Training and Education in Professional Psychology
2009, Vol. 3, No. 2, 94–105

1931-3918/09/$12.00 DOI: 10.1037/a0014516

A Critical Analysis of Cross-Cultural Research and Testing Practices: Implications for Improved Education and Training in Psychology

Barbara M. Byrne
University of Ottawa

Thomas Oakland
University of Florida, Gainesville

Frederick T. L. Leong
Michigan State University, East Lansing

Fons J. R. van de Vijver
Tilburg University

Ronald K. Hambleton
University of Massachusetts, Amherst

Fanny M. Cheung
The Chinese University of Hong Kong

David Bartram
SHL Group Ltd., London

Psychological research that involves cross-cultural comparisons has increased considerably during the last decade and is expected to escalate further. Given its growing popularity within mainstream psychology, cross-cultural research no longer can be considered the sole domain of experts trained in this specialization. Concomitant with this expansion, important methodological advances in quantitative psychology (e.g., measurement, statistical analysis, and research design) impact the study of cultural differences. The purpose of this article is to heighten awareness of important methodological advances among psychologists being prepared for or engaged in teaching, research, consultation, or other forms of practice that focus on diverse cultural groups. Credible and unbiased research findings coupled with psychometrically sound selection and use of assessment instruments contribute importantly to attaining the gold standard for all psychological research and testing practices. This article highlights methodological advancements and other issues that bear importantly on both the preparation and subsequent practices of psychologists in ways that promote credibility and lessen bias.

Keywords: education and training in cross-cultural assessment, quantitative training, testing practices across culture, psychometric training, test use across culture

Cross-cultural psychology as a scholarly discipline "has grown from a whisper and a hope circa 1960 into a large and thriving intellectual enterprise circa 2000" (Segall, Lonner, & Berry, 1998, p. 1108). The rapid increase in racial-ethnic, religious, and other forms of cultural diversity within many nations has contributed to psychology's mounting interest in cultural diversity, one that has spawned a virtual explosion of research and testing practices that compares cultural groups within one country as well as between

BARBARA M. BYRNE is professor emeritus in the School of Psychology, University of Ottawa, Canada. As a quantitative psychologist, her research interests focus on construct validity issues related to measuring instruments, theoretical concepts, and causal networks, and their equivalence across mono- and multicultural groups, and sound application of factor analysis and structural equation modeling.

THOMAS OAKLAND is a professor of Educational Psychology at the University of Florida. His research interests include adaptive behavior, test development and use, professional standards, and the international status of school psychology.

FREDERICK T. L. LEONG is professor of psychology at Michigan State University in the Industrial/Organizational and Clinical Psychology programs and serves as the director of the Center for Multicultural Psychology Research. He has authored or co-authored over 115 articles in various psychology journals, 70 book chapters, and edited or co-edited 10 books.

FONS J. R. VAN DE VIJVER holds a chair in cross-cultural psychology at Tilburg University, The Netherlands, and North-West University, South Africa. He has over 275 publications, mainly about cross-cultural methods.

RONALD K. HAMBLETON'S research interests are in the area of methodology for the development and validation of educational and psychological assessments: Specifically, test translation methodology, applications of item response theory, computer-based testing, setting performance standards, and score reporting.

FANNY M. CHEUNG received her PhD in psychology from the University of Minnesota. She is currently professor of psychology and chairperson of the Department of Psychology at The Chinese University of Hong Kong. Her research interests include cross-cultural personality assessment, Chinese psychopathology, and gender issues.

DAVID BARTRAM is research director for SHL Group Ltd. and special professor in Occupational Psychology and Measurement at the University of Nottingham's Institute for Work, Health, and Organizations.

Authors listed four through seven are in the order of their commentary presentation.

CORRESPONDENCE CONCERNING THIS ARTICLE should be addressed to Barbara M. Byrne, 3301 Diamond Key Court, Punta Gorda, FL 33955-4656. E-mail: bmbch@uottawa.ca or bmbyrne@comcast.net

two or more countries. At least three recent trends distinguish much of the current research from prior research in this area. First, psychologists with little training in cross-cultural psychology increasingly are conducting cross-cultural research. Second, such research often involves the use of tests[1] and other data gathering methods that first must be adapted for use in other cultures. Finally, questions being addressed through research generally require knowledge and application of advanced research designs and statistical analyses. These three conditions importantly impact the preparation and practices of psychologists.

In recognition of this rapidly changing methodological landscape in cross-cultural research, Division 52 (International Psychology) of the American Psychological Association (APA) invited the first author to establish a joint Division 52/Division 5 (Div52/5; Evaluation, Measurement, & Statistics) Task Force charged with the task of identifying methodological aspects of cross-cultural research considered to need an update. Specifically, the purposes of this joint Task Force were threefold: (a) to seek input from experts in both cross-cultural/international psychology and quantitative psychology regarding a need to update existing methodological practices in cross-cultural/international research; (b) to identify particular areas of methodological weakness; and (c) to elaborate on how these perceived weaknesses serve to impede, distort, degrade, or otherwise limit the generalization of research findings.

Selection of Div52/5 Task Force members was based on two criteria: (a) internationally recognized expertise in measurement/methodological issues applied to cross-cultural/international research and (b) applied expertise spanning the broad areas of psychometrics, research design, and statistical analysis. In addition, the Task Force included both practitioners and academicians. This selection process resulted in the formation of a seven-member Task Force, each of whom is a contributor to this article. Although collectively the Div52/5 Task Force members have expertise on a wide range of psychology specializations, they are bound by two common threads: all have substantial research and/or practical experience working with diverse cultural groups as well as have a common interest in and experience resolving methodological difficulties associated with this type of research and/or practice.

The intent of this paper is to articulate the findings of this Div52/5 Task Force, albeit in a form that highlights methodological aspects important to cross-cultural research. However, we contend the issues discussed in this paper have implications for a broader segment of psychologists, including those dedicated to teaching and clinical and mental health practices, and engaged in issues important to cultural differences and diversity. Thus, our broad and primary objective is to heighten awareness of important advances that can impact the design and analyses of research as well as testing and other clinical practices that involve cross-cultural issues. This objective includes the recommendation to review current training practices, and make revision, when needed, that future psychologists are sufficiently prepared to meet the challenges of cross-cultural work.

This article is structured around three main sections. Section 1 highlights three overarching methodological concerns of cross-cultural research that serve to contaminate the credibility of such inquiry. It begins by elaborating these overarching methodological concerns because their resulting limitations are linked to more specific methodological deficiencies identified later in the article. Section 2 discusses four specific methodological aspects of cross-cultural research and testing practices that require attention. These topics are presented in the form of commentaries each of which reflects the particular concern of its author, an internationally recognized expert in the area. Section 3 summarizes information presented in the two preceding sections and suggests ways by which preservice and in-service training programs may provide information and opportunities to acquire skills critical to the conduct of research and testing practices that involve multicultural populations.

Overarching Methodological Concerns in Cross-Cultural Research and Testing Practices

All specific methodological concerns inherent in cross-cultural research and testing practices discussed in this article are linked in one way or another to three core and broad-based issues: (a) assumption of equivalent psychological meaning and factorial structure of a measuring instrument across cultural groups, (b) failure to take into account the hierarchical nature of the cultural data, and (c) lack of familiarity with ethical issues governing test adaptation and test use in diverse cultural settings.

Issues of Structural and Measurement Equivalence

Structural and measurement equivalence issues constitute a common methodological concern in cross-cultural research. When a psychological construct (e.g., self-concept) is measured in different cultural groups, one cannot assume that the meanings of the scores will be identical and that the scores subsequently can be compared across groups. Indeed, cross-cultural comparisons of these scores are interpretable only in light of evidence that the meaning and dimensional structure of the construct (e.g., self-concept) as well as the items comprising the measuring instrument are group-equivalent (van de Vijver & Leung, 1997). Thus, instruments and the psychological constructs they are designed to measure must be tested statistically for their equivalence across the cultural groups being studied. Unfortunately, many who are engaged in cross-cultural research continue to compare level (e.g., mean) scores without first testing the extent to which both the instrument and the psychological meaning and structure of its underlying constructs are group equivalent.

Scholars in cross-cultural psychology have long grappled with this equivalence issue (e.g., Johnson, 1998; Poortinga, 1995; van de Vijver & Leung, 2000; van de Vijver & Tanzer, 1997) and have proposed various terms to help clarify its various forms (e.g., conceptual, construct, functional, linguistic, and metric). However, in essence, each of the proposed classifications is part of two primary equivalence categories: *structural equivalence* and *measurement equivalence*. Structural equivalence (also termed *conceptual or construct equivalence*) is concerned with the extent to which the meaning and dimensional structure of a psychological construct are identical across cultural groups. Measurement equivalence is concerned with the extent to which both the item content and psychometric properties (i.e., validity and reliability) of the instrument are similar across groups.

[1] Throughout this article, the terms *test*, *instrument*, and *scale* are used interchangeably.

Cultural differences in societal structures, values, and socialization practices contribute to differences found in the meaning and/or structure of a measured construct and the perception of its related item content. Thus, societal qualities alone can contribute importantly to structural and measurement nonequivalence (e.g., differences in the psychological meaning and/or dimensional structure of a construct as well as differences in the perception of item content and/or psychometric properties of the measuring instrument).

An understanding of the concept of *structural nonequivalence* may be promoted by presenting an example from a study designed to test statistically for the equivalence of an instrument developed to measure the physical (i.e., self-perceived physical ability and appearance) and social (i.e., self-perceived social relationship with peers and parents) self-concepts of Australian and Nigerian adolescents (Byrne & Watkins, 2003). Evidence of structural nonequivalence was claimed on the basis of the inequality of construct relations (i.e., latent factor correlations) across groups. Specifically, the latent factor correlation between the perceived ability and perceived appearance dimensions of physical self-concept differed significantly across culture, with their intercorrelations being higher for Australian than for Nigerian adolescents. This finding may derive from the differing societal values held by these two adolescent groups. Self-perceived physical attractiveness may be defined by Australian adolescents in terms of a superior body physique and by Nigerian adolescents in terms of beautiful facial features.

In contrast to structural nonequivalence, *measurement nonequivalence* signals discrepancies in the operation of particular items (e.g., differential perception of item content) across groups, which then impacts associated links to the underlying constructs (i.e., the factor loadings). These discrepancies ultimately lead to differential validity and/or reliability of the measuring instrument across groups. Measurement nonequivalence typically arises as a consequence of method and/or item bias, which can derive from many sources. Method bias, for example, can arise from particular characteristics of the instrument (e.g., response styles such as acquiescence or extremity ratings) or from the manner of its administration (e.g., communication problems between interviewer and interviewee). Item bias can occur as a consequence of differences in the appropriateness of item content (e.g., use of a term or colloquialism that is not understood in at least one of the cultural groups), inadequate item formulation (e.g., unclear wording), or unsuitable item translation. (For an elaboration of these biases, readers are referred to van de Vijver & Leung, 1997 and van de Vijver & Poortinga, 2005.)

Although biases associated with item translation often are discussed under the topic of linguistic equivalence, they represent a special case of measurement equivalence wherein the original instrument (i.e., source instrument) is translated into another language (i.e., the target instrument) and then awaits the testing of its structural and measurement equivalence. Modifications of a source instrument for subsequent use in a culture that differs from the one in which it was developed involves a comprehensive and rigorous series of procedures that test statistically for the validity of its scores within the new cultural context and for its structural and measurement equivalence with the source instrument. The term *test adaptation* is used to describe this more advanced approach to the development and use of translated instruments (Hambleton, Merenda, & Spielberger, 2005).

This review of structural and measurement equivalence underscores the importance for psychologists who wish to conduct cross-cultural research to be aware of issues involved in testing across different cultural groups as well as of the methodological tools needed to address these issues. Structural equation modeling represents one of the key methodological approaches to testing for the equivalence of both psychological constructs and measuring instruments across cultural groups. Item response theory, discussed in more detail later in this article, provides an alternative means of testing for the equality of only test items.

Hierarchical Structure of the Data

By their very nature, cross-cultural data are hierarchically structured: Individuals are nested within culture (M. W.-L. Cheung, Leung, & Au, 2006; van de Vijver & Leung, 1997). Thus, this hierarchical (or multilevel) structure must be considered through the use of statistical analyses that have the capability to focus on the individual (i.e., disaggregated) as well as the group (i.e., aggregated) levels of the structure. Failure to address this data structure reality results in analyses that assume a single-level focus. In other words, analyses are conducted either at the individual (lower) or at the country (higher) level. With the individual-level approach, individuals form the unit of analysis and researchers focus on the extent to which relations among variables within culture are similarly related across culture. In contrast, with the culture-level approach, countries form the unit of analysis, and research centers on comparisons of means across national groups with no assumption that variable relations found across cultures necessarily hold within each culture.

Regardless of whether cross-cultural researchers conduct their analyses at the individual or at the country level, there is an implicit assumption that both the psychological meaning of the construct and its underlying factorial structure are equivalent across these two levels. However, unless tested statistically, one is unable to know the extent to which this assumption of construct equivalence is supported. For example, M. W.-L. Cheung et al. (2006), in testing for the equivalence of a hypothesized five-factor structure of social axioms across individual and country levels, found only one of the five factors (Social Cynicism) to be level-invariant. Although the other four factors (Social Complexity, Reward for Application, Religiosity, and Fate Control) were strongly correlated at the country level, they were virtually uncorrelated at the individual level. Provided with such findings, van de Vijver and Poortinga (2002) contended that different constructs were needed to describe individual and country differences (see van de Vijver & Poortinga, 2002 van de Vijver, van Hemert, & Poortinga, 2008, for a description of other possible nonequivalent outcomes across individual and country levels).

The above example underscores the importance of not assuming test interpretations and the underlying structure of a psychological construct across two or more cultures are the same. Indeed, problematic repercussions at both the individual and country levels may be inevitable when the nested structure of cross-cultural data is ignored. These difficulties arise from the disaggregation and aggregation approaches to the analyses, respectively (see van de Vijver et al., 2008, for a substantive elaboration of this issue; see

Julian, 2001, for a statistical explanation). Thus, unless this multilevel structure of the data is taken into account, findings based on single-level analyses conducted either at the individual or country level must be considered dubious (e.g., van de Vijver & Poortinga, 2002). This caveat notwithstanding, a review of the literature reveals that, except for a few recent studies (e.g., M. W.-L. Cheung & Au, 2005; M. W.-L. Cheung et al., 2006; van de Vijver & Watkins, 2006), virtually all cross-cultural research has been conducted as single-level analyses, mostly at the individual level (Leung, 1989; van de Vijver & Poortinga, 2002).

The use of structural equation modeling within the framework of a multilevel model provides the most rigorous approach to testing for cross-level equivalence. This strategy allows the researcher to consider both levels of the hierarchically structured data simultaneously. However, one critically important aspect of this analytic approach is the need for large country-level samples. With approximately 220 countries in the world, many of which are small and still developing (M. W.-L. Cheung & Au, 2005), this requirement when applied to cross-cultural data is excessively restrictive and quite unrealistic. Indeed, somewhat small country-level samples (e.g., *N*s of 27 and 40) have been used satisfactorily (see M. W.-L. Cheung & Au, 2005; M. W.-L. Cheung et al., 2006, respectively). Alternatively, a less demanding multilevel approach based on exploratory factor analysis (van de Vijver & Poortinga, 2002) has been found to work well with country-level samples ranging from *N*s of 15 to 25 (F. J. R. van de Vijver, personal communication, September 23, 2007).

On the other hand, researchers may need to take a completely different approach to their analyses when data comprise a small number of cultural groups; analysis of covariance procedures serve this purpose well. For example, Shebani, van de Vijver, and Poortinga (2008) hypothesized that cross-cultural differences in the short-term memory digit span of Dutch and Libyan children could be explained by Baddeley's (1997) phonological loop model, which predicts that speakers of languages with longer digits (measured as pronunciation speed) can hold fewer digits in their short-term memory. Consistent with Baddeley's model because Arab digits are longer than Dutch digits, Libyan children pronounced fewer digits than Dutch children in a fixed amount of time thereby resulting in a longer digit span for Dutch children. Using analysis of covariance, cross-cultural performance differences no longer were significant after correction for differences in pronunciation speed (see van de Vijver & Leung, 1997, for details related to these small sample procedures; Byrne & Watkins, 2003, for an illustrated application).

Researchers comparing psychological variables across cultural groups should not ignore the nested effects of the data. Although applications of multilevel modeling based on structural equation modeling increasingly are reported in journals for educational, sociological, and organizational data, such applications in psychology involving cross-cultural data are virtually nonexistent with two recent exceptions (M. W.-L. Cheung & Au, 2005; M. W.-L. Cheung et al., 2006). Likewise, the number of studies using exploratory factor analysis to examine multilevel equivalence and the influence of contextual variables as possible sources of cultural influence are rare. This paucity seemingly indicates a lack of knowledge and expertise in the use of structural equation modeling in general and multilevel modeling in particular as well as an insufficient understanding of the need to explain the possible influence of culture on group differences.

This dearth in the literature may be understandable given the somewhat limited scope of training among researchers who have acquired an interest in cross-cultural issues as well as those who specialized in cross-cultural psychology during their graduate preparation. Nevertheless, efforts can be made to update student preparation and psychologists' skills and knowledge to be consistent with standards advocated by cross-cultural methodologists who recommend heightened awareness to the importance of testing for cross-level equivalence (see Leung, 1989; van de Vijver & Leung, 1997, 2000; van de Vijver & Poortinga, 2002; van de Vijver et al., 2008).

Ethical Issues Related to Test Adaptation and Use

The profession of psychology and those countries in which it practices are linked through an unwritten social contract whose broad principles are clear. A country agrees to establish and fund institutions that enable psychology to select and prepare its neophytes, to allow psychologists to define and license its practice, and to fund research. In turn, psychology is expected to serve all members of the society well by addressing critical national issues. Its ethics code communicates the ways psychology will serve society.

Research and other forms of scholarship constitute one of psychology's most enduring and important contributions to society. However, psychological research, as with other psychological practices, is subject to both legal and ethical standards in those countries in which research is conducted. Ethics codes may help define acceptable procedural and methodological issues associated with these forms of research. Procedural issues include confidentiality, protection from harm, informed consent, plagiarism, and publication credit. Methodological issues include competence in conducting research and the proper selection and use of tests and other data gathering procedures.

The competence of many professionals engaged in cross-cultural research may be limited. As noted earlier, few psychologists are trained in these research methods. The number who acquires such competence through supervised experience or consultation is unknown yet likely to be small. Most engaged in these forms of research are likely to rely on their foundation knowledge of research methods and ongoing research experiences together with personal study. This reliance may be somewhat insufficient, may lead to inadequacies in the manner in which research is conducted, including data acquisition and interpretation, and may not meet prevailing ethics standards for professional competence.

Psychologists engaged in cross-cultural research should honor available ethics codes in those countries in which their research is conducted. However, most countries do not have psychological associations and thus do not have ethics codes that address psychological research. Moreover, psychology's two major international psychological associations, the International Union of Psychological Sciences and the International Association of Applied Psychology, have not developed ethics codes.

The International Test Commission (ITC; www.intestcom.org) developed two guidelines that address issues important to cross-national research: those for test adaptations (Hambleton et al., 2005) as well as for computer-based and Internet-delivered testing

(Bartram & Hambleton, 2006; Coyne & Bartram, 2006). Ethical issues associated with test adaptation are discussed elsewhere (Oakland, 2005). In addition, the ITC Guidelines on Test Use (Bartram, 2001) discuss five broad and important behaviors associated with good conduct in test use. Comprehensive standards and guidelines that address cross-cultural research are not readily available. Thus, psychologists must rely on codes from the few countries that have developed them and use sound professional judgment when they are not available.

Among the more than 70 psychological associations that are members of the International Union of Psychological Sciences, some but not all have ethics codes. Among those that have codes, many do not address testing issues. A survey of test standards in 31 ethics codes representing 35 countries found approximately one-third do not address test use. Furthermore, among the codes that address test use, relatively few adopt standards analogous to those addressed in the American Psychological Association's (APA; 2002) *Ethical Principles of Psychologists and Code of Conduct* (Leach & Oakland, 2007). However, neither the APA Code nor commonly used textbooks on law and ethics in psychology (e.g., Koocher & Keith-Spiegel, 2007) discuss international research related issues. Nonetheless, professors who teach courses on ethics as well as research design and statistics are encouraged to rely on APA's, 2002 *Ethical Principles of Psychologists and Codes of Conduct* as well as supplementary readings that address ethical issues. Thus, issues directly important to cross-cultural research should be considered for inclusion in subsequent revisions of the APA code as well as in textbooks that address ethical issues in psychology.

The following standards from the 2002 *Ethical Principles of Psychologists and Codes of Conduct* (APA, 2002) should be highlighted during the preparation of psychologists engaged in cross-cultural work: boundaries of competence (Standard 2.01) and maintaining competence (Standard 2.03), consultation (Standard 4.06), delegation of work to others (Standard 2.05), documentation of professional and scientific work and maintenance of records (Standard 6.01), test construction (Standard 9.05), interpreting assessment results (Standard 9.06), assessment by unqualified persons (Standard 9.07), obsolete tests and outdated test results (Standard 9.08), maintaining test security (Standard 9.11), informed consent to research (Standard 8.02), and plagiarism (Standard 8.11). Adherence to these and other ethical standards will help ensure high stakes decisions based on test results are supportable.

Commentaries on Specific Methodological Concerns in Cross-Cultural Psychological Research and Testing Practices

Methodological concerns noted earlier are discussed through the views of four Div52/5 Task Force members. The following commentaries highlight limitations commonly found in cross-cultural psychological research and test use and propose remedies that can be implemented through education and training to better prepare psychologists and other mental health professionals who have an interest in this form of inquiry. The commentaries are presented in an order that flows from more general to more specific training issues. The first commentary discusses foundation issues associated with the early training of cross-cultural psychologists. The second commentary discusses the need to update training from translating tests to adapting them followed by a rigorous judgmental and empirical review. The third commentary retains a focus on adapted instruments and alerts readers to difficulties associated with using imported tests in a culture for which the dimensional structure of the construct being measured may be invalid. The fourth commentary presents a very specific example of how the various issues addressed in this article may be applicable to multinational test use, especially in industrial/organizational psychology and how the dearth of adequately trained psychologists led this author's multinational organization to structure its own training program.

Commentary on Foundation Issues Associated With the Early Training of Cross-Cultural Psychologists

Clients served by psychologists increasingly are culturally diverse. Gone are the days when this diversity can be treated in a "colorblind" way and when test use and treatments seen as adequate for mainstream society are deemed equally adequate for members of all ethnic groups. Thus, psychologists increasingly are being encouraged to scrutinize the adequacy of their procedures in light of diversity issues. For example, digit span subtests of omnibus intelligence tests are an inadequate measure of short-term memory if the client has insufficient mastery of the testing language.

Psychologists often are not trained to conduct interviews and use tests in ways that are culturally appropriate (Sattler, 1998). In addition, testing practices often reflect various explicit and implicit references to the culture of the test developer. Language and other communication issues can easily arise when client and test developers do not share a common cultural background. Thus, training programs that utilize scholarship on models of cultural differences (e.g., individualism—collectivism; Hofstede, 2001), acculturation (e.g., current models of biculturalism; Sam & Berry, 2006), and psychometric issues of comparability (van de Vijver & Leung, 1997) can aid in promoting foundation clinical and methodological knowledge and experiences that help address these issues.

Some professional standards that address assessment (e.g., American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), and training issues (APA, 2003) applicable to cross-cultural practice exist, albeit substantially more are needed. Due to their limited availability, personal judgments and views of psychologists can have a relatively large influence on outcomes. Van de Vijver and Leung (2000, p. 34) referred to these preconceived views as cognitive biases or the "*partis pris*" that impact cross-cultural assessment, treatment, and research. Two common forms of cognitive biases that impede progress and professional practice are: (a) the imbalance between acceptance and rejection of cross-cultural score differences, and (b) Euro American dominance in cross-cultural assessment; let us examine these biases in more detail.

Imbalance between acceptance and rejection of cross-cultural score differences: Are results attributable to the client or the test? Psychologists commonly ponder interpretations of unexpected test scores collected cross culturally (e.g., those that are very low or high). Experience shows that these interpretations often vary depending on the domains being assessed. For example,

score differences on the social domain often are accepted at face value whereas those on the cognitive domain often are rejected as measurement artifacts (cf. Faucheux, 1976). This practice leads to an imbalance between uncritical acceptance and rejection of cross-cultural score differences. In fact, no scientific evidence supports the notion that cross-cultural differences based on psychometrically sound measures of attitudes are credible whereas those based on psychometrically sound measures of cognitive abilities are not credible. This perspective is particularly difficult to reconcile in light of the accepted notion that cross-cultural differences in socialization practices are substantial and have an impact on the expression of many psychological traits.

Euro American dominance in cross-cultural assessment: Is the instrument appropriate for this client? Instruments developed in Western countries commonly are believed to have universal applicability. However, the uncritical administration of Western instruments to non-Western clients can lead to incorrect inferences. For example, although cross-cultural evidence that the Five-Factor Model of personality may be universal is impressive (McCrae, Terracciano, et al., 2005), work on the Chinese Personality Assessment Inventory (F. M. Cheung et al., 2001) has identified a sixth factor that may be needed to describe its social aspects within the Chinese society (e.g., F. M. Cheung, 2004, 2006). The possible existence of a sixth personality dimension exemplifies a potential problem arising from the use of Euro American methods in psychology and suggests that Western findings may impart an incomplete picture of the true structure of personality in non-Western as well as possibly Western societies. The need to be critical of the instruments we use is apparent. A recommendation to use only those instruments shown to yield reliable and valid scores within a multicultural context is self-evident and consistent with existing standards (e.g., American Educational Research Association et al., 1999).

Quality graduate programs in psychology generally make some effort to ensure their students are exposed to the discipline of psychology, not merely to the discipline's methodological dimensions. Thus, their students take courses on the social foundations of behavior, social psychology, developmental psychology, and in other ways acquire an understanding of the biological, environmental, and cultural impact on behavior. In addition, efforts to encourage students to live and work abroad, even somewhat briefly, can help promote desired sensitivity to culture and its impact on attitudes and behavior. Persons engaged in cross-cultural work must have a broad understanding of human growth and development and an appreciation of the impact of cultural attitudes, values, and behavior. Graduate programs that lack this broad foundation are unlikely to successfully prepare their students for such work.

Commentary on Need to Update Training on the Adaptation of Tests

The yearly number of cross-cultural research articles increased from approximately 600 in 1978 to approximately 1,500 in 2003 (van de Vijver, 2006). Clearly, the rate of interest in cross-cultural research has grown enormously. Concomitantly, many psychologists engaged in cross-cultural research have not kept abreast of advances in psychological testing methods. Improvements in testing methodology offer the potential for increasing the validity of many tests used in cross-cultural psychological research and practice. Three important testing methods can assist graduate students and others to improve their engagement in cross-cultural services: modern test theory and practices, computer-based testing, and test adaptation methodology. Each is discussed below.

Modern test theory and practices. Modern test theory, more commonly known as *item response theory*, offers important features that include greater flexibility in test design, test evaluation, and test analysis than classical test theory and associated methods (Hambleton, Swaminathan, & Rogers, 1991). Within the framework of classical test theory, item statistics (e.g., item difficulty, item discrimination, coefficient alpha) are dependent on the examinee sample in which they are obtained. For example, the difficulty level of an item on an achievement test is determined by the proportion of respondents who answer the item correctly and do not consider the abilities of the respondents in the calculations. In contrast, tests developed on item response theory models allow for the separation of respondent characteristics (i.e., observed scores) from item characteristics (i.e., item statistics), thereby overcoming this major limitation of classical test theory. Consequently, models based on item response theory are being used increasingly in the construction of tests, including large scale international assessments (e.g., Program for International Student Assessment; Trends in International Mathematics and Science Study). Details related to both the principles of item response theory and its many advantages over classical test theory can be found elsewhere (e.g., Hambleton et al., 1991).

Many aspects of item response theory make it a desirable methodology for use in cross-cultural psychological research (van de Vijver & Leung, 1997). Three are discussed here. First, estimates of the item statistics are independent of the group status on a particular latent trait, and estimates of the latent traits are independent of items comprising tests. Thus, the use of an identical test (apart from possibly a cultural and language test adaptation) is no longer necessary when conducting comparisons across different cultural groups (van de Vijver & Leung, 1997). Second, item response theory allows for tests of goodness-of-fit between sample data and the hypothesized factorial structure of a measuring instrument. When these differ across groups, considerable information is learned that impacts on test score interpretations across cultural groups. Finally, item response theory provides an excellent statistical framework for identifying test items that may be unfair or differentially valid across comparison groups as a consequence of gender, ethnicity, level of education, religion, culture, and other background variables. Those engaged in test development commonly check for the equivalence of items across targeted samples. This analytic process is termed *differential item functioning* within the context of item response theory (as opposed to measurement nonequivalence within the context of structural equation modeling).

Although some psychologists engaged in cross-cultural practices may have limited interest in using item response theory to build new tests, their understanding of the basic tenets of item response theory will expand their repertory of skills for later use and enable them to become more informed consumers of information. Concepts such as classical test reliability, standard error of measurement, true score, and the classical test model are being replaced by item response theory concepts such as test informa-

tion, conditional standard errors, latent trait scores, and item characteristic curves. Test publishers and some cross-cultural researchers are using item response theory as a framework for building and evaluating tests.

Computer-based testing. Administration of assessment scales and tests via computers is becoming increasingly widespread (Bartram & Hambleton, 2006). These advantages include ease and swiftness of administration and scoring, increased test security, potential to assess high level thinking skills, and their use to facilitate research internationally. Computer based-tests allow researchers and test administrators to monitor an examinee's performance, leading to the identification of optimal test items to be administered. For example, let us assume an examinee is performing very well on a test of academic aptitude. More detailed information, acquired in an efficient manner, about the examinee's academic aptitude is acquired by focusing on more difficult test questions. Testing effectiveness and efficiency are served by matching the items' difficulty levels with an examinee's ability level. Computer-based testing is becoming more common in the assessment of achievement, aptitudes, personality, and health issues such as quality of life.

Cross-cultural psychologists are likely to experience changes in the ways tests are administered and scored in the coming years. Thus, their knowledge and use of computer testing has the potential to significantly advance their clinical and research interests in cross-cultural issues.

Updated methods of test adaptation. In 1992, the ITC embarked on an ambitious effort to develop guidelines for adapting tests because of a widely held perception that test adaptation practices used in cross-cultural research were seriously flawed (see Hambleton et al., 2005). Test adaptation methodology has advanced greatly since 1992. For example, the following three somewhat commonly used test translation strategies now are questioned due to their inadequacies: (a) use of nonprofessional translators, (b) exclusive use of back translation methods, and (c) use of small empirical studies based on persons who are bilingual.

Cross-cultural research may benefit from the availability of the ITC Guidelines for Test Adaptation. They focus on a two-pronged methodological approach that, when used together, can provide a thorough validation of the test adaptation (Hambleton et al., 2005): the use of multiple translators in various test adaptation designs (e.g., forward and backward methods) and the use of various statistical procedures. Empirical evidence derives from investigations of construct equivalence of tests across language and cultural groups as well as method and item biases as noted earlier. Cross-cultural researchers need to be able to use many new statistical methods and judgmental designs (e.g., backward and forward translation designs) to properly translate and adapt their research instruments for use in other languages and cultures. With a backward translation design, the translation goes from source language to target language and is back-translated to the source language. The quality of the translation is judged by comparing the original source language and the back-translated source language versions of the instrument. With a forward translation design, the source and target language versions of the instrument are compared. Older methods of test adaptation are simply not rigorous enough today to ensure valid instrumentation for cross-cultural research.

Training. Graduate students in cross-cultural psychology programs as well as others interested in engaging in cross-cultural work need courses that address modern advances in psychometric and statistical methods, including item response theory, computer-based testing, test adaptation methodology, and structural equation modeling. This training is likely to come from psychometric methods courses usually taught in graduate programs in education and psychology. Although these courses currently may not be highly attended, they have become essential for the practice of psychology cross-culturally. This exposure will help dispel myths about testing and test development (e.g., Cronbach's coefficient alpha for polytomously scored data or the Kuder–Richardson Formula 20 for binary scored data always are the appropriate approaches for reporting reliability evidence, test norms always are needed to interpret tests), promote skills needed for test development and interpretation.

In addition to university courses, many regional, national, and international professional meetings offer workshops that include an emphasis on modern psychometic methods. The text by van de Vijver and Leung (1997) that addresses research design and statistical methods in cross-cultural psychology has been a mainstay for over a decade. In addition, newer books are beginning to emerge (e.g., Bartram & Hambleton, 2006; Hambleton et al., 2005) that extend these perspectives to the psychometric training needs of graduate students and psychologists interested or engaged in cross-cultural issues.

Commentary on Use of Imported Tests

Countries in which psychology is in its early stage of development frequently need to use imported tests. Local researchers and practitioners in these countries often are unfamiliar with professional ethics, copyright requirements, and psychometric methods important to the use of imported tests and cross-cultural assessment. Test users who assess clients from other cultures also may not be aware of cross-cultural differences in test results. Such information is not readily available in the research literature or in test manuals. Risks of misusing tests are obvious (F. M. Cheung, Leong, & Ben-Porath, 2003).

Similar problems are found in cross-cultural research that uses tests or scales developed in English and then translated into different languages. In this regard, most cross-cultural studies mention only the use of back translation, a technique that is not considered to be sufficient. Although little may be known as to why particular items do not work well in certain scales, these item-level differences may affect both the content and the construct validity of the scales. F. M. Cheung (2004) discussed the methodological problems of the "transport and test" approach (Church & Lonner, 1998) in cross-cultural assessment. Test users engaged in cross-cultural work need to consider a number of practical and methodological issues that, although well known to researchers specifically trained in cross-cultural psychology, tend not to be known by other researchers and practitioners trained in different areas of psychology or different disciplines.

Local psychologists using imported tests often encounter practical constraints of limited available resources and expertise. Possible solutions include promoting awareness of the issues in cross-cultural assessment among test users and working with the local psychological associations to promote cross-cultural testing and ethical standards. For example, the editorial board of *Acta Psychologica Sinica*, the top psychology journal in China, requires

authors who submit manuscripts to the journal to stipulate that copyright permission for any test used and test translation contained in the manuscript has been obtained. Likewise, authors and publishers of copyrighted tests can facilitate cross-cultural translation and adaptation by providing guidance and assistance to test translators. Stronger partnerships among the original test author and test translators help provide better instruments in cross-cultural assessment. Cross-cultural work with the revised Minnesota Multiphasic Personality Inventory (MMPI–2) exemplifies how test authors can facilitate test translation and adaptation and promote training and cross-cultural research that leads to the high standards of its different language versions (see Butcher, 1996, 2004). The MMPI–2 international conferences and workshops exemplify venues where practitioners can gain knowledge of cross-cultural differences in their interpretation of test results.

A broader issue relates to the "ethnocentric" approach of trying to map cross-cultural data to the established models of psychological constructs originating in Western culture. As such, cross-cultural data commonly are made to conform to established etic (i.e., universal) models when they are inconsistent with the prevailing theory. Cross-cultural researchers should explore alternative models that may help explain the data more appropriately within the framework of the indigenous context.

The Chinese Personality Assessment Inventory, an indigenously derived personality measure for the Chinese culture, exemplifies how this caveat can be a source of enlightenment. Research findings that failed to replicate the purportedly universal Five-Factor Model of personality led F. M. Cheung et al. (2001) to reexamine the configuration of personality constructs within the context of Chinese society. Research results revealed six rather than five factors in a joint factor analysis that included the Big Five measure and the Chinese Personality Assessment Inventory, in which the indigenous factor of the Chinese Personality Assessment Inventory did not load on any of the Big Five factors (F. M. Cheung et al., 2008). Recent studies conducted in Europe and Asia also have identified additional factors beyond the Big Five, thus raising the questions of whether a dominant taxonomy such as the Five-Factor Model can be interpreted adequately within more culturally relevant frameworks and whether adoption of a universal taxonomy of personality traits is theoretically and empirically viable.

The cultural contexts that inform the knowledge base in assessment cannot be ignored as psychological assessment becomes globalized. The cultural perspective in assessment should be mainstreamed in psychology rather than being marginalized as peripheral interests. Cross-cultural training in assessment should become an integral part of graduate and professional training of all students, not only those interested in cross-cultural psychology.

Commentary on Multinational Test Use in Industrial/Organizational Practices

Multinational organizations assess their current and potential employees in ways that involve comparing people from diverse national, linguistic, and cultural backgrounds. This task raises a key question: When does the impact of culture on test scores matter? Attempts to answer this question necessarily generate three linked questions: (a) How do we define culture?, (b) What effects do cultural differences have on scores?, and (c) How do we aggregate data across cultures most appropriately? These three questions bear critically on the credibility of cross-cultural findings in general, particularly when they involve the norming of scores within and across national boundaries. For assessment purposes, culture matters only when it is related to some group level effect that is sufficiently large to result in misinterpretation of individual level scores.

The need to distinguish the terms *culture, nationality*, and *language* constitutes one of the major challenges when trying to define culture. A review of the literature reveals a wide range of descriptions pertinent to culture. Examples include shared values (e.g., a common understanding of the important aspects of life), shared cognitions (e.g., a common way of perceiving and making sense of the world), shared knowledge (e.g., a shared understanding of what constitutes "common sense" and assumptions of what others know), shared standards or cultural norms (e.g., social behavior, dress codes, expression of emotions in public), and shared language. Thus, while developing test norms, the test developer constantly is presented with the dilemma of whether samples should be aggregated according to nationality, language, or culture.

International test developers seek to establish norms both within and across national boundaries. Problems associated with definitions (e.g., how do we define a national culture?) and aggregation (e.g., who should be included in the sample?) complicate these norming efforts at both levels of inquiry. We turn first to the within-country (i.e., national) level. The practice of norming tests using national samples typically entails some acknowledgment of the ethnic component, yet often with no analysis of effect sizes associated with these cultural demographics. The unit of analysis and the level of data aggregation should not be defined in terms of some arbitrary political construct (e.g., nation) unless there is evidence that doing so corresponds to a single culture or homogeneous group. Definition of the unit of analysis should be tied to the operational definition of culture and to the basic notion of relative homogeneity within and heterogeneity between groups.

Consider now the aggregation of data from norm groups across countries and languages when defining the qualities of an appropriate reference group. The key question to answer here is stated clearly by Cronbach (Cronbach, 1990, p. 127): "Does the norm group consist of the sort of persons with whom [the candidate] should be compared?" Cronbach makes clear that an individual's score need not be compared with those from his or her own demographic group. An additional consideration is whether the test should be based on broad or narrow comparison groups. The broader the comparison group, the greater the degree of aggregation required.

The use of aggregated norms can benefit the test developer in at least two important ways: (a) by reducing country-related sample biases in which they are present and (b) by not concealing the effects of cultural differences that are hidden in the use of culture-specific norm groups. On the other hand, in the presence of true language biases, the use of aggregated norms has the potentially negative effect of treating country differences as a true form of bias rather than as a function of translation bias.

The following suggested practices address aggregation issues and their impact on testing practices across cultures. First, establish construct equivalence across cultures and obtain a clear understanding of the demographic composition of the various national, linguistic, and/or cultural norm groups. Second, whereas comparison of scores with the same (multinational, multicultural or multilingual) norm tends to accentuate cultural differences, the use of only country, language, or culture specific norms can reflect the relative levels of

underlying traits within that population without removing any cultural differences. Third, within international contexts, inferences from scores should take into account both an individual's position relative to his or her specific cultural norm as well as the effects of other relevant specific or aggregated norms on that individual's scores. Finally, comparisons between each language norm and an aggregated multinational norm show where individual profiles will diverge. This information should be made available to the interpreter, either through the norm information for the different groups or through qualitative data that identify where a particular pair of countries or norms differs.

Given the inadequacy of traditional academic curricula in applied psychology to provide both sufficient depth of psychometric knowledge and breadth of understanding regarding how psychometrics interact with measurement issues in the real world, my organization finds it necessary to conduct its own training of these basic and advanced skills for those psychologists entering the profession as test designers and developers. Although one would not anticipate the latter to be acquired outside of a practical testing environment, it seems perfectly reasonable to assume that the basic knowledge and skills of psychometrics are more widespread. Nonetheless, national differences do exist. For example, although a number of institutions provide excellent psychometrics training in the United States, Spain, and the Netherlands, the same cannot be said for the United Kingdom and much of the rest of Europe in which few, if any, academic courses do more than furnish a basic introduction to psychometrics for applied psychologists in training. This reality comes at a time when the demands on the technical expertise of test designers and developers as well as test users are becoming greater due to the advent of new types of Internet-delivered instruments and increasing use of tests for cross-cultural comparisons. The issue of when and how to aggregate data into norm groups, as discussed above, is just one of many that arises when large-scale cross-cultural assessments are conducted.

Current testing practices demand that psychologists and other assessment professionals engaged in cross-cultural comparisons have a thorough understanding of the various effects that language and culture can have on test design and item responses as well as knowledge of the tools needed to empirically assess these effects at both the item and scale levels. In addition, they need to understand the importance of linking judgmental and empirical procedures together in the adaptation of assessment instruments for use cross culturally. These requirements can be met by being knowledgeable of the methodological strategies of structural equation modeling and item response theory. Taken together, these requirements clearly point to a need for training that addresses the pragmatics of how tests are used in the real world, not just how they should be used in an ideal one. This need is best met when those offering training within academic environments work in concert with practitioners in the profession so that science and practice can each reinforce each other.

Implications for Updated Professional Training of Psychologists Conducting Cross-Cultural Research and Testing Practices

Two parallel trends in psychological research that emerged during the last decade were identified above, one associated with cross-cultural psychology and the other with quantitative psychology. A review of the psychological literature during the last decade reveals a rapid escalation in the number of cross-cultural studies and advances in methodology, including statistical applications (e.g., structural equation modeling, item response theory). Unfortunately, their development has not been mutually interactive. The use of advanced statistical procedures can help resolve many problems encountered in cross-cultural research (e.g., structural and measurement equivalence, adequacy of translation, detection of item bias). However, those conducting cross-cultural studies rarely use these strategies. This oversight suggests an absence of knowledge of the extent to which these statistical methods can address important issues as well as an absence of training in their appropriate use when engaged in cross-cultural research.

The results of a study of the availability of measurement and other quantitative courses in psychology published 17 years ago by Aiken, West, Sechrest, and Reno (1990) suggested little to no advancement, perhaps even a decline, during the preceding two decades. A replication of this study (Aiken, West, & Millsap, 2008) reported continued evidence of inadequate quantitative training in psychology PhD programs, particularly with respect to research design. However, Rossen and Oakland's (2008) recent survey of measurement and quantitative courses in 192 APA-accredited doctoral level professional programs (i.e., in clinical, counseling, school, and combined) reported significant improvements in professional psychology preparation during the last two decades. Most programs require introductory methods courses. Many programs offered advanced methods courses as electives although few were required. For example, a course in structural equation modeling was required by 18% and offered as an elective in 46%. Item response theory was required in 11% and offered as an elective in 22%. Few differences exist between PhD and PsyD programs with respect to the number and type of research methods courses offered.

Van de Vijver and colleagues (van de Vijver & Leung, 1997, 2000; van de Vijver & Poortinga, 2002) dedicated their careers to cross-cultural psychology, including the preparation of cross-cultural psychologists. They and other leaders have urged cross-cultural psychologists and others engaged in multicultural assessments to address the many methodological issues highlighted in this article. These recommendations are important for those interested in conducting cross-cultural research or striving to obtain a more informed understanding of the method and results sections found in published articles derived from this form of research.

This article has emphasized the importance of efforts to improve the academic preparation of graduate students, particularly in advanced courses, and to update the training of psychologists in research methods to promote the credibility and use of results derived from cross-cultural research. For practitioners, these advancements translate into a need for improved professional training in the proper use of tests and interpretation of their results when assessments involve members of diverse cultural groups. To this end, Geisinger and Carlson (1998) suggested a model of instruction that comprised several modules, each of which addressed a specific aspect of this assessment process. Although the reality of journal space limitations precludes the inclusion of specific examples of applied training pertinent to the issues raised in this article, we are hopeful that readers will find the suggested pedagogically oriented citations that follow to be both helpful and informative in the revision and/or selection of professional training programs. These citations address the broad issues of instrument equivalence, test bias, normative equivalence, and indigenous instrument development.

Instrument Equivalence

The key concern in the use of instruments across diverse cultural groups is the extent to which item content is similarly perceived and their underlying constructs similarly structured across groups. Evidence of nonequivalence related to both can be determined through the application of structural equation modeling; at least three pedagogical variants of this technique are worthy of mention here. First, for a detailed introduction and annotated application of this procedure across cultural groups, see Byrne (2008). Second, for an extensive explanation of this methodology in testing for the construct validity of scores derived from an instrument developed in one country (United States) and then adapted for use in another (China) as they relate to (a) use within China, and (b) equivalence with scores from the original American instrument can be found in Byrne, Stewart, and Lee (2004) and Byrne, Stewart, Kennard, and Lee (2007), respectively. Finally, for readers who may wish to be "walked through" these techniques, Byrne (1998, 2001, 2006) partners explanations of the process with detailed reviews of input and output files related to the three most commonly used structural equation modeling programs.

Differential item functioning, a term used to describe nonequivalent test items, is typically identified within the framework of an item response theoretical approach. A 1998 pedagogically oriented article by Clauser and Mazor as well as two popular books on the topic (Camilli, & Shepard, 1994; Holland & Wainer, 1993) should be helpful to interested readers in comprehending both the issues and application of this approach to the detection of nonequivalent test items.

Test Bias

Evidence of nonequivalent scores related to a measuring instrument derives from some type of bias associated either with the construct, the data collection or administration methods, and/or the items themselves. For an excellent comprehensive discussion and explanation of test bias as it relates to the testing of culturally diverse groups, readers are referred to van de Vijver and Poortinga (2005).

Normative Equivalence

As noted earlier, when an instrument is adapted for use in a culture that differs from the one in which it was originally developed, there is a need to reestablish the norms relative to the new culture and, if appropriate, to equate them with the original version (Geisinger, 1994). When instruments are developed for use in multiple cultures, such as those developed by multinational test companies, procedures become substantially more complex. Typically, details related to these development and standardization processes are presented in the accompanying technical manual. One such document outlining the steps followed in developing the Occupational Personnel Questionnaire 32 (OPQ32; Bartram, Brown, Fleck, Inceoglu, & Ward, 2006) can be downloaded from www.shl.com/opqtechnicalmanual. A second example of this multicultural standardization process is outlined by Schmitt, Kihm, and Robie (2000) as it relates to development of the Global Personality Inventory.

Indigenous Instrument Development

Two pedagogical papers detailing the many and varied steps involved in establishing an indigenous instrument are recommended. The first of these illustrates the steps taken in establishing the Chinese Personality Assessment Inventory (F. M. Cheung et al., 1996); the second (F. M. Cheung et al., 2008) illustrates how a combined emic-etic approach was used in developing culturally relevant Openness scale for this personality instrument.

Psychologists and other mental health professionals interested in acquiring or updating their training programs to include methodological approaches suggested in this article can of course also obtain needed information from books, workshops, training seminars, and by taking graduate courses. In addition to the Byrne (1998, 2001, 2006), Holland and Wainer (1993), and Camilli and Shepard (1994) books noted earlier, Hambleton et al. (1991) and Embretson and Reise (2000) discussed the basic concepts and applications of item response theory, and Hambleton et al. (2005) discussed ways to adapt measuring instruments for use across cultures. The APA's Continuing Education Office regularly offers half-day and full-day workshops at its annual conventions; many topics addressed in this article have been presented in these training sessions. The APA also offers in-depth advanced training seminars on methodological topics related to those addressed in this article. Many workshops offered during the ITC conference, the International Association of Applied Psychology, the International Congress of Psychology, and the European Congresses of Psychology also are pertinent to cross-cultural psychologists.

Changes in graduate preparation, including practicum and internships, are needed to materially improve the preparation of psychologists to engage in and become informed consumers of cross-cultural research. A basic set of cross-cultural research competencies, as illustrated by the methodological issues discussed in this paper, may need to be added to the core curriculum. Although all graduate students may benefit from such preparation, realism dictates that some will benefit more than others, given their personal and career interests. Thus, a two-pronged effort is needed: To increase the number of graduate programs that specialize in cross-cultural methods for those more keenly interested in this specialization, and to encourage graduate students whose major interests lie elsewhere to take foundation courses that at least prepare them to be informed consumers of cross-cultural research.

Change in higher education often is slow and always requires the involvement of various parties, including faculty interested in and prepared to teach advanced courses on the issues addressed in this article, administrators who support such changes, and graduate students who either are required or elect to obtain recommended preparation. In addition, institutional, state, and national policies also may require revision, including efforts by the APA's Commission on Accreditation. Efforts to promote needed changes also must consider various trends impacting quantitative methods, including low enrollment of graduate students who specialize in quantitative methods, advanced courses offered as electives rather than required, as well as declining numbers of professors able to teach advanced courses. Thus, despite the somewhat urgent need, changes to psychology program curricula represent a major undertaking that will require several years to fully implement. Finally, professional organizations dedicated to the advancement of research methods are encouraged to form a joint task force to establish best practice guidelines for conducting and reporting cross-cultural research.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63,* 32–50.

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45,* 721–734.

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington DC: Author.

American Psychological Association. (2002). *Ethical principles of psychologists and codes of conduct.* Washington DC: Author.

American Psychological Association. (2003). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist, 58,* 377–402.

Baddeley, A. (1997). *Human memory: Theory and practice* (Rev. ed.). Hove, England: Psychology Press.

Bartram, D. (2001). International guidelines for test use. *International Journal of Testing, 1,* 93–114.

Bartram, D., Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *Technical manual* (Pt. 1, chapters 1–6). Thames-Ditton, England: SHL Group.

Bartram, D., & Hambleton, R. (Eds.). (2006). *Computer-based testing and the Internet.* West Sussex, England: Wiley.

Butcher, J. N. (Ed.). (1996). *International adaptations of the MMPI–2: A handbook of research and applications.* Minneapolis: University of Minnesota Press.

Butcher, J. N. (2004). Personality assessment without borders: Adaptation of the MMPI–2 across cultures. *Journal of Personality Assessment, 83,* 90–104.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring* (Rev.ed.). Minneapolis, MN: University of Minnesota Press.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Byrne, B. M. (2006). *Structural equation modeling with EQS* (2nd ed.): *Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psichothema, 20,* 872–882.

Byrne, B. M., Stewart, S. M., Kennard, B. D., & Lee, P. (2007). The Beck Depression Inventory II: Testing for measurement equivalence and factor mean differences across Hong Kong and American Adolescents. *International Journal of Testing, 7,* 1–17.

Byrne, B. M., Stewart, S. M., & Lee, P. W. H. (2004). Validating the Beck Depression Inventory–II for Hong Kong Community Adolescents. *International Journal of Testing, 4,* 199–216.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34,* 155–175.

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

Cheung, F. M. (2004). Use of Western- and indigenously-developed personality tests in Asia. *Applied Psychology: An International Review, 53,* 173–191.

Cheung, F. M. (2006). A combined emic–etic approach to cross-cultural personality test development: The case of the CPAI. In Q. Jing, H. Zhang & K. Zhang (Eds.), *Psychological science around the world* (Vol. 2, pp. 91–103). London: Psychology Press.

Cheung, F. M., Cheung, S. F., Zhang, J. X., Leung, K., Leong, F. T. L., & Yeh, K. H. (2008). Relevance of openness as a personality dimension in Chinese culture. *Journal of Cross-Cultural Psychology, 39,* 81–108.

Cheung, F. M., Leong, F. T. L., & Ben-Porath, Y. (2003). Psychological assessment in Asia: Introduction to the special section. *Psychological Assessment, 15,* 243–247.

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

Cheung, F. M., Leung, K., Zhang, J. X., Sun, H. F., Gan, Y. Q., Song, W. Z., et al. (2001). Indigenous Chinese personality construct: Is the Five Factor Model complete? *Journal of Cross-Cultural Psychology, 32,* 407–433.

Cheung, M. W.-L., & Au, K. (2005). Applications of multilevel structural equation modeling to cross-cultural research. *Structural Equation Modeling, 12,* 598–619.

Cheung, M. W.-L., Leung, K., & Au, K. (2006). Evaluating multilevel models in cross-cultural research: An illustration with social axioms. *Journal of Cross-Cultural Psychology, 37,* 522–541.

Church, A. T., & Lonner, W. J. (1998). The cross-cultural perspective in the study of personality. *Journal of Cross-Cultural Psychology, 29,* 32–62.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–44.

Coyne, I., & Bartram, D. (Eds.). (2006). ITC guidelines on computer-based and internet-delivered testing [Special issue]. *International Journal of Testing, 6*(2).

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper Row.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Faucheux, C. (1976). Cross-cultural research in experimental social psychology. *European Journal of Social Psychology, 6,* 269–322.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Hambleton, R., Merenda, P., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

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

Hofstede, G. (2001). *Culture's consequences* (2nd ed.). Thousand Oaks, CA: Sage.

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

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national research. *Zuma Nachrichten Spezial, 3,* 1–40.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8,* 325–352.

Koocher, G., & Keith-Spiegel, P. (2007). *Ethics in psychology* (3rd ed.). New York: Oxford University Press.

Leach, M., & Oakland, T. (2007). Ethics standards impacting test development and use: A review of 31 ethics codes impacting practices in 35 countries. *International Journal of Testing, 7,* 71–88.

Leung, K. (1989). Cross-cultural differences: Individual-level vs. culture-level. *International Journal of Psychology, 24,* 703–719.

McCrae, R. R., Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology, 88,* 547–561.

Oakland, T. (2005). Selected ethical issues relevant to test adaptations. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 65–92). Mahwah, NJ: Erlbaum.

Poortinga, Y. (1995). Use of tests across culture. In T. Oakland & R. K. Hambleton (Eds.), *Interpersonal perspectives on academic assessment* (pp. 187–206). Norwell, MA: Kluwer.

Rossen, E., & Oakland, T. (2008). Graduate preparation in research methods:

The current status of APA-accredited professional programs in psychology. *Training and Education in Professional Psychology, 2,* 42–49.

Sam, D. L., & Berry, J. W. (Eds.). (2006). *The Cambridge handbook of acculturation psychology.* Cambridge, England: Cambridge University Press.

Sattler, J. (1998). *Clinical and forensic interviewing of children and families.* Seattle, WA: Author.

Schmitt, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology, 53,* 153–193.

Segall, M. H., Lonner, W. J., & Berry, J. W. (1998). Cross-cultural psychology as a scholarly discipline: On the flowering of culture in behavioral research. *American Psychologist, 53,* 1101–1110.

Shebani, M. F. A., van de Vijver, F. J. R., & Poortinga, Y. H. (2008). Memory development in Libyan and Dutch school children. *European Journal of Developmental Psychology, 5,* 419–438.

van de Vijver, F. J. R. (2006, July). *Toward the next generation of instruments in cross-cultural testing: Recent developments in translations and adaptations.* Invited address at the ITC Conference on Test Adaptations, Brussels, Belgium.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

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

van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Research, 33,* 141–156.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263–279.

van de Vijver, F. J. R., van Hemert, D. A., & Poortinga, Y. H. (2008). Conceptual issues in multilevel models. In F. J. R. van de Vijver, D. A. Hermert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures.* Mahwah, NJ: Erlbaum.

van de Vijver, F. J. R., & Watkins, D. (2006). Assessing similarity of meaning at the individual and country level: An investigation of a measure of independent and interdependent self. *European Journal of Psychological Assessment, 22,* 69–77.

Received February 19, 2008
Revision received September 8, 2008
Accepted September 25, 2008 ■