Cross-Cultural Assessment

van de Vijver, F.J.R.

Published in:
Applied Psychology: An International Review

Publication date:
2002

Document Version
Peer reviewed version

Cross-Cultural Assessment: Value for Money?

Fons van de Vijver

Tilburg University, the Netherlands

Abstract

Psychological assessment increasingly involves the application of tests in different cultural contexts, either in a single country (involving migrants) or in different countries. After a description of the main characteristics of such cross-cultural assessment, its costs and benefits are addressed. It is argued that these vary for the stakeholders involved: clients, persons or institutions hiring psychological expertise, psychology as a profession, and society at large. Although a comprehensive analysis of costs and benefits would require empirical study, it can be safely concluded that in the long run all stakeholders gain from cross-cultural assessment. In the near future the demand for cross-cultural assessment will increase, due to the growing internationalization of business and the increasing need of migrant groups for culture-informed psychological services. The role of psychologists in this process is twofold: we need explicit standards of appropriate cross-cultural assessment and we should communicate these standards effectively to all stakeholders in order to develop and maintain a high quality of testing and service delivery.

Cross-Cultural Assessment: Value for Money

Cross-cultural research is thriving. There is a consistent increase of publications dealing with cross-cultural issues. In PsycLit, the electronic database of psychological publications, the number of publications dealing with cross-cultural differences has grown in the last ten years, both in absolute number and in its relative contribution to the total number of publications in the database. This trend is not surprising and is undoubtedly fueled by societal developments of the last decades. A more globally operating economy, internationalization of education, large labor migration streams, and massive numbers of refugees have given cross-cultural encounters a more prominent place in various societies. A report by the International Labour Organisation, issued in March 2000, predicted further increases in international labor migration; more specifically, if the disparity in affluence between the rich and poor countries continues to grow as in the last decade, an increase in labor stream from the poor to the rich countries can be expected.

The consequences for psychologists are variegated. The present paper focuses on assessment; innovations are needed to accommodate the growing number and changing nature of test applications in different cultural groups. For example, selection and personnel psychologists need to develop and administer tests for use in multicultural settings as well as assessment procedures for expatriates; in clinical assessment there is a need for culture-informed instruments; school psychologists may want to assess the language development of bilingual children simultaneously in both languages. In general, three types of instruments can be envisaged in cross-cultural

assessment. The first comprises instruments with a known reliability and validity in western groups. To what extent instruments retain these properties after translation has to be determined empirically. Personality questionnaires by Eysenck are examples of instruments that have been translated and validated in various cultural groups (e.g., Barrett, Petrides, Eysenck, & Eysenck, 1998; Eysenck, Barrett, & Eysenck, 1985). “Blind” applications of western instruments in new cultural populations, in which there is no concern for the applicability of the instrument and (the establishment of) the psychometric properties in the new context, are not considered here, simply because they constitute bad practice.

Second, new instruments can be developed that are designed to function in a cross-cultural context. This was the idea behind the so-called “culture-free” (Cattell, 1940), “culture-fair” (Cattell & Cattell, 1963), and, more recently, “culture-reduced” tests (Jensen, 1980). The claim that there are psychological tests that are not affected by cultural factors has been criticized (e.g., Frijda & Jahoda, 1966). Still, the idea that some tests are more suited for cross-cultural assessment than others because of particular features such as their format, mode of administration, or item contents, still underlies much test design and data analysis in cross-cultural psychology.

Third, when existing instruments are invalid, unreliable, or do not cover the target construct in other cultural groups, culture-specific instruments need to be developed (e.g., Cheung et al., 1996). The instruments may be newly assembled or based on minor or major adaptations of existing tests. The Minnesota Multiphasic Personality Inventory provides a good example of the latter; the instrument contains various implicit references to the American

culture of the test designers and extensive adaptations are often required before it can be used in other languages and cultures (e.g., Lucio, Reyes-Lagunes, & Scott, 1994).

The present paper uses cross-cultural assessment as a generic term for all of the above applications. So, cross-cultural assessment refers here to all issues arising in the application of psychological instruments, either in a single country in the assessment of migrant groups, or in the assessment of individuals from at least two countries. This may involve the application of the same test in various cultural groups or different culture-specific tests. It is essential that the tests used have demonstrated their appropriateness (reliability, validity, and equivalence) in all cultural groups involved. Cross-cultural assessment refers not just to the instrument and its characteristics, such as item rubrics and reliability, but also to the application of norms (if applicable) and all other issues arising in the usage of tests involving at least two different cultural groups.

The present article addresses and important aspect of cross-cultural assessment: Is it cost-effective? Cross-cultural assessment has various costs. Norms of new tests may have to be made or existing norms may require adaptations for cultural groups that were not included in the original norm sample. Establishing norms in a multicultural group may require complex sampling schemes and logistics in order to examine the performance of various ethnic groups. The cost of cross-cultural assessment has always to be compared to the considerably easier and cheaper solution in which the assessor acts as if the administration procedure and norms of an instrument as previously established in one cultural group, can be seamlessly transferred

to another group. So, the question can be asked as to whether cross-cultural assessment adds sufficient incremental value to the bad, but common practice of straightforward applications of existing tests and their norms to warrant the additional effort.

An examination of these costs is preceded by a description of the theoretical background of cross-cultural assessment; the core concepts in cross-cultural assessment, bias and equivalence, are discussed in the next section. The third section places assessment in the framework of game theory and identifies four players (stakeholders): the client, the person or institution hiring the psychological expertise, psychology as a profession, and the larger society. Costs and benefits of cross-cultural assessment for all players are discussed in the fourth part. It is argued that benefits outweigh costs for all stakeholders. In the fifth part trends in cross-cultural assessment are described. Finally, conclusions are drawn.

Cross-Cultural Assessment

Main Issues

From a theoretical perspective the most characteristic features of cross-cultural assessment are bias and equivalence. An item or test is biased if it does not measure the same psychological propensity (trait, ability, attitude) across cultural groups; bias challenges the construct validity of an item or test. An empirical example, given by Hambleton (1994, p. 235), is the test item "Where is a bird with webbed feet most likely to live?" The Swedish translation of the English "bird with webbed feed" into "bird with swimming

feet" provides a much stronger clue to the solution than the English original item.

Equivalence refers to the level at which item or test scores can be compared across cultural groups, and, hence, to the calibration of scales. If there is no bias, scores are equivalent across cultural groups. Equivalent scores can be compared, both across and within groups. A score of 10 on an unbiased scale for depression has the same psychological meaning in all cultural groups studied. Bias and equivalence are closely related concepts. Bias refers to factors that show a differential impact on scores in cultural populations, while equivalence involves the implications of bias on the scope for comparing scores. If bias occurs, the equivalence of scores is challenged. Scores of American and Swedish pupils on the item about swimming feet were incomparable. (For a more extensive description of equivalence, see Poortinga, 1989, and Van de Vijver & Leung, 1997a, b.)

Three types of bias can be distinguished (cf. Van de Vijver & Leung, 1997a, b; Van de Vijver & Poortinga, 1997). The first is construct bias. A measure shows construct bias if the construct measured is not identical across cultural groups. Identity can be challenged by a lack of overlap in behaviors associated with the construct in the cultures studied. An example comes from research in personality on the Five-Factor Model. McCrae and Costa (1997) found considerable evidence for the universality of the structure as found in US subjects, among German, Portuguese, Hebrew, Chinese, Korean, and Japanese samples. However, Cheung et al. (1996) found that the Five-factor model leaves out aspects of psychological functioning that are deemed salient by Chinese. Interpersonal factors like “harmony” and “losing

face” are frequently observed in free person descriptions in the latter group, but are not represented in the Five-Factor Model. The Chinese study shows that this model may well be universal (meaning here the cross-culturally “greatest common denominator of personality”), but it may not be comprehensive.

Method bias refers to the presence of nuisance variables due to method-related factors. Three types of method bias can be envisaged. First, incomparability of samples on aspects other than the target variable can lead to method bias (sample bias). For instance, cultural groups often differ in educational background and, when dealing with mental tests, these differences can confound real population differences on a target variable. Method bias also refers to problems deriving from instrument characteristics (instrument bias). A well-known example is stimulus familiarity. Deregowski and Serpell (1971) asked Scottish and Zambian children in one condition to sort miniature models of animals and motor vehicles, and in another condition to sort photographs of these models. Although no cross-cultural differences were found for the actual models, the Scottish children obtained higher scores than the Zambian children when photographs were sorted. A final type of method bias arises from administration problems (administration bias). Communication problems between testers and testees (or interviewers and interviewees) can easily occur, especially, when they have different first languages and cultural backgrounds (cf. Gass & Varonis, 1991). Interviewees´ insufficient knowledge of the testing language and inappropriate modes of address or cultural norm violations on the part of the interviewer can

seriously endanger the collection of appropriate data, even in structured interviews.

At first sight it may seem counterintuitive to include sample and test administration features in a list of biasing factors. However, it should be kept in mind that bias is not an instrument property but a characteristic of a cross-cultural comparison of an application of an instrument; as a consequence, instrument properties can induce bias, but so can participants' characteristics and the way a test is administered.

Item bias is the last type of bias. An item of, say, an anxiety scale is said to be biased if persons with the same level of trait anxiety, but coming from different cultures, are not equally likely to endorse the item. Reasons for such differential response patterns may be, among other things, poor translations or inappropriateness of item contents (such as in the example about the swimming feet, mentioned before). As a hypothetical example, the item "Are you afraid when you walk alone on the street in the middle of the night?" may be responded to differently by persons depending on the safety of their neighborhood, even when the persons would have equal total scores on the questionnaire. This type of bias, also known as differential item functioning or DIF (Berk, 1982; Holland & Wainer, 1993), has been extensively studied by psychometricians.

Stakeholders

When cross-cultural assessment, as is often the case, involves migrants, it is sometimes called multicultural assessment. The present evaluation of cost-effectiveness primarily deals with this type of cross-cultural

assessment, although the reasoning that follows also applies to cross-national research. Cross-cultural assessment can be seen as a complex enterprise involving four different stakeholders, each with their own interests. Depending on the context of application, the first is called the client, patient, participant, or applicant. The group of research participants is less relevant here because of their typically small personal interest in the outcome of the assessment. Members of mainstream groups also belong to this category, although they often do not constitute the target population. The clients may ask for vocational guidance, clinical assessment, or they may be eligible for some desirable treatment (e.g., job application or school admission).

The second party is the person or institution hiring the psychological expertise for some purpose. The client may be the hiring party, but in regular selection situations, the two parties are clearly distinct and have different, often incompatible interests in the assessment procedure.

The psychological profession is the third stakeholder. In principle practitioners and researchers of all branches of psychology are represented here. The last stakeholder is the larger society. The interest of society at large should not be underrated. Like all societies, pluralistic societies have a vast interest in the well-being of their members; issues such as intergroup relations and access of cultural groups to institutions such as education and the labor market are important elements in this well-being.

Costs and Benefits of Cross-Cultural Assessment: Traditional Views

Traditionally, the utility of cross-cultural assessment has been associated with fairness (e.g., Cronbach, 1984; Hunter, Schmidt, & Hunter,

1979; Petersen & Novick, 1976; Sackett & Wilk, 1994; Schmidt, & Hunter, 1977). The fairness approach puts an emphasis on predictive validity (i.e., (in)equality of regression lines of majority and minority groups). The main outcomes of the tradition are twofold. First, various definitions of and approaches to fairness have been developed (e.g., Petersen & Novick, 1976). For instance, in quota hiring the proportion of applicants to be hired from each cultural group is agreed upon prior to the selection procedure. It is the task of the selection officer to select the best applicants for each group. As an alternative, in an “equal marginal risk” model the person with lowest score from each group hired has the same probability of success in the job. The model produces the largest number of successful workers (cf. Cronbach, 1984, p. 389). If the regression lines describing the relationship between predictor and criterion are identical, this amounts to “color-blind hiring” (i.e., hiring the applicants with the highest score, irrespective of cultural background). Second, the problem of “validity generalization”, the seemingly erratic fluctuations of predictive validity coefficients across independent studies, has been solved (Schmidt, & Hunter, 1977). A combination of results, applying meta-analytic techniques, usually points to the similarity of regression lines of majority and minority groups, thereby (seemingly) providing an empirical basis for “color-blind hiring”.

There is another line in the literature in which fairness issues are treated in a different way. These approaches involve score adjustments of members of minority groups. Various procedures have been proposed, such as within-group norming (i.e., establishing norms per cultural group) and bonus points. As an example of the latter, Mercer (1979; see also Cronbach,

1984, p. 211ff.) designed a system for “correcting” test scores of a child (such as scores on the WISC) based on information of the socioeconomic background of the child. Scores of White children were typically shifted downwards, while scores of Hispanic children and Black children were boosted by the “correction”. Another example is the sliding band (Cascio, Outtz, Zedeck, & Goldstein, 1991). The procedure defines score bands, beginning with the top score. The band consists of all scores that do not differ significantly from the top score. Within the band all observed scores are assumed to reflect equal proficiency for the prospective job. Within this band, sores are not significantly different from the highest score. Minority applicants of the band are then selected first, followed by the choice of majority group members with the highest score. The band is then ”slided” to a lower score (one less than the top score) and the procedure is repeated until the target number of chosen applicants is reached. If (and only if) members of minorities have scores within bands of eligible candidates, they get a preferential treatment.

Both validity generalization and score adjustment procedures show problems. Validity generalization studies suffer from two problems. The first is the exclusive focus on predictive validity and the underrating of the importance of equivalence issues. In a study involving migrants (mainly of Turkish, Moroccan, and Caribbean descent) and autochthonous applicants Netherlands, te Nijenhuis (1997) reported similar results among applicants of blue-collar jobs at the Dutch railways. When it is realized that the level of mastery of the Dutch language varies across mainstream and migrant applicants, his findings may seem counterintuitive despite their

correspondence with Schmidt and Hunter's. From a cross-cultural perspective these studies are characterized by a one-sided emphasis of predictive validity and an insufficient recognition that common bias sources may challenge both the predictor and the criterion. The reason for the similarity of predictive validity coefficients in groups of migrants and mainstream Dutch can be easily seen if it is realized that the lower linguistic proficiency of the Turkish group affects both current test performance and later job success. In other words, similarity of regression lines across groups does not yet make a test equivalent across groups. In particular method bias should be scrutinized. In a soon to be defended PhD thesis, Helms-Lorenz (2000) reports substantial generational increases in IQ among primary-school children of Dutch migrant groups. Because such a jump was not matched by a similar increase among the Dutch children, it is likely that method bias played an important role and that the shift has to be accounted for by factors like education and better mastery of the Dutch language.

There is another problem with (American) validity generalization studies (but similar results may also apply elsewhere). On mental tests the difference between Blacks and Whites is 1 SD, while in actual job performance the difference is smaller, often not more than .5 SD. Some important source of intergroup score differences at the predictor is not at represented in the criterion. If we would use a job tryout as predictor, then more minority applicants would be hired. Now, bias is defined as differences in scores on an instrument that do not generalize to a domain of interest (Van de Vijver & Leung, 1997a, b). This smaller score difference on a target (job success) than on a source (mental test) would clearly qualify as bias. More

specifically, it would be method bias as bias derives from the test method (the application of a cognitive test).

The score adjustment tradition has been challenged on psychometric grounds. Thus, Cronbach criticized Mercer's score corrections because no data were provided indicating that the "corrected" scores showed a higher validity; as an example, no data were provided to demonstrate that the "corrected" scores better predict school performance or more adequately reflected the intellectual abilities. In a similar vein, Schmidt (1991) criticizes the usage of sliding bands:

> [The usage] appears to be based on the belief that if two scores are not statistically significantly different, then the best estimate is that they are equal. This belief is incorrect: Regardless of statistical significance, the statistically best estimate is always that the individual with the higher obtained score has the higher true score. Therefore, if the test has criterion-related validity, the statistically best estimate is that the individual with the higher obtained score will have the higher job performance. (Schmidt, 1995, p. 267)

He concludes that "all banding models are simply attempts to reduce adverse impact while minimizing utility losses" and that banding is "statistically (and thus economically) suboptimal" (p. 266).

In my view, the discussion about fairness and score adjustments is hampered by the ambiguity of the role of psychometrics in both traditions. In the score adjustment tradition there seems to be an assumption that adjustments should have a psychometric justification. The question of the need for score adjustments refers to societal injustice, due to the skewed

distribution of money and resources across the different layers of a society. The perceived desirability of correcting (or not correcting) for these differences cannot be justified on psychometric grounds; more importantly, no psychometric rationale is needed. Whether or not a society will legalize and accept affirmative action is the outcome of a complex interaction of relevant stakeholders such as political parties, representatives of the justice system, employers, employees, and minority interest groups. The only contribution of psychometrics to this debate can and should be the specification of models to identify bias and group-dependent scoring schemes. Implicit rules of conduct or explicit legislature, inspired by the public debate, define the typically narrow operational boundaries of psychometric procedures.

The role of psychometrics in the fairness tradition also deserves some comments. This tradition capitalizes on economical utility (cf. Zedeck, Outtz, Cascio, & Goldstein, 1991), thereby overrating the interest of the hiring party in cross-cultural assessment. As will be shown below, all stakeholders involved in cross-cultural assessment have their interests; as a consequence, the utility of this assessment is the sum of the utilities of all parties involved.

Cost-Effectiveness for Stakeholders

Let us take a closer look at the interests of all stakeholders involved. The cost-effectiveness for clients can be easily determined. The efforts required are not large; a cross-cultural assessment procedure may entail more time and tests to be administered than conventional testing procedures (Table 1). Various factors can influence the length of the testing time: test instructions may be longer, more examples may be included, and the

administration of additional instruments may be required. The benefits for clients have to do with the increase in level of service delivery and the possibly higher validity of inferences based on the test scores. The benefits for the group of clients are substantial, though at individual level the procedure may mean that a person is not hired.

The cost of cross-cultural assessment for the commissioning institution is financial, because such a procedure may take longer and the institution may have to indirectly pay for the research to validate the instruments utilized. If an existing assessment procedure would be poor, the cost of poor selection should also be taken into account, such as turnover, additional training, a larger load on supervisors; models to estimate such costs have been described by Cascio (1982, 1987). As a benefit, the institution may get a more valid picture of the abilities or personality of an applicant. Moreover, in some countries, such as the USA, evidence based on cross-culturally validated instruments constitute acceptable evidence for judges and jurors and hence, the usage of these tests can lead to decisions that can be defended in court.

Both the costs and benefits of cross-cultural assessment for psychology as a profession are multifaceted. The costs are considerable: There is a need to develop adequate assessment procedures; norm tables established in a mainstream group, may no longer apply; the quality of widely used tests has to be empirically examined in groups not included in the norm study, et cetera. In general, many aspects that can be taken for granted in test administrations to individuals from the mainstream group may no longer hold and it requires investments in the form of new research to examine bias in and equivalence of these instruments.

The main benefit of cross-cultural assessment for psychology as a profession is the higher level of quality of service delivery. Actual and by external parties perceived quality standards benefit from sensitivity to and knowledge of the multicultural nature of western societies. Claims that we can deal with cultural heterogeneity can be more validly made. Decisions based on instruments that are validated in a cross-cultural context can be easier defended in court, even if such a practice does not yet imply that every case leads to a for psychologists favorable outcome. Finally, research into cross-cultural assessment has had and may continue to have various spin-offs, such as the development of various statistical techniques to assess item bias and equivalence, deeper insight in the psychological characteristics of various groups and, generally, in the role of culture in human behavior.

The societal costs and benefits of cross-cultural assessment are not clearly defined. There are potential costs, which may materialize in a particular case. Cross-cultural assessment is typically less transparent and accessible to society than conventional assessment. In multicultural societies the use of tests that have been developed with an open eye for the linguistic and cultural aspects of the target groups, may convey the message of a favorable or unfavorable treatment of some cultural groups. Proper communication about these procedures is important. A psychological attitude by the general public against unequal treatment may adversely impact on race relations. It is doubtful, however, whether cross-cultural assessment alone suffices to create such a climate in society; it is more likely that in a society in which there is little support for equal opportunity programs, cross-cultural assessment will not be strongly supported. Conversely, in a society

with support for a multicultural policy, cross-cultural assessment will not be controversial.

Some types of cross-cultural assessment do not compare scores across cultures (e.g., clinical assessment as part of therapy intake) and it is difficult to see why there would be any objections against this noncomparative test usage in any society.

The benefits for society come from at least two sources. First, cross-cultural assessment better reflects the daily reality of multicultural societies and increasing internationalization (globalization) of these societies. Second, a society benefits when its members better realize their potential and cross-cultural assessment can (admittedly only a little) help to reach this goal.

In addition to the costs and benefits of cross-cultural assessment for all parties, one can also determine the costs and benefits of not using it. These costs and benefits are the opposite of the aspects mentioned before. For example, the continued use of tests with an unknown validity is cheap (as there are no development costs) but does not reflect the multicultural nature of society. Considering the cost of not using cross-cultural assessment may become more common when this type of assessment is more widespread. For example, the cost for all parties is more visible when adequate cross-cultural tests are available but not used.

Cross-Cultural Assessment as a Four-Party Game

What can be concluded from this complicated picture? The first obvious conclusion is that not all stakeholders and individual members of these constituencies have convergent interests. Conflicts of interest can occur

between stakeholders; for example, the development of possibly costly cross-cultural assessment procedures creates jobs for psychologists but has to be paid for by hiring institutions. The interest of the hiring party and society can also conflict. If a law enforces affirmative action, some employees benefit while others “pay” for it. Another example can be found in Van Beek’s (1993) work. He found that Dutch employers are unlikely to hire members of migrant groups, even when the groups were matched on several job-relevant characteristics, such as education. Apart from the question whether this tendency points to discrimination or a realistic anticipation on the problems with foreign employees, a serious societal problem is created when most or all employers follow this practice. It blocks the entrance of foreign groups to the labor market, with unavoidable consequences for the income and economic opportunities of these groups, thereby eventually creating potential sources of interethnic tensions. If many employers follow this practice, the larger society is the big loser, because it has to pay the bill for all costs due to these tensions. That Van Beek’s findings are still valid, is indirectly illustrated by labor market developments of the last years. During the last years the Dutch economy has gone through a period of rapid expansion. Many new jobs have been created. The expansion of the labor market has not affected all kinds of jobs equally; the number of high-skilled jobs grew much faster than did the number of low-skilled jobs. As a consequence, many migrant workers (with often little schooling) did not gain at all from the economic spurt. Ironically, the strong growth of the latter years has increased the difference of relative positions of natives and migrants.

The economic growth of the last years has led to a small, bit clearly noticeable migration stream. Some companies who experience problems in filling high-skilled positions recruit personnel from elsewhere. The free movement of labor within the borders of the European Union provides a legal basis for this practice. A second group of migrants comes from countries like South Africa. They are often descendants of Dutch ancestors and have a Dutch passport. To my knowledge no systematic study of the assessment procedures of these groups have been carried out. The combination of a cultural background that is relatively close to the Dutch culture and a high education, characteristic for this groups, makes it unlikely that issues of fairness and bias will play an important role in their assessment.

Conflicts of interests in cross-cultural assessment can also occur (and are probably even more likely) among individual members of the constituencies. The interests of the minority and majority group members can be antithetical in cross-cultural assessment procedures.

Game theory is proposed here as a framework to examine the actions and decisions of all stakeholders. Game theory distinguishes between various types of games (e.g., Dixit & Skeath, 1999; Fink, Gates, & Humes, 1998). The type that comes closest to cross-cultural assessment involving the four stakeholders is a four-party “chicken” game, a slight variation on the well-known prisoners’ dilemma game. The name is derived from a game that was played in the fifties by American teenagers. Two persons drive their cars toward each other. The first to swerve loses the game and becomes the “chicken”. If both decide to swerve, there is no gain; if both decide not to swerve there is loss for both (car damage and personal injuries). There is only

differential gain and loss for both parties if only one party swerves. Cross-cultural assessment has a payoff matrix that is somewhat comparable to a chicken game (Table 2). The four players of cross-cultural assessment have the choice to favor (to swerve) or disfavor cross-cultural assessment. Each choice has advantages and disadvantages. Based on the assumption of rational behavior by the players, game theory predicts that each party will choose the option(s) that maximize(s) its own payoff, taking into account that all other parties also maximize their output. If all cross-cultural assessment would be compulsory (strictly enforced) or all parties would favor it, then there will be no differential gain incurring from applying cross-cultural assessment for any employer or psychologist (although, obviously, all stakeholders would gain). A typical feature of chicken games is the large disparity in payoff for parties when one party chooses to cooperate (applying a cross-cultural test) while other parties prefer make the opposite choice. In this case the latter players get a huge benefit, which is paid for by the party making the cooperative choice (e.g., by developing a cross-cultural test). If some employers decide to ask psychologists to develop cross-culturally validated tests, the “freerider” who uses these tests without having contributed to the costs (e.g., by illegal copying) receives a large benefit.

Characteristic of a multiparty chicken game is the payoff function of Figure 1. If this would display the payoff for employers, it would mean that when hardly any employer favors cross-cultural assessment, it becomes attractive for an employer to make a deviant choice. It can give a lot of free publicity and create goodwill for the employer if he/she is prepared to hire a migrant. From the perspective of the employee, being hired as the only

migrant in a company may be a mixed blessing, particularly when colleagues see the decision to hire the migrant not as being based on merit. When most employers apply cross-cultural testing, it is (again) interesting for the individual employer to make the deviant choice, because he/she may not to pay the cost of a more expensive assessment procedure. Application of this type of assessment will lead to some profit for most parties. The reluctance of some hiring parties to implement cross-cultural assessment suggests that the perceived benefit is small, if present at all. Moreover, the defective choice in the chicken's game, amounting to a refusal to apply cross-cultural testing and to persevere in applying conventional tests with unknown applicability in a cross-cultural context, is likely to lead to a negative payoff for all players. It is clear that for some individual members of all stakeholders the option not to advocate the use of cross-cultural assessment serves their interest well. As a consequence, there is a need to develop guidelines for appropriate cross-cultural testing. Depending on the country and customary way to settle conflicts in these areas, either national laws or rules by national psychological organizations may be more appropriate. Sackett and Wilk (1994) discuss the legal environment of score adjustments during the last 15 years in the USA. Their case description illustrates how the recommended practice for psychologists has been influenced by a book by a National Academy of Sciences committee that supported quota hiring, followed by a public outcry against this practice, and a subsequent provision in the civil Rights Act of 1991 that reflected the public opinion. Court cases in latter years led to a further specification of how the Act should be interpreted.

As illustrated, game theory provides a useful analogy to understand the dynamics of cross-cultural testing; yet, its limitation should be acknowledged. In game theory players should have the opportunity to choose among all options available. This assumption does not hold true for all parties in cross-cultural assessment. Theoretically speaking, a migrant may challenge or even sue a psychologist for not using appropriate tests. In practice, however, lawsuits for such bad practice are uncommon in many countries.

Another problem in applying a game-theoretical perspective on cross-cultural assessment is the partial incommensurability of the interests of the stakeholders, such as the material costs of the employers and the increased quality of service delivery of psychologists. By estimating all costs and benefits it would be fairly easy to determine the overall cost-effectiveness of cross-cultural assessment. These estimated costs and benefits would also allow to fill the cells of the payoff matrix of the chicken game as displayed in Table 2 (extended to four parties) with realistic values, thereby enabling the prediction of strategies vis-à-vis cross-cultural assessment of all parties. However, even without such empirically based cost estimates, it is fairly obvious that the relative costs for all stakeholders are low and that the interest of no stakeholder is put in jeopardy by favoring cross-cultural assessment, while the benefits are considerable. The statement that the relative costs are low does not mean that these costs are always low in an absolute sense. On the contrary, the development and validation of a large selection battery that should cater for many different cultural populations can easily become a multimillion dollar venture. From the perspective of the hiring party (who will

have to do the investment), the cost-effectiveness of such an investment has to be related to the possible costs of malpractice lawsuits, material or immaterial judicial penalties or settlements, and publicity around such lawsuits.

Trends in the Near Future

The main costs and benefits of cross-cultural assessment for the four stakeholders are not susceptible to strong temporal fluctuations and do not differ considerably across western countries. As a consequence, there may be a firm basis for a continued interest in cross-cultural assessment:

1. Western societies continue to be multicultural and the issue of valid assessment in culturally heterogeneous groups will only become more prominent.
2. The traditionally assimilationist perspective on acculturation will give way to pluralistic outcomes. In various western societies particular ethnic groups have become or will become so big, that they can easily establish and maintain their own institutions (such as schools and churches). In sociology this is called “institutional completeness” and involves the presence of all institutions that are needed for cultural transmission. A good example can be found in large American cities, such as Los Angeles and San Francisco. If the projected demographic trend would borne out, these cities will have a larger Hispanic than Anglo-American population in twenty years. When cultural groups are so large, they will not experience much pressure to assimilate. Rather, maintaining their own cultural

identity (separation) or pursuing a combination of the original and mainstream identity (integration) become viable options. Consequently, migrant groups will increasingly call for culture-specific services.

3. The scientific interest in cross-cultural comparisons will continue. In the past much cross-cultural research was carried out by researchers who devoted at least a substantial part of their career to these topics; nowadays the vast majority of research is carried out by “sojourners,” researchers who are interested in cross-cultural research as an extension of their intracultural work, often for not more than one or two projects (Van de Vijver & Leung, 2000). Although the currently increasing interest in cross-cultural studies may level off in the future, it can be expected that cross-cultural research will become part of all major paradigms of psychology and that an adequate test of a theory or assessment device will always include the study of cross-cultural variation.
4. The internationalization of business life will continue. The globalization of the economic market and the increasing number of internationally operating companies will lead to growing numbers of sojourners and expatriates. During the last decade implications of these developments were clearly visible, such as the development and implementation of intercultural communication training and of assessment procedures for sojourners and expatriates. Further developments in these areas and the establishment of a firmer

empirical base underlying these procedures can be expected in the future.

All these trends point in the direction of an increasing prominence of cross-cultural assessment. Even if not all trends would materialize, this assessment will still become more important and widespread. Cross-cultural assessment is also facilitated by the introduction of guidelines for appropriate test usage, such as the AERA/APA/NCME Standards (2000) and by the introduction of ratings of the cross-cultural applicability of an instrument in test qualification systems (e.g., Bartram & Coyne, 1999; information about the international project on test use guidelines can also be downloaded from the Website of the International Test Commission at http://cwis.kub.nl/~fsw_1/itc).

In conclusion, cross-cultural encounters will become more frequent and prominent in the near future. The interest in and societal relevance of cross-cultural assessment will continue to grow. Cross-cultural assessment is coming of age. These developments create new challenges for cross-cultural measurement and psychology as a profession, which are explored in the next section.

Enhancing the Quality and Impact of Cross-Cultural Assessment

In order to insure that the increased interest in cross-cultural assessment will make a positive and lasting contribution both to psychology as a discipline and society at large, two types of measures should be taken by those involved in cross-cultural assessment. The first focuses on the discipline itself and is aimed at enhancing the quality of assessment and service delivery (internal function). The second is more externally oriented

and refers to the communication of these measures to other parties involved in cross-cultural assessment (external function).

The internal function is aimed at enhancing the quality of assessment and service delivery. The quality of cross-cultural assessment will benefit from the stipulation of quality standards (or in a less prescriptive terminology, recommendable practices). In addition to country-specific topics, these standards should contain at least three elements:

1. Psychometric standards, focusing on equivalence and the absence of bias, in addition to common norms of reliability and validity;
2. Standards of appropriate test usage, specifying, among various other things, what to do when no relevant norm data for an instrument are available;
3. A description of the roles of each of the four stakeholders of cross-cultural assessment (client, hiring party, psychologist, and society). Judicial aspects of this type of assessment may require special attention.

There are two recent examples of sets of assessment standards that explicitly deal with cross-cultural testing. The first is the new version of the AERA/APA/NCME Standards (2000), in which there are chapters on bias and equivalence. Relevant issues of testing in a multicultural context are discussed. The introduction of standards of assessment in a multicultural context in the present version is an important step in achieving a uniform set of rules to apply to such assessment. From an international perspective, the main limitation of the new standards is their focus on the USA. Another example can be found in the work of an international committee, formed on

the initiative of the International Test Commission (ITC) and headed by Ron Hambleton (University of Amherst, MA) (Hambleton, 1994; Hambleton, Merenda, & Spielberger, 2000; Van de Vijver & Hambleton, 1996). A set of 22 guidelines have been formulated, describing the issues in cross-cultural assessment in more detail and paying more attention to the specification of practical suggestions on how to meet the guidelines. Both the AERA/APA/NCME Standards and the ITC Guidelines provide good templates for other countries that can either be adopted in an integral way or adapted to meet the local needs.

In the Netherlands some tests have been developed, explicitly aimed at application in a multicultural population. For example, the Leertest voor Etnische Minderheden (Hessels, 1993, 1996) is a measure of learning potential for Turkish and Moroccan children from five to eight years. The test has been constructed with the aim to minimally rely on knowledge of the Dutch language and culture (in many cases these children begin to learn Dutch when they go to Kindergarten). Norms for Turkish, Moroccan, and native Dutch children are available. The RAKIT, a Dutch intelligence for children at primary school age, is another example. Norms were already available for native children. New norms have been established for Moroccan and Turkish children.

The availability of standards of appropriate cross-cultural assessment will facilitate the second (internal) task of psychology: we have to determine to what extent psychological instruments and common assessment practices follow these standards. In various countries there is an urgent need to review the adequacy of commonly employed psychological instruments for use

among migrant groups. Let us consider the Netherlands, which in this respect is a typical example. Systematic studies of bias in common psychological tests are fairly recent (Van de Vijver & Bleichrodt, 2000). There is still a long way to go before empirical data about the cross-cultural suitability of all commonly applied psychological tests will be available. It is realistic to assume that at least some of these instruments will show so many sources of bias and inequivalence that their suitability is seriously diminished. The results of our stock-taking exercise will be an increased awareness of the need to develop new instruments.

The effectiveness of cross-cultural assessment is enhanced when in addition to the internal role that was described above, the external role (i.e., communication with society) is also taken seriously. Psychologists will have to communicate to all parties involved (clients, hiring party, and the larger society) what we mean with appropriate tests and testing, what we can and cannot do with cross-cultural assessment, and what we mean with standards of good practice.

There are two problems to overcome in this communication. The first is that cross-cultural assessment of in particular intellectual abilities has started a public debate at various times in the past, most recently in the discussion about Herrnstein and Murray's (1994) book "The bell curve." It does not require much deliberation to conclude that psychology as a discipline is the big loser in such public debates. The implicit message to society is unambiguous: psychologists do not agree on these topics and they are not able to settle the issue on scientific grounds, thereby creating an atmosphere of lack of professionalism that extends far beyond the topic of the debate.

Emphasizing the need for standards may help to overcome the resistance of a not very receptive audience.

The second problem is the external pressure to meet unrealistically high standards. In order to be competitive, psychologists may be tempted to overplay their hand and to promise more than they can deliver. The limitations of cross-cultural assessment are numerous: in most countries only a few tests (if any at all) have been scrutinized for suitability in cross-cultural assessment, for some small cultural groups almost no validated tests may be available, we are still largely unable to take acculturation status into account in dealing with test scores, even our best test procedures are not at all culture-free, psychological assessment only marginally affects (if at all) the relationships and mutual views of mainstream and minority groups, et cetera.

Conclusion

The question of the cost-effectiveness of cross-cultural assessment was addressed. It was argued that for all stakeholders involved (client, party hiring psychological expertise, psychologist, and larger society), the relatively small loss will be outweighed by the considerable potential gains, such as an increased level of service delivery. The major forces behind the current interest in cross-cultural assessment such as large migration streams and the internationalization of business life can be expected to remain prominent for quite some time. Therefore, cross-cultural assessment will become more important in the future.

When, like historians often do, we simplify a continuous development to a set of discrete events, hoping that the gain in clarity outweighs the loss of

accuracy, it can be conjectured that cross-cultural assessment is in a transitional stage. In the first stage cross-cultural psychology, borrowing heavily from mainstream psychology and cultural anthropology, was mainly an activity of a relatively small group of specialists. They shared dissatisfaction with the “acultural” nature of mainstream psychology, and its implicit view of pan-cultural generalizability of findings obtained in western laboratories, not infrequently obtained among anything but random samples of the general population. The study of human behavior without attention for the cultural factor is, to borrow Jahoda’s metaphor, like Hamlet without the prince of Denmark. In the second stage cross-cultural assessment becomes better integrated in psychology and most of the research in this area is not carried out by specialists, but by persons who have an interest in the topic without devoting their whole career to it. The domain of applications of cross-cultural assessment will probably broaden in the next decade. The number of cross-cultural studies will continue to grow. In order to facilitate this growth, it is important that we establish and communicate what we mean with adequate assessment and testing practice in cross-cultural assessment. The challenge is to meet the demands of “the market” while maintaining quality standards of “the discipline” to the extent possible. By adhering to such a set of standards the quality of research and service delivery will increase and the criteria to assess the quality will become clearer.

References

AERA/APA/NCME. (2000). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Barrett, P.T., Petrides, K.V., Eysenck, S.B.G., & Eysenck, H.J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. Personality and Individual Differences, 25, 805-819.

Bartram, D., & Coyne, I. (1998). Variations in national patterns of testing and test use: The ITC/EFPPA International Survey. European Journal of Psychological Assessment, 14, 249-260.

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

Cascio, W.F. (1987). Applied psychology in personnel management. Englewood Cliffs, NJ: Prentice Hall.

Cascio, W.F., Outtz, J., Zedeck, S., & Goldstein, I.L. (1991). Statistical implications of six methods of test score use in personnel selection. Human Performance,4, 233-264.

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

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

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

Cronbach, L.J. (1984). Essentials of psychological testing (4th ed.). New York: Harper & Row.

Deregowski, J.B., & Serpell, R. (1971). Performance on a sorting task: A cross-cultural experiment. International Journal of Psychology, 6, 273-281.

Dixit, A., & Skeath, S. (1999). Games of strategy. New York: Norton.

Eysenck, H.J., Barrett, P., & Eysenck, S.B. (1985). Indices of factor comparison for homologous and non-homologous personality scales in 24 different countries. Personality and Individual Differences, 6, 503-504.

Fink, E.C., Gates, S., & Humes, B.D. (1998). Game theory topics: Incomplete information, repeated games, and N-player games. (Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-122). Thousand Oaks, CA: Sage.

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

Gass, S.M., & Varonis, E.M. (1991). Miscommunication in nonnative speaker discourse. In N. Coupland, H. Giles, & J.M. Wiemann (Eds.), Miscommunication and problematic talk. Newbury Park, CA: Sage.

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

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (Eds.) (2000). Adapting educational tests and psychological tests for cross-cultural assessment. Mawhaw, NJ: Erlbaum. (forthcoming)

Helms-Lorenz, M. (2000). Cultural influence on cognitive test performance. PhD thesis, Tilburg University, the Netherlands.

Herrnstein, R.J., & Murray, C. (1994). The bell curve. Intelligence and class structure in American life. New York: Free Press.

Hessels, M.G.P. (1993). Leertest voor Etnische Minderheden. Theoretische en empirische verantwoording [Learning Potential Test for Ethnic Minorities: Theoretical and empirical background]. Rotterdam: RISBO.

Hessels, M.G.P. (1996). Ethnic differences in learning potential test scores: Research into item and test bias in the Learning potential test for Ethnic Minorities. Journal of Cognitive Education, 5, 133-153.

Holland, P.W., & Wainer, H. (Eds.) (1993). Differential item functioning. Hillsdale, NJ: Erlbaum.

Hunter, J.E., Schmidt, F.L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 721-735.

McCrae, R.R., & Costa, P.T. (1997). Personality trait structure as a human universal. American Psychologist, 52, 509-516.

Jensen, A.R. (1980). Bias in mental testing. New York: Free Press.

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

Mercer, J.R. (1979). Technical manual. System of Multicultural Pluralistic Assessment. New York: Psychological Corporation.

Petersen, N.S., & Novick, M.R. (1976). An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13, 3-29.

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

Sackett, P.R., & Wilk, S.L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. American Psychologist, 49, 929-954.

Schmidt, F.L. (1991). Why all banding procedures in personnel selection are logically flawed. Human Performance, 4, 265-277.

Schmidt, F.L., & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Te Nijenhuis, J. (1997). Bias in intelligence testing of immigrants in the Netherlands. Amsterdam: Free University.

Van Beek, K.W.H. (1993). To be hired or not to be hired, the employer decides: Relative chances of unemployed job-seekers on the Dutch labor market. PhD thesis.

Van de Vijver, F.J.R., & Bleichrodt, N. (Eds.) (2000). Diagnostiek bij allochtonen. Lisse, the Netherlands: Swets. (forthcoming)

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

Van de Vijver, F.J.R., & Leung, K. (1997a). Methods and data analysis of comparative research. In J.W. Berry, Y.H. Poortinga, & J. Pandey

(Eds.), Handbook of cross-cultural psychology (2nd ed., vol. 1). Boston: Allyn & Bacon.

Van de Vijver, F.J.R., & Leung, K. (1997b). Methods and data analysis for cross-cultural research. Newbury Park, CA: Sage.

Van de Vijver, F.J.R., & Leung, K. (2000). Methodological issues in psychological research on culture. Journal of Cross-Cultural Psychology, 31, 33-51.

Van de Vijver, F.J.R., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, 13, 29-37.

Zedeck, S., Outtz, J., Cascio, W.F., & Goldstein, I.L. (1991). Why do “testing” experts have such limited vision? Human Perfomance, 4, 297-308.

Author Note

Correspondence concerning this article should be addressed to Fons van de Vijver, Department of Psychology, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands; phone: +31 13 466 2528; fax: +31 13 466 2370; e-mail: fons.vandevijver@kub.nl

Comments on an earlier version of this article by Robert A. Roe are gratefully acknowledged.

The manuscript is based on an invited Address to Division 2 (Division of Psychological Assessment and Evaluation) of the International Association of Applied Psychology, San Francisco, August 10, 1998.

Table 1. Costs and Benefits of Cross-Cultural Assessment for All Parties Involved

Party	Costs	Benefits
Clients	Elaborate assessment may be required	Higher quality of service, increased validity
Hiring institution	Assessment procedure may be more involved and expensive	Higher face validity, though not necessarily higher predictive validity
Psychological profession	• Need to develop adequate assessment and decision procedures • Need for additional training in psychology courses • Various difficult assessment issues will arise, such as the question of which norms apply to an acculturating group in which individuals vary in terms of acculturation status • Quality of tests in multicultural settings has to be determined	• Higher level of professionalism • Leads to the development of instruments and practices that can be defended in court • Better insight in role culture in human behavior
Society	may “oversensitize” society for racial issues and adversely impact on racial relations	• Better reflects the daily reality of multicultural societies and growing internationalization • Society will benefit when its members can better achieve their potential

Table 2. Hypothetical Payoff Matrix in a Two-Person Chicken Game, Indicating Payoffs Associated with (Not) Favoring Cross-Cultural Assessment

		Party 2	
		Favor	Do not favor
Party 1	Favor	2, 2	1, 4
	Do not favor	4, 1	0, 0

Note. The numbers in the cells indicate the gains associated with a particular choice for the first and second party, respectively.

Figure Caption

Figure 1. Payoff in a multiplayer chicken game (after Dixit & Skeath, 1999, p. 365)