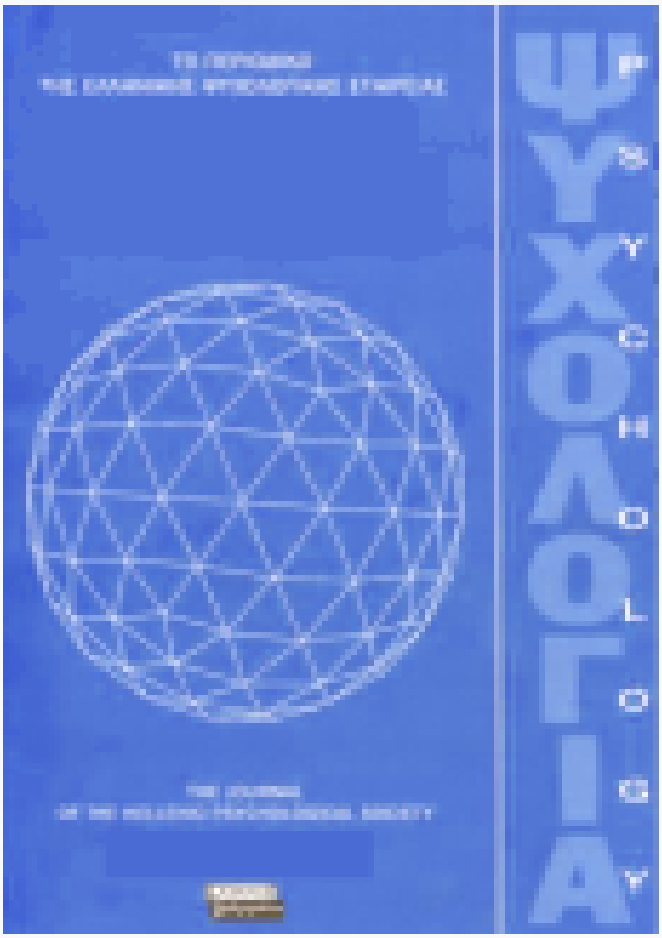
Psychology: the Journal of the Hellenic Psychological Society

Vol 9, No 2 (2002)



Key issues in cross-cultural assessment

Fons J. R. Van De Vijver

doi: 10.12681/psy_hps.24061

To cite this article:

Van De Vijver, F. J. R. (2020). Key issues in cross-cultural assessment. *Psychology: The Journal of the Hellenic Psychological Society*, *9*(2), 203–211. https://doi.org/10.12681/psy_hps.24061

Key issues in cross-cultural assessment

FONS J. R. VAN DE VIJVER
Tilburg University, the Netherlands

ABSTRACT Key issues in cross-cultural assessment are discussed. The concepts of bias and equivalence are described as pivotal aspects in cross-cultural assessment. A taxonomy of bias and equivalence is described. It is argued that issues in cross-cultural assessment are relevant not just in cross-national studies but in any translation and adaptation of western instruments, even when these are not meant for cross-national comparison. The problems to be dealt with in the process of such translations and adaptations are described on the basis of four hypothetical studies.

Key words: Bias, Cross-cultural assessment, Equivalence.

Recent developments in cross-cultural assessment

The often-heard truism that the world is becoming smaller may not strike the reader as reflecting a profound insight in the physical reality of our planet. Yet, when the truism is interpreted in its metaphorical sense, as referring to the psychological space, we touch upon a salient trend that impacts on many of us and will continue to do so in the future. Cross-cultural encounters are or become part of everyday life in Western societies. In some cases these are virtual encounters, such as the introduction of western technology in a nonwestern society, the adaptation of goods or services for a foreign market (e.g., changes of taste of drinks when exporting to other markets). In other cases there are real encounters between individuals with a different cultural background. A report by the International Labour Organisation, issued in March 2000, predicted 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. Other examples are the growing numbers of expatriates (persons working for a company with temporary assignments abroad), email communication, student exchange, tourism, etc.

These developments are studied in psychology, particularly in cross-cultural psychology. The areas in which cross-cultural psychology has to deal with the shrinking world are numerous but can be reduced to essentially two types of applications. The first involves cross-national applications, such as the translation of instruments and the adaptation of clinical treatments. The second involves the application of cross-cultural insights in multicultural societies. The latter are societies with members of various cultural backgrounds. All western societies are multicultural nowadays. For several of these (such as Scandinavia, Germany and the Benelux countries), multiculturalism is recently novel and they clearly need

Address: Fons J. R. van de Vijver, Department of Psychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Tel.: *31-13-4662528, Fax: *31-13-4662370. E-mail: fons.vandevijver@kub.nl

time to adjust their traditionally monocultural perspective. Applications of cross-cultural insights in these societies can amount to the design of adequate tests (e.g., in education), the assessment of acculturation, and the development of culture-sensitive treatments.

The present paper focuses on the implications of the "shrinking world" on assessment. In the first section, four hypothetical (but realistic) examples of studies are described in which cross-cultural assessment plays a role. Based on the issues emerging in the discussion of the examples, the two key concepts in this type of assessment, bias and equivalence, are introduced in the second section. In the third section I describe issues of test translations and adaptations, distinguishing between comparative and noncomparative test usage. Conclusions are drawn in the final section.

Four hypothetical examples

A psychologist from Zimbabwe, let us call her Joan, wants to design an intelligence test in Shona, the language spoken by many Zimbabweans, and rather than starting from scratch she would like to adopt parts of an American intelligence test. The test is of the "omnibus type", meaning that various mental functions are measured, such as verbal reasoning, memory, spatial skills, and mental arithmetic. The aim of the test is not a comparison of test performance of Zimbabweans with Americans but the development of a test that can measure intelligence in the main local language and that can help in educational and vocational counseling.

The second example is more comparative in nature. A large international company that specializes in telecommunication is doing better in some countries than in others and the management is interested in differences in "company culture" in its various national branches. The services of a psychologist, let us call him Andrew, are hired to set up an international survey among company employees in the various subsidiaries to measure company communication, job satisfaction, and hierarchies in manager–employee relationships. Andrew's company has developed a survey questionnaire for assessing these views that has been administered in a cross-cultural study comparing employees from the U.K., Japan, and the U.S.A. The current study has to be done in several other countries.

The third example involves a Greek psychologist (let us call her Maria) who is interested in Greek personality structure. She wants to design a questionnaire that reflects the implicit personality theories of Greek people. She will do this by first carrying out a local survey in which she asks a random sample of Greek adults to describe their partner, a close family member, and a neighbor. Based on these free descriptions, she wants to compose an instrument that captures Greek personality in all its facets.

The fourth example introduces Guillermo, a Hispanic school psychologist in Chicago, who has been asked by the local educational authority to design an instrument for assessing the cognitive skills of preschoolers in the (multicultural) Chicago population. The existing English-language test does a poor job in assessing the cognitive skills of the nonnative speakers.

What are the cross-cultural challenges facing these four psychologists?

That all four studies have salient cross-cultural components is not obvious; only the second study seems to be genuinely cross-cultural, as it involves a data collection in more than one country. Let us first take a closer look at Joan who develops an intelligence test for use in Zimbabwe. Cross-cultural work is relevant for Joan because it gives her a good starting point. There is ample evidence that western tests show a factorial structure that is also found in applications among nonwestern groups. Cross-cultural studies have confirmed that a hierarchical

structure of intelligence (see, e.g., Carroll, 1993) that has been reported in western groups can also be found in nonwestern groups. There is also evidence that the common western finding that intelligence test scores predict school performance (e.g., Ghiselli, 1956) is also valid elsewhere. These observations do not imply that Joan should copy a western instrument without any concern about its suitability in a Zimbabwean context. Rather, she may find it necessary to change stimuli, response formats or even complete subtests because of unfamiliarity of stimulus material, inappropriateness of item content, or other factors that threaten the suitability of the test in Zimbabwe. Nonwestern applications of western tests help to get insight in the specific weaknesses of western tests, such as their capitalization on scholastic knowledge. In short, knowledge of the cross-cultural literature will help her to appreciate the strength and weakness of western tests.

Andrew's study of the international telecommunication company is an example of an important new area of research: topics in international business. Some questions Andrew will have to solve are: Is the existing survey questionnaire adequate for the countries in which it has not yet been administered? Will the instrument measure the same constructs across the countries? How can this issue be addressed statistically? Are there country-specific items in the instrument that reduce its adequacy? Can the instrument be adequately translated in the various languages of the study (e.g., colloquialisms may threaten the translatability)? Should the translation be close (literal) or should minor or major parts of the instrument be adapted?

Maria, who is interested in Greek personality structure, has to face quite different cross-cultural issues. Studies of indigenous personality have been carried out in China (e.g., Cheung et al., 1996; Cheung & Leung, 1998) and the Philippines (e.g., Guanzon-Lapeña, Church, Carlota, & Katigbak, 1998); Maria's study extends this research line to Greece. She can adopt the procedures of the earlier studies. Her work involves cross-cultural aspects in two ways. She can compare her results with those obtained in China and the Philippines, addressing the question of cross-cultural similarities and differences in personality structure. Second, there is a fair chance that Maria will decide to administer other personality questionnaires in addition to her newly developed Greek inventory. She will be interested in the question to what extent Greek personality differs from the standard structure reported in other countries. In order to find out what is unique for the personality of a county, one will need to know the commonalities with other cultures. Therefore, she may decide one or more measures that have been applied in various countries and that have shown a stable personality structure. Well known examples are the Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975) and the NEO-FFI-R (Costa & McCrae, 1992), based on the Five-Factor Model of Personality. There is evidence for cross-cultural stability of the factorial structure underlying both instruments (Eysenck & Eysenck, 1983; McCrae & Costa, 1997).

Guillermo, a school psychologist in Chicago, deals with educational issues in multicultural societies. It is a recurrent finding that common western tests may not be suitable for use in multicultural societies. For example, instruments assessing mental abilities often show undesirable cultural and verbal loadings. A test of mental arithmetic can easily become an implicit test of word knowledge when the children studied differ in their proficiency in the testing language. Not all children who are enrolled for the first time, may have an adequate knowledge of the language and culture of the test. Yet, teachers and counselors may like to get some insight in the intellectual capabilities of the child. The administration of a standard instrument is then rather uninformative, unless one is specifically interested in the child's knowledge of the mainstream language and culture. Cultural sensitivity of service delivery to these children will increase their quality. It is a

major challenge to design instruments that can be used in a culturally heterogeneous group. Ideally, a poor knowledge of the culture and language of the test developer should not influence performance on such instruments. Practically, very few cross-cultural psychologists will maintain that there are such "culture-free" and "culture-fair" tests. Yet, the implicit agenda of these test movements which began more than 60 years ago to reduce unwanted sources of cross-cultural score differences (e.g., Cattell, 1940; Cattell & Cattell, 1963) is still highly important and has not lost salience since it was first expressed.

Bias and equivalence

There are a few recurrent themes in the cross-cultural topics to be dealt with by our four colleagues. These issues are clearest in Andrew's cross-cultural study but can also be traced in the other studies. The first one deals with the question to what extent the instruments measure the same in each cultural group. Does Andrew's inventory assess manager-employee relationships in each country? It may well be that scores are not directly comparable across cultures because of the presence of nuisance factors. For example, items may have country-specific contents, which render them inadequate for cross-cultural comparison, items may have been inadequately translated, the countries studied may show differences in response styles such as social desirability and acquiescence. Similarly, implicit knowledge of the mainstream American culture may be assumed in Guillermo's test of cognitive skills of migrant children. These threats of the comparability of scores are known as bias (van de Vijver & Leung, 1997a, b). More technically, bias refers to all factors that impact on test scores and do not belong to the construct under study. Equivalence is a closely related concept that refers to the influence of bias on the comparability of test scores.

Following Van de Vijver and Leung (1997a, b), three sources of bias in cross-cultural research are distinguished (see also Van de Vijver, 2001). The first is called *construct bias*; it occurs when the construct measured is not identical across groups or when behaviors that constitute the domain of interest from which items are sampled, are not identical across cultures. Triandis and Vassiliou (1972) have argued that *philotimo* is a person-describing adjective that is unique to the Greek language and culture. Let us assume that the claim is correct and that the term refers to an underlying set of behaviors that are associated with one another in Greece but not in any other language. Maria's study of the implicit theory of Greek personality could then find a *philotimo* factor or scale that would not be found in any other country. The absence of this factor in existing personality inventories would then point to construct bias when the cross-cultural comparison involves Greeks.

An important type of bias, called *method bias*, can result from sample incomparability, instrument characteristics, tester and interviewer effects, and the method (mode) of administration. In general, method bias is a label for all sources of bias emanating from aspects that are described in the method section of empirical papers. Important sources of method bias are differential stimulus familiarity (in mental testing) and differential social desirability (in personality and survey research). Method bias constitutes an important source of alternative explanations of observed cross-cultural differences. Schooled and unschooled participants almost always show a difference in stimulus familiarity and test-wiseness which has virtually unavoidable consequences for observed scores: at least some of the observed differences will be due to differential test exposure, which is unrelated to the construct under study. Both Joan and Guillermo may have to deal with children with a widely different educational background, which may impact on their test scores. Andrew's comparative study of employees will have to deal with cross-national differences in social desirability. The salience of

this factor will increase with the difference in Gross National Product (wealth) of the participating countries. Van Hemert, Van de Vijver, Poortinga, and Georgas (in press) found a strong negative correlation between the Gross National Product of a country and its score on the Lie Scale of the Eysenck Personality Questionnaire.

Finally, the last type of bias refers to anomalies at item level; it is called *item bias* or *differential item functioning*. According to a definition that is widely used in psychology, an item is biased if persons with the same standing on the underlying construct (e.g., they are equally intelligent) but coming from different cultural groups, do not have the same average score on the item. The score on the construct is usually derived from the total test score. If a geography test administered to pupils in Greece and the Netherlands, contains the item "What is the capital of Greece?", Greek pupils can be expected to show higher scores on the item than Dutch students, even when pupils with the same total test score would be compared. The item is biased because it favors one cultural group across all test score levels. Of all bias types, item bias has been most extensively studied. Various psychometric techniques are available to identify item bias (e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993).

Four different types of equivalence can be envisaged (cf. Van de Vijver & Leung, 1997a, b). The first type is labeled *construct nonequivalence*. It amounts to comparing "apples and oranges" (e.g., the comparison of Chinese and Western filial piety, discussed above). Because there is no shared attribute, no comparison can be made. The second is called *structural* (or *functional*) *equivalence*. An instrument administered in different cultural groups shows structural equivalence if it measures the same construct in these groups. Structural equivalence has been examined for various cognitive tests (Jensen, 1980; Van de Vijver, 1997), Eysenck's Personality Questionnaire (Barrett, Petrides, Eysenck, & Eysenck, 1998), and the so-called Five-Factor Model of personality (McCrae & Costa, 1997). Structural equivalence does not presuppose the usage of identical instruments across cultures.

The third type of equivalence is called *measurement unit equivalence*. Instruments show this type of equivalence if their measurement scales have the same units of measurement and a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the same measurement units in each culture). At first sight it may seem unnecessary or even counterproductive to define a level of equivalence with the same measurement units but different scale origins. Why would scales have different origins across cultural groups? The need for the concept of measurement unit equivalence may become clear by looking at the impact of differential social desirability or stimulus familiarity on cross-cultural score differences in more detail. Suppose that the Raven test has been administered in literate and illiterate groups. It is not farfetched to assume that cross-cultural differences in stimulus familiarity will affect the scores. The literate subjects are expected to show higher scores and to have a larger stimulus familiarity. At least some of the observed score differences may have to be accounted for by differential stimulus familiarity. The latter will obscure real cross-cultural differences. When the relative contribution of both sources cannot be estimated, the interpretation of group comparisons of mean scores remains ambiguous. A correction for differential familiarity would be required to make the scores comparable. It may be noted that the basic idea of score corrections that are needed to make scores fully comparable is also applied in covariance analysis, in which score comparisons are made after the disturbance created by concomitant factors (bias in the context of the present chapter) is statistically controlled for.

Only in the case of *scalar (or full score) equivalence* direct comparisons of scores can be made; it is the only type of equivalence that allows

for statistical tests that compare means (such as *t* tests and analyses of variance). This type of equivalence assumes the same interval or ratio scales across groups and the absence of any type of bias. Conclusions about which of the latter two types of equivalence applies are often difficult to draw and can easily create controversy. For example, racial differences in intelligence test scores have been interpreted as largely due to valid differences (scalar equivalence) and as mainly reflecting measurement artifacts (measurement unit equivalence).

Structural, measurement unit, and scalar equivalence are hierarchically ordered. The third presupposes the second, which presupposes the first. As a consequence, higher levels of equivalence are more difficult to establish. It is easier to demonstrate that an instrument measures the same construct in different cultural groups (structural equivalence) than to demonstrate numerical comparability across cultures (scalar equivalence). On the other hand, higher levels of equivalence allow for more precise comparisons of scores across cultures. Whereas in the case of structural equivalence, only factor structures and nomological networks (Cronbach & Meehl, 1955) can be compared, measurement unit and full score scalar equivalence allow for more fine-grained analyses of cross-cultural similarities and differences. It is only in the latter that mean scores can be compared across cultures in *t* tests and analyses of (co)variance.

The three A's of translations: Application, adaptation, and assembly

All four psychologists can use existing measures in their study, either for the whole battery as part of a larger battery (such as Maria's use of existing personality questionnaires in addition to a newly developed scale). There are three options in the process of translating instruments: application (i.e., close translations), adaptations (i.e., close translation of parts and alterations of other parts), and assemblies of a new instrument (i.e., designing a completely new instrument). Which option is chosen has important ramifications for all stages of a study. A distinction can be made between comparative and noncomparative test usage. Andrew's study is primarily comparative as the multinational company is interested in a comparison of country scores. Similarly, Guillermo's test will have to be applied among pupils from various ethnic groups; consequently, his test should allow for cross-cultural score comparisons. Maria's aim is not comparative. Rather, the aim of designing an inventory assessing Greek personality does not involve any explicit comparisons with foreign data. The only comparative moment may be in the establishment of structural equivalence of the new, Greek inventory and existing questionnaires. Similarly, Joan's test of cognitive skills for Zimbabwean children is only implicitly comparative. Her first aim is to design an instrument that has a good reliability and validity (both construct and predictive).

Demands imposed on instruments are higher in comparative than in noncomparative test usage. In particular when scores have to be compared across cultures and full score equivalence has to be assumed, applications are almost always used. Standard statistical techniques to compare means such as *t* tests and analyses of variance assume the same interval or ratio scale in each cultural group, which can easily be achieved only when applications are used. Some statistical techniques are available that deal with adaptations (partly dissimilar stimuli), such as item response theory (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) and structural equation modeling (e.g., Bollen, 1989; Maruyama, 1998). When assemblies are used, numerical scores comparisons across cultures cannot be carried out.

The choice of a translation technique (between the three A's) may seem straightforward from a statistical perspective: if we do not want to challenge numerical score comparability across

cultures even prior to the data collection, application seems to be the best choice. Unfortunately, the statistical viewpoint entails only one kind of consideration that is relevant in cross-cultural studies. The level of score comparability in a cross-cultural study is the outcome of a number of decisions and statistical properties of the data. The first question to be considered in deciding on a translation strategy should be the nature of the bias to be expected. If there is a fair chance that construct bias may challenge the results (e.g., when Maria would only apply various non-Greek tests to assess Greek personality), an assembly may be the best choice. A choice of an application would be misleading in such a case. Cheung and her colleagues found that the Big-Five Model of Personality is supported in China but that the five factors do not capture all systematic variation in person descriptions of Chinese people. In addition to the five factors found in much Western research, other factors were found dealing with relational characteristics, such as relationship harmony. An application of the Big-Five and the observation of structural equivalence of the inventory in a Chinese and an American group would lead to the incorrect conclusion that the five factors would provide a good description of personality in both groups. In sum, applications are simple and straightforward but this simplicity can quickly turn into a disadvantage when bias occurs (in particular construct bias).

Let us take a closer look at which translation strategies could be utilized by the four psychologists of our example. Joan's assignment to design a cognitive skill test for Zimbabwe gives her considerable freedom in choosing a translation option. Some subtests may be closely translated; a well-known example is the Digit-Span of the Wechsler tests. This test of short-term memory is almost always literally translated. Vocabulary, a test of word knowledge, is almost never literally translated. Rather, word frequency lists are used to compile new lists of words, which are then pilot tested. Adaptations are also widely used in intelligence testing. For example, it is common to adapt items of mental arithmetic to the local currencies. The conversion (e.g., from US dollars to Greek drachmas) may make it necessary to change the item contents in order to maintain the original calculation underlying the item.

Andrew needs inventories that yield scores that can be compared across cultures. Therefore, he will be mainly interested in inventories that can be closely translated and show good psychometric properties in all countries involved. Some instruments may be adapted if close translations would be inadequate. In practice there is a fair chance that Andrew does not know at beforehand to what extent his measures are adequate in all countries. If that is the case, he will probably "blindly" administer the inventories. In the data analyses he can scrutinize structural equivalence and item bias, but it will be impossible to demonstrate the presence or absence of construct bias.

Maria may start from an existing questionnaire and attempt to adapt it to the Greek context. More likely, however, she will start with a field study in which she asks adult Greeks to describe persons in their environment, look for commonalities in these descriptions and formulate items on the basis of these descriptions. This part of the study exemplifies the usage of assemblies. In addition to the new inventory, she will have to administer an existing list in order to be able to pinpoint differences and similarities with findings reported elsewhere. She may prefer to alter as little as possible in the translation process in order to get a measure of Greek personality according to non-Greek standards for referential purposes.

Guillermo may want to adapt tests from other languages (e.g., a translation of an English inventory in Spanish to be used among Hispanics). In addition, he may want to change existing tests so as to improve their appropriateness for a multicultural group (e.g., by removing or changing items with a contents that are too culture-specific).

Conclusion

Cross-cultural assessment has become more important during the last decades. Whereas in the early days of cross-cultural psychology there was an emphasis on the comparison of western and nonwestern samples, current studies typically involve western countries. The increased interest in cross-cultural studies can be expected to continue in the foreseeable future. An important line of development in this assessment involves the shift of explicit comparisons to more implicit comparisons; test administrations in multicultural groups often involve such implicit comparisons. Psychologists working in multicultural societies need to have a basic knowledge of cross-cultural assessment. It would be unrealistic to assume that normed tests will be available for all cultural groups in a multicultural society. For example, the cultural composition of the Dutch society is so heterogeneous and the various ethnic groups so numerous that national norms can be developed only for the most frequently employed tests for the largest groups. For many cultural groups there will be no norms for any psychological test. Applications of tests in these groups (unavoidably in the main language) have to be considered with great caution. The blind application of majority norms to these groups may lead to incorrect inferences about the testee's personality or cognitive skills. The reduced suitability of tests in these groups should be routinely acknowledged in reporting results. Rather than denying that there is any problem in the assessment procedure, it is more prudent to acknowledge the limitations of the instrument.

Cross-cultural psychology has long been a branch of psychology in which research was carried out by specialists in the field. In the last decades there is a clear change of trend. Most cross-cultural research is carried out nowadays by psychologists who do not spend a lifetime in this type of research but who are interested in cross-cultural issues and who see cross-cultural work as a natural extension of their monocultural work. From my perspective it is reassuring to see that most research is done by nonspecialists, as it shows the widespread interest in «the cultural factor» in human behavior. At the same time the trend also has a potential downside. Not all cross-cultural researchers appreciate that cross-cultural research has to deal with a number of specific methodological issues that are more or less salient in monocultural research. It is important that cross-cultural researchers are aware of the challenges of their studies. The current article has described methodological issues to be dealt with in cross-cultural research. Hopefully, the present article helps to increase the interest in cross-cultural research and the awareness of the specific issues of this research.

References

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.

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

Camilli, G., & Shepard, L. N. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies.* Cambridge, UK: Cambridge University Press.

Cattell, R. B. (1940). A culture-free intelligence test. *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. (1998). Indigenous personality measures. Chinese examples. *Journal of Cross-Cultural Psychology, 29,* 233-248.

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.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire.* San Diego, CA: EdITS.

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, pp. 41-69). Hillsdale, NJ: Erlbaum.

Ghiselli, E. E. (1956). *The validity of occupational aptitude tests.* New York: Wiley.

Guanzon-Lapeña, M. A., Church, A. T., Carlota, A. J., & Katigbak, M. S. (1998). Indigenous personality measures: Philippine examples. *Journal of Cross-Cultural Psychology, 29,* 249-270.

Hambleton, R. K., & Swaminathan H. (1985). *Item response theory: Principles and applications.* Dordrecht, The Netherlands: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

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

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

Maruyama, G. M. (1998). *Basics of structural equation modeling.* Thousand Oaks, CA: Sage.

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

Triandis, H. C., & Vassiliou, V. (1972). A comparative analysis of subjective culture. In H. C. Triandis (Ed.), *The analysis of subjective culture* (pp. 299-335). New York: Wiley.

Van de Vijver, F. J. R. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology, 28,* 678-709.

Van de Vijver, F. J. R. (2001). Research methods. In D. Matsumoto (Ed.), *Handbook of culture and psychology* (pp. 77-97). Oxford: Oxford University Press.

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, pp. 257-300). Boston: Allyn & Bacon.

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

Van Hemert, D. D. A., Van de Vijver, F. J. R., Poortinga, Y. H., & Georgas, J. (in press). *Structure and score levels of the Eysenck Personality Questionnaire across individuals and countries. Personality and Individual Differences.*