From Poortinga and Van de Vijver (1987). Reprinted by permission from Sage Publications.

the increase is not monotonic. After reaching an upper limit, $\hat{\rho}^2_{sc}$ is going down when a still larger number of items are biased. In contrast, $\hat{\rho}^2_c$ (the generalizability coefficient for the main effect of Culture) is a monotonic function of the number of biased items. It should be noted that the simulations mentioned here were carried out under the assumption that the bias favored one group systematically. The more items are biased, the larger this shift from the interaction component to the main effect for culture will be. This situation should not be considered uncommon, since in cross-cultural psychology mostly Western tests are used to compare the performance of Western and non-Western samples. Bias effects will frequently disadvantage these latter samples (Van de Vijver & Poortinga, 1985).

A main effect can reflect either bias, valid psychological differences (that is, differences in the domain of generalization), or a combination of both. A similar argument holds for ρ^2_{sc}; this coefficient also need not be indicative of bias only, but can also reflect psychological differences (Van de Vijver & Poortinga, 1985). Therefore, the classical item bias paradigm is inadequate in its exclusive interpretation of the *SC*-component as bias and of the *C*-component as evidence for valid differences. In our opinion, a more balanced interpretation and explanation of the *SC*- and *C*-components is one of the essential tasks of the cross-cultural psychologist.

The criticisms expressed toward the item bias are not restricted to the analysis of variance model. Any other common model of bias analysis is subject to similar difficulties in interpretation. Even in item response theory, often considered as the psychometrically most advanced model to study item bias, it is impossible to arrive at unbiased intergroup comparisons when most or all items of the test are biased against a single group.

In general, it will be very difficult to distinguish bias from real differences when the only data at hand are the test data; in the last section of this chapter a model will be outlined that includes context variables as explanatory variables for the observed cross-cultural differences.

Uniform and Non-Uniform Bias

As noted before, in the item bias detection techniques a distinction has been made between conditional and unconditional methods. Within the conditional methods, Mellenbergh (1982) has introduced a further distinction, uniform and nonuniform bias. Per item, a Persons (Cultures) by Ability data matrix is composed. An item is nonuniformly biased when

both the interaction between ability and culture and the main effect for culture are significant; an item is uniformly biased when only the main effect for culture is significant; if both effects are nonsignificant, the item is said to be unbiased.

In the design suggested by us, ability can be introduced as an additional factor, thereby forming a Persons (Cultures) by Stimulus by Ability design. The method of analysis is then a conditional one, with total test score being used as a separate independent variable.

An Example

Our approach to the analysis of score equivalence in intergroup comparisons will be illustrated with a set of data previously reported by Van de Vijver and Poortinga (1982). The data were collected on three samples: Indian students, Dutch students, and Dutch army conscripts. All subjects were males; each group consisted of 32 subjects.

The subjects answered 43 items of the Strength of Excitation Scale in the third experimental edition of the Temperament Inventory (Strelau, 1972). Each question of this inventory has three response alternatives: affirmative (2 points), undecided (1 point), and negative (0 points).

In the group of Indian students, a split-half reliability coefficient of .72 was observed; for the Dutch students this was .78; and for the Dutch soldiers .75.

In table 10–3 the results of the analysis of variance are given, together with the estimated components of variance. The *SC*-component was found to be significant, while the significance level of the main effect for culture, computed by means of a quasi-F ratio, was .09. However, as noted earlier, these *F*-ratios are not of primary interest here.

From the estimated variance components, the generalizability coefficients were computed. The value of $\hat{\rho}^2_{sc}$ (unit length) was .08, which exceeds the proposed criterion of .05. This means that the Temperament Inventory does not constitute a score equivalent scale for these samples from the Netherlands and India.

In subsequent analyses, potential sources of this lack of comparability were investigated; this was done by eliminating subjects or samples from the data set. The results of the analyses are presented in table 10–4. First, the two Dutch groups were taken together, thereby defining Dutch males from approximately 18 to 25 years as our population of interest. The value of $\hat{\rho}^2_{sc}$ was .02, a value which was also found for $\hat{\rho}^2_{c+sc}$. This means that within the population of young Dutch males the Temperament Inventory could be taken to yield score equivalent results. As a next step, the scores

Table 10–3. Results of the Analysis of Variance and the Estimated Components of Variance ($\hat{\sigma}^2$)

Source	*SS*	*df*	*MS*	*F prob.*	$\hat{\sigma}^2$
S	444.18	24	10.67	5.34 .00	.0903
C	16.52	2	8.26	2.50 .09	.0036
P,PC	172.11	93	1.85	3.37 .00	.0303
SC	168.01	84	2.00	3.64 .00	.0453
SP,SPC,E	2146.06	3906	0.55		.5494

Table 10–4. Estimated Generalizability Coefficients

Groups[1]	$\hat{\rho}^2_{sc}$	$\hat{\rho}^2_{c+sc}$	*Comments*
is,ds,dc	.08	.08	All subjects
ds,dc	.02	.02	Only Dutch subjects
is,ds	.11	.11	Only students
dc	.07	.17	Only Dutch conscripts; high vs. low scorers
dc	.01	.01	Only Dutch conscripts; random split

[1] *is* = Indian students; *ds* = Dutch students; *dc* = Dutch conscripts.

of the Dutch and the Indian student samples were taken together, thereby defining male students of approximately 18 to 25 years as the population of interest. The value of $\hat{\rho}^2_{sc}$ was .11, clearly indicating the presence of bias.

In analyses not reported here, it was observed that many items had high endorsement rates; this led to the hypothesis that the lack of score equivalence was caused by ceiling effects. In order to test this hypothesis, one of the groups, the Dutch army conscripts, was split up into two subsamples, one with low scorers and the other group with high scorers. For these subgroups $\hat{\rho}^2_{sc}$ was .07, while for a random split of the conscripts in two subgroups a value of .01 was observed. The value of .07 seems to be high enough to conclude that ceiling effects are at least one reason for the lack of equivalence in the total data set, although there is hardly any doubt that other sources also have played a role. It may be noted that this split in high and low scorers is somewhat similar to a conditional procedure, as discussed previously. The limited number of subjects in each cultural group prohibited a finer distinction of ability in more than two score levels.

Explaining Cross-Cultural Differences

A careful analysis of bias within a given data set can provide important cues about how observed cross-cultural differences should be interpreted

(Malpass & Poortinga, 1986; Poortinga & Van der Flier, 1988). In the previous section, it was argued that the choice between bias and valid cross-cultural differences can be very complicated and definitely requires more than a simple inspection of item bias statistics. There is still another problem: in most cross-cultural studies the interpretation of the data is post hoc and, hence, tentative. In general it will be impossible to provide decisive reasons why a particular post hoc interpretation should be preferred. The existence of this problem has been recognized in cross-cultural psychology and the need for testable theories about these differences which allow less ambiguous conclusions has been emphasized (e.g., Malpass, 1977; Segall, 1986; Whiting, 1976). In this section a methodological framework will be presented which can help to structure efforts of testing such theories.

The basic idea behind our approach is that not only should the dependent variable be measured on which a researcher anticipates a difference. In the design, measurements should also include the postulated antecedent conditions, which presumably have led to an observed intergroup difference.

Since in many instances only the dependent variables are clearly identifiable and the boundary between independent variables and bias variables is fluent and somewhat arbitrary, we shall use the term *context variables* for both.

Before continuing, a brief digression on these context variables is necessary. In view of the broad range of variables related to culture, no restriction is imposed on the domain from which context variables can be recruited. On the contrary, there can be sociological variables, like mode of subsistence or socioeconomic status, measured, for instance, by family income. Other context variables will be of a more psychological nature: for example, educational background. But also physical, physiological, or economical variables can be relevant. Neither are there restrictions on the methods employed to gather data about the context variables. Self-report inventories, judgmental methods, or even external referents like the Human Research Area Files which are based on ethnographic descriptions of cultural communities (Narroll et al., 1980) are acceptable.

A Multiple Regression Model

The strategy proposed here amounts to a check on the contribution of all context variables to the observed cross-cultural differences on the

dependent variable. The model is presented here in the form of a multiple regression equation, although this is not the only possible choice.[2]

The statistical technique used is a hierarchical regression analysis. In the first step of the analysis, the context variables are entered as predictors. For simplicity of presentation, mutual independence of the predictors will be assumed, although it is not required by the model. The first step provides information as to whether the context variables significantly contribute to the variance in the dependent variable, which in general will be an item or test score. In the second step, culture is added to the equation as a predictor. This step gives information about the size of the remaining culture effects after elimination of the effects of the context variables.

First Step. Suppose a test is administered to people in a number of cultures and, in addition to this, data on a single context variable was also gathered. In the first step of the hierarchical regression analysis the independent variable used to predict the test score is the context variable. The regression equation for this simple linear regression model with one predictor K as context variable, is given by:

$$X_{pk} = a + b_k K_p + E_{pk} \tag{5}$$

where
X_{pk} is the observed test score of individual p;
a is the intercept;
b_k is the regression coefficient of the context variable K;
K_p is the value of individual p for the context variable K. When more than one context variable is used, the term $b_k K_p$ will consist of the sum of these, each predictor having its own b_k;
E_{pk} is the error component.

The contribution of the context variable to the variance in the dependent variable is evaluated by means of the multiple correlation coefficient which can be tested for significance (e.g., Cohen & Cohen, 1983, formula 3.6.1). When this coefficient does not differ from zero, the context variable does not explain score differences across groups and, hence, gives no insight into the nature of the observed intergroup differences.

Second Step. A significant multiple correlation coefficient indicates that the context variable at hand is a valid predictor of cross-cultural differences. Even though this can be very important from a theoretical point of view, as it suggests that a determinant of intergroup differences has been identified, it is only one side of the coin. The question still open is how much of the total group differences on the dependent variable remains

after a correction for the impact of the context variable has been carried out. This information is provided in the second step of the hierarchical analysis, in which culture is introduced as a predictor.

Culture is a nominal variable that can enter a regression equation in three ways: by dummy coding, effect coding, or contrast coding (Cohen & Cohen, 1983, ch. 5). The choice is immaterial for the present purpose as all three lead to the same multiple correlation. The regression equation for the second step of the analysis is:

$$X_{p(c)k} = a' + b_k K_p + b_{kc} C_c + E_{p(c)k} \tag{6}$$

where
$X_{p(c)k}$ is the observed test score;
a' is the intercept;
b_{kc} is the regression coefficient for culture after removal of the effects of the context variable;
C_c is the culture effect;
$E_{p(c)k}$ is the error component.

In this model, context variables and culture are successively entered in the analysis, and the effect of culture is computed after the scores have been corrected for the effects of the context variables. The size of the multiple correlation coefficient of the second analysis gives information about how much can still be gained in prediction by including additional context variables should they be available. The difference between the multiple correlation coefficients of the first and second step in the analysis can be tested for significance (e.g., Cohen & Cohen, 1983, formula 4.4.1).

Context variables are the more valuable, as they explain a larger part of the score variance in the dependent variable. For a given value of the multiple correlation in the first analysis, the explanatory power of the context variable is highest when the introduction of culture in the second analysis does not increase the multiple correlation significantly.

An Implication

A nontrivial and at first sight paradoxical consequence of the replacement of the *C*-component by the *K*-component is that the *C*-component, the traditional index of cross-cultural differences, appears not to be of primary interest in cross-cultural psychology; rather, in the present approach the *C*-component represents the cultural differences not yet explained by context variables. *Cultural differences which have multiple interpretations*

are in themselves rather meaningless; they only form the starting point for further analysis. It is the task of the cross-cultural psychologist to minimize the size of the C-component by replacing it with explanatory context variables rather than to demonstrate the presence of any C-component. In other words, cross-cultural psychologists should emphasize the interpretation and explanation of cross-cultural differences by means of context variables rather than the mere documentation of these differences in the form of a (significant) *C*-component.

A Brief Digression on Context Variables

It may seem attractive to use nominal classifications like race, nationality, cultural group, or language as context variables. For instance, McNemar (1975) states that race should be included in the regression equation as it is more often than not a significant predictor of the variables in which psychologists are interested, e.g., job success and school performance. Whatever their attractiveness, nominal classifications are methodologically invalid context variables. The major problem is their mutual interchangeability. Any score difference between groups can be ascribed to a difference in culture, or in religion, or in language, or in race, or to any combination of these. The choice is arbitrary and not logically compelling. In a regression analysis, each of these variables can be given the same coding and, hence, no distinction can be made between them. Interchangeability of the position of groups is no longer possible when context variables are measured on a scale of at least an ordinal level.

Another type of context variable that seems intuitively attractive is a psychological test that is similar to the test under study. After all, the best predictor of the score on the target test will be the subject's score on a parallel test. From a theoretical point of view, such a context variable does not yield much information beyond that provided by the original instrument. Both are likely to be affected by the same sources of intergroup variance. Therefore, the status of a test that resembles the dependent variable as context variable will often be debatable. We can carry this argument still a step further. When groups differ in test-wiseness, almost any test will show a difference. It will be test-wiseness rather than the presumed constructs underlying the tests which should be considered as the agent behind the score differences. As the dependent variable is more dissimilar to a context variable—for instance, when it is a sociological or economic measure—it becomes more unlikely that corresponding differ-

ences on the dependent and context variables are a function of a common source of error.

An Example

The data on the Strength of Excitation Scale reported earlier formed part of a larger cross-cultural project on the cultural invariance of basic personality parameters (Poortinga & Van de Vijver, 1987). In another part of the project the habituation of the orienting reflex (OR) was studied, measured by the subjects' skin conductance response (SCR). Pure tones of 500 Hz and a duration of 1 second were presented at intervals of 20 seconds. Two identical sessions were held for each subject. Four samples were involved in the study, in a 2 (groups) by 2 (cultures) design: Indian students, illiterate Indians of the Juang group, Dutch students, and Dutch military conscripts. Here we shall be concerned with the responses to the first stimulus presentation in each session.

It is instructive for our argument to consider what would be concluded on the basis of a "classical analysis," such as an analysis of variance. Two of these analyses were carried out, one per test session. The results are presented in table 10–5. In the first session, neither the independent variables (group and culture) nor their interaction turned out to be significant while, in the second session, both main effects yielded significant values. It may seem tempting to interpret the (admittedly rather weak) cross-cultural differences in terms of genuine psychological differences between the cultures at hand.

In the experiment various measurements had been collected on the "state of arousal" of the subjects in the experimental situation. One of these was used as context variable, namely the extent of spontaneous fluctuations in the skin conductance recorded during periods of rest at the beginning and at the end of each experimental session.

Prior to the analyses the question should be addressed whether spontaneous SCR meets the requirements imposed on context variables. A major objection might be that this variable constitutes a "parallel" measure of the dependent variable. However, the skin conductance recorded during rest and during the experimental task differ in one crucial aspect: the presence or absence of an external stimulus. This makes the former an adequate measure of pre-experimental individual differences in arousal.

In the first of the two regression analyses the impact of the rest SCR on the OR was evaluated. The results are presented in table 10–6. The multiple correlation coefficient was highly significant. In the second

Table 10–5. Significance Levels of the Analysis of Variance

Source	*Session 1*	*Session 2*
Culture	.19	.01
Group	.71	.01
Culture x Group	.10	.42

Table 10–6. Results of the Hierarchical Regression Analysis

Statistic	*Session 1*	*Session 2*
SMC (Rest)	*.382**	*.210**
Increments in SMC	G. 013	G .007
(second analysis)	GxC.014	C .021
	C .015	CxG.022

$^{*}p < .01$.

Notes: SMC = Squared multiple correlation coefficient; G = Group; C = Culture; CxG = Culture by group internation.

analysis an additional set of independent variables was introduced, namely culture, group, and their interaction. In this analysis no remaining intercultural differences were found. The increase in squared multiple correlation from the first to the second analysis was less than 0.03 for each of the two sessions. It is important to note that the size of the factor "culture," significant in the analysis of variance, was rendered nonsignificant in the second analysis, that is, after a correction for differences in spontaneous fluctuations in skin conductance.

In conclusion, differences in spontaneous fluctuations in skin conductance accounted for cross-cultural differences in the size of the initial OR. This makes it unlikely that these differences were caused by a differential sensitivity in the two cultures for the impact of the stimulus on the neuropsychological apparatus.

It could be argued that the cultural differences in fluctuations in skin conductance should not be taken for granted as done here but require further study. We concur with this view. The fluctuations in SCR can be considered a first hypothesis to account for the cross-cultural differences observed which can be gradually shaped and refined in later studies. The major focus of the present approach—the replacement of the vague concept of culture by one or more specific variables—remains intact whatever hypothesis should prove to be the most valid explanation.

A Final Remark

The Zeitgeist of cross-cultural psychology can be described as a "difference climate," i.e., an ideological atmosphere in which the documentation of intergroup differences in psychological functioning is considered to be the major task. In this chapter we have suggested a shift in orientation. Cross-cultural psychologists should try to explain rather than explore cultural differences. Some methodological and psychometric tools that facilitate this orientation have been described in this chapter.[3]

Notes

1. In an analysis of variance the researcher has to decide which factors in the design are random and which factors are fixed (Hays, 1973). The formulas of table 10–2 are based on an all-random model. This choice deserves some comment, in particular for the factor culture. The researcher's interest and conclusions usually go beyond the particular cultures included in a study. Treating culture as a fixed factor implies that only the cultures involved are taken to be of interest. On the other hand, the choice of particular cultures for a study is usually motivated by convenience and the presumed presence of certain characteristics and not by random selection, as would be required for a random factor. When the researcher wants to make generalizations about a dimension on which the cultures under study are assumed to vary, it seems appropriate to consider culture as a random factor.

2. Although this will not be elaborated here, there is a close link between the analysis of variance model of the previous section and the multiple regression model (e.g., Cohen & Cohen, 1983; Pedhazur, 1982).

3. The reader is referred to Campbell (1961) and Holland and Rubin (1983) for related approaches aimed at maximizing the interpretability of group differences.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika* 38:123–140.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test for scholastic aptitude. *Journal of Educational Measurement* 10:95–105.

Benson, J. (1987). Detecting item bias in affective scales. *Educational and Psychological Measurement* 47:55–67.

Berk, R. A. (ed.) (1982). *Handbook of methods for detecting item bias.* Baltimore: Johns Hopkins University Press.

Biesheuvel, S. (1949). Psychological tests and their application to non-European peoples. In G. B. Jeffery (ed.), *The yearbook of education.* London: Evans, pp. 87–126.

Bijnen, E. J., Van der Net, Th. Z. J., & Poortinga, Y. H. (1986). On cross-cultural comparative studies with the Eysenck Personality Questionnaire. *Journal of Cross-Cultural Psychology* 17:3–16.

Boorstin, D. J. (1985). *The discoverers.* New York: Random House.

Browne, M. W. (1978). The likelihood ratio test for the equality of correlation matrices. *British Journal of Mathematical and Statistical Psychology* 31:209–217.

Campbell, D. T. (1961). The mutual methodological relevance of anthropology and psychology. In F. L. K. Hsu (ed.), *Psychological anthropology: Approaches to culture and personality.* Homewood, IL: Dorsey Press, pp. 333–352.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement* 28:61–75.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (ed.), *Educational measurement*, 3rd ed. New York: Macmillan, pp. 201–219.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral instruments* New York: Wiley.

David, K. H. (1974). Cross-cultural uses of the Porteus Maze. *Journal of Social Psychology* 92:11–18.

Deregowski, J. B., & Serpell, R. (1971). Performance on a sorting task: A cross-cultural experiment. *International Journal of Psychology* 6:271–281.

Dixon, W. J. (1981). *BMDP statistical software.* Berkeley: University of California Press.

Eysenck, H. J. (1984). The effect of race on human abilities and mental test scores. In C. R. Reynolds & R. T. Brown (eds.), *Perspectives on bias in mental testing.* New York: Plenum, pp. 249–262.

Eysenck, H. J., & Eysenck, S. B. G. (1983). Recent advances in the cross-cultural study of personality. In J. N. Butcher & C. D. Spielberger (eds.), *Advances in personality assessment*, Vol. 2. Hillsdale, NJ: Erlbaum, pp. 41–70.

Golding, S. L. (1975). Flies in the ointment: Methodological problems in the analysis of the percentage of variance due to persons and situations. *Psychological Bulletin* 82:278–288.

Greenfield, P. M. (1966). On culture and conservation. In J. S. Bruner, R. R. Olver, & P. M. Greenfield (eds.), *Studies in cognitive growth.* New York: Wiley, pp. 225–256.

———. (1979). Response to Wolof "magical thinking." *Journal of Cross-Cultural Psychology* 10:251–256.

Hays, W. L. (1973). *Statistics for the social sciences.* London: Holt, Rinehart and Winston.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (eds.), *Principals of modern psychological measurement.* Hillsdale, NJ:

Horn, J. L. (1967). On subjectivity in factor analysis. *Educational and Psychological Measurement* 27:811–820.

Horn, J. L., & Knapp, J. R. (1973). On the subjective character of the empirical base of Guilford's structure-of-intellect model. *Psychological Bulletin* 80:33–43.

Erlbaum, pp. 3–26.
Humphreys, L. G., Ilgen, D., McGrath, D., & Montanelli, R. (1969). Capitalization on chance in rotation of factors. *Educational and Psychological Measurement* 29:259–271.
Hutchins, E. (1980). *Culture and inference.* Cambridge, MA: Harvard University Press.
Irvine, J. T. (1978). Wolof "magical thinking": Culture and conservation revisited. *Journal of Cross-Cultural Psychology* 9:300–310.
Irvine, S. H. (1979). The place of factor-analysis in cross-cultural methodology and its contribution to cognitive theory. In L. Eckensberger, W. Lonner, & Y. H. Poortinga (eds.), *Cross-cultural contributions to psychology.* Lisse: Swets & Zeitlinger, pp. 300–343.
Irvine, S. H., & Carroll, W. K. (1980). Testing and assessment across cultures: Issues in methodology and theory. In H. C. Triandis & J. W. Berry (eds.), *Handbook of cross-cultural psychology*, Vol. 2. Boston: Allyn & Bacon, pp. 181–244.
Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.
Kaiser, H. F., Hunka, S., & Bianchini, J. C. (1971). Relating factors between studies based upon different individuals. *Multivariate Behavioral Research* 5:409–422.
Kraemer, H. C. (1981). Extension of Feldt's approach to testing homogeneity of coefficients of reliability. *Psychometrika* 46:41–45.
Kristof, W. (1963). The statistical theory of stepped up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika* 28:221–238.
Lancy, D. E. (1983). *Cross-cultural studies in cognition and mathematics.* London: Academic Press.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.
Malpass, R. S. (1977). Theory and method in cross-cultural psychology. *American Psychologist* 32:1069–1079.
Malpass, R. S., & Poortinga, Y. H. (1986). Strategies for design and analysis. In W. J. Lonner & J. W. Berry (eds.), *Field methods in cross-cultural psychology.* Newbury Park, CA: Sage, pp. 47–83.
Mann, C. W. (1940). Mental measurement in primitive communities. *Psychological Bulletin* 37:366–395.
Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2-statistics. *Journal of Educational Measurement* 18:229–248.
McNemar, Q. (1975). On so-called test bias. *American Psychologist* 30:848–851.
Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics* 7:105–118.
————. (1989). Item bias and item response theory. *International Journal of Educational Research* 13:127–143.
Mercer, J. R. (1984). What is a racially and culturally nondiscriminatory test? A sociological and pluralistic perspective. In C. R. Reynolds & R. T. Brown

(eds.), *Perspectives on bias in mental testing*. New York: Plenum Press, pp. 293–356.

Narroll, R., Michick, G. L., & Narroll, F. (1980). Holocultural research methods. In H. C. Triandis & J. W. Berry (eds.), *Handbook of cross-cultural psychology*, Vol. 2. Boston: Allyn & Bacon, pp. 479–521.

Neimark, E. D. (1975). Intellectual development during adolescence. In F. D. Horowitz (ed.), *Review of child development research*, Vol. 4. Chicago: University of Chicago Press, pp. 541–594.

Ombrédane, A., Robaye, F., & Plumail, H. (1956). Résultats d'une application répétée du matrix-couleur à une population de Noirs Congolais. *Bulletin du Centre d'Etudes de Recherches Psychotechniques* 5:129–147.

Ord, I. G. (1970). *Mental tests for pre-literates*. London: Ginn.

Pedhazur, E. (1982). *Multiple regression in behavioral research*, 2nd ed. New York: Holt, Rinehart & Winston.

Pick, A. D. (1981). Cognition: Psychological perspectives. In H. C. Triandis & W. Lonner (eds.), *Handbook of cross-cultural psychology*, Vol. 3. Boston: Allyn & Bacon, pp. 117–154.

Poortinga, Y. H. (1971). Cross-cultural comparison of maximum performance tests. *Psychologia Africana Monograph* 6.

Poortinga, Y. H., & Foden, B. I. M. (1975). A comparative study of curiosity in black and white South African students. *Psychologia Africana* Monograph 8.

Poortinga, Y. H., & Malpass, R. S. (1986). Making inferences from cross-cultural data. In W. J. Lonner & J. W. Berry (eds.), *Field methods in cross-cultural psychology*. Newbury Park, CA: Sage, pp. 17–46.

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

Poortinga, Y. H., & Van der Flier, H. (1988). The meaning of item bias in ability tests. In S. H. Irvine & J. W. Berry (eds.), *Human abilities in cultural context*. Cambridge: Cambridge University Press, pp. 166–183.

Porteus, S. D. (1917). Mental tests with delinquents and Australian aboriginal children. *Psychological Review* 24:32–42.

Porteus, S. D. (1965). *Porteus Maze Test: Fifty years of application*. Palo Alto, CA: Pacific Books.

Price-Williams, D. R. (1962). Abstract and concrete modes of classification in a primitive society. *British Journal of Educational Psychology* 32:50–61.

Reuning, H., & Wortley, W. (1973). Psychological studies of the Bushmen. *Psychologia Africana*, Monograph Supplement, 7.

Rock, D. A., Werts, C., & Grandy, D. (1982). *Construct validity of the GTE Aptitude Test across populations* (ETS Research Report 81–57). Princeton, NJ: Educational Testing Service.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980) Biased item detection techniques. *Journal of Educational Statistics* 5:213–233.

Schwarz, P. A. (1961). *Aptitude tests for use in developing nations*. Pittsburgh: American Institute for Research.

Scribner, S., & Cole, M. (1981). *The psychology of literacy.* Cambridge, MA: Harvard University Press.

Segall, M. H. (1986). Culture and behavior: Psychology in global perspective. *Annual Review of Psychology* 37:523–564.

Serpell, R. (1979). How specific are perceptual skills? *British Journal of Psychology* 70:365–380.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparisons of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics* 6:317–375.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics* 9:93–128.

Strelau, J. A. (1972). A diagnosis of temperament by nonexperimental techniques. *Polish Psychological Bulletin* 3:97–103.

Super, C. M. (1981). Behavior development in infancy. In R. H. Munroe, R. L. Munroe, & B. B. Whiting (eds.), *Handbook of cross-cultural human development.* New York: Garland STPM Press, pp. 181–270.

Ten Berge, J. M. F. (1977). *Optimizing factorial invariance.* Groningen: VRB Drukkerijen.

Van de Vijver, F. J. R. (1984, December). *Group differences on structured tests.* Paper presented at the Advanced Study Institute, Athens.

———. (1988). Systematizing the item content in test design. In R. Langeheine & J. Rost (eds.), *Latent trait and latent class models.* New York: Plenum, pp. 291–307.

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

———. (1985). A comment on McCauley and Colberg's conception of cross-cultural transportability of tests. *Journal of Educational Measurement* 22:157–161.

Van de Vijver, F. J. R., Daal., M., & Van Zonneveld, R. (1986). The trainability of formal thinking: A cross-cultural comparison. *International Journal of Psychology* 21:589–615.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika* 47:123–140.

Van der Flier, H. (1972). Evaluating environmental influences on test scores. In L. J. Cronbach & P. J. D. Drenth (eds.), *Mental tests and cultural adaptation.* The Hague: Mouton, pp. 447–452.

———. (1980). *De vergelijkbaarheid van individuele testprestaties.* Dissertation. Lisse: Swets & Zeitlinger.

Van der Flier, H., Mellenbergh, G. J., Adèr H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement* 21:131–145.

Vernon, P. E. (1979). *Intelligence: Heredity and environment.* San Francisco: Freeman.

Whiting, B. (1976). The problem of the packaged variable. In K. F. Riegel & J. A. Meacham (eds.), *The developing individual in a changing world.* The Hague: Mouton, pp. 303–309.