PRACTICAL RESOURCES for the Mental Health PROFESSIONAL

Culture and Children's Intelligence

Cross-Cultural Analysis of the WISC-III

Edited by
James Georgas
Lawrence G. Weiss
Fons J.R. van de Vijver
Donald H. Saklofske

The sponsoring editor for this book was Nikki Levy, the senior developmental editor was Barbara Makinster, and the senior project manager was Paul Gottehrer. The cover was designed by Cathy Reynolds. Composition was done by Cepha Imaging Pvt. Ltd., Bangalore, India and the book was printed and bound by Maple-Vail, York, PA.

Cover photo credit: Banana Stock © 2003.

This book is printed on acid-free paper. ♾

Academic Press
An imprint of Elsevier Science
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.academicpress.com

Academic Press
84 Theobald's Road, London WC1X 8RR, UK
http://www.academicpress.com

Library of Congress Catalog Card Number: 2003104541

International Standard Book Number: 0-12-280055-9

PRINTED IN THE UNITED STATES OF AMERICA
03 04 05 06 07 7 6 5 4 3 2 1

18

Methodology of Combining the WISC-III Data Sets

Fons J. R. van de Vijver

Department of Psychology
Tilburg University
Tilburg, The Netherlands

Kostas Mylonas

Department of Psychology
The University of Athens
Athens, Greece

Vassilis Pavlopoulos

Department of Psychology
The University of Athens
Athens, Greece

James Georgas

Department of Psychology
The University of Athens
Athens, Greece

This chapter consists of three parts. In the first part an overview is given of which items were adapted and which items were closely translated per subtest in each of the countries. This presentation provides the background for the statistical analyses reported in Chapter 19; however, the overview is also interesting in its own right. It provides insight in the judgmental bias of the subtest, which refers to nonstatistical procedures to identify bias, based on a content analysis of the items. All local test development teams had to address two questions: (1) which American items were expected to be transferable to a new linguistic and cultural context without major alterations and (2) which items were assumed to require adaptations. As a consequence, country comparisons of the number of adapted items of the 11 subtests provide information about the judgmental bias in these subtests. The second part of this chapter describes (in a largely nontechnical way) the statistical analyses that are reported in Chapter 19. Conclusions are drawn in the third part.

OVERVIEW OF TEST ADAPTATIONS PER COUNTRY

In Chapter 17, a distinction was made between three ways of translating Wechsler Intelligence Scale for Children—Third Edition (WISC-III) subtests: applications (i.e., close translations of items), adaptations (i.e., change of item contents in order to enhance the suitability of an item for a particular cultural context), and assemblies (i.e., the development of a completely new instrument, needed when the original instrument would be entirely inappropriate in the new culture). Aggregated across countries, the vast majority of the items in the WISC-III adaptations described in this book, about 90%, has been closely translated or simply copied (in the case of pictorial stimuli), while a small minority of the items, about 10%, has been adapted. Assemblies were not used.

A detailed overview of the similarity of each subtest in each country to the U.S. subtest is presented in the Appendix. The tables indicate for each subtest whether the item was closely translated or adapted. In the case of a close translation, an item may appear in a different place in the item order. The order of the items is always determined by the empirically observed difficulty order. As a consequence, the order in a specific country may differ somewhat from the order in the U.S. subtest. The tables in the Appendix contain information about both the nature of the translation (application or adaptation) and the rank order of the closely translated items. So, the tables have three types of items: closely translated items with the same position in the item order in the U.S. as in a target country, closely translated items that have moved to a different place in the item rank order, and adapted items.

From a bias perspective, an interesting feature of the Appendix involves the proportion of adapted items across subtests and countries. An overview of these proportions is given in Table 18.1. The rows of the table present the country names in ascending order of proportions of adapted items. Analogously, the columns present the subtests in increasing order of their proportion of adapted items. A comparison of countries shows that, as could be expected, the smallest number of adaptations were found in the three English-speaking countries (Australia, Canada, and the UK). The largest proportions are found in Japan, The Netherlands and Flanders, and France. The rank order of the countries has some face validity in that culturally and geographically proximate countries tend to be close to one another. However, there was one notable exception. Whereas Korea and Taiwan are close to each other (with relatively low numbers of adapted items), Japan had the highest proportion of all countries. The reason behind the deviant position of Japan is not clear.

A column-wise comparison of Table 18.1 shows that the items of Object Assembly and Digit Span are identical in all countries. Block Design, Mazes, Picture Arrangement, and Picture Completion are identical in all countries except Japan; the test constructors in each country apparently judged that the performance subtests were culturally appropriate in their cultures. Thus, the Performance

TABLE 18.1 The Proportion of Adapted Items per Subtest and Country (in Order of Increasing Row and Column Means)

Country[a]	Object assembly	Digit span	Block design	Mazes	Picture arrangement	Picture completion	Arithmetic	Similarities	Comprehension	Information	Vocabulary	*Mean*
AUS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CAN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UK	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.07	0.00	0.01
LITH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.23	0.03
SLO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.17	0.23	0.04
TAI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.10	0.30	0.04
KOR	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.05	0.17	0.20	0.23	0.07
GRE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.37	0.33	0.08
SWE	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.43	0.43	0.09
GER	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.11	0.17	0.17	0.37	0.11
FRA	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.16	0.56	0.33	0.77	0.17
DUT	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.29	0.53	0.32	0.77	0.20
JAP	0.00	0.00	0.08	0.10	0.21	0.40	0.08	0.37	0.28	0.57	0.93	0.28
Mean	0.00	0.00	0.01	0.01	0.02	0.03	0.08	0.09	0.15	0.22	0.35	0.09

Note: The country of reference is the U.S.; cell means represent the proportion of subtest items that were adapted from the U.S. subtest version.

[a]Country codes: AUS: Australia; CAN: Canada; DUT: The Netherlands and Flanders (Dutch language area); FRA: France and Francophone Belgium; GER: Germany, Austria, and Switzerland (German language area); GRE: Greece; JAP: Japan; KOR: South Korea; LITH: Lithuania; SLO: Slovenia; SWE: Sweden; TAI: Taiwan; UK: United Kingdom; U.S.: United States of America.

subtests showed fewer adaptations than the Verbal subtests. The largest score was obtained by Vocabulary (with 35% of adapted items).

A possibly less apparent, though salient feature of Table 18.1 is the overall patterning of the proportions; going from left to right and from the top to the bottom, numbers tend to increase. Although there are various distortions of the consistency, there appeared to be quite some agreement among the local teams about which U.S. subtests needed more and which subtests required less adaptation. An intraclass coefficient, measuring consistency across countries (with subtests as replications) showed a highly significant value of 0.92 ($p < 0.001$). An intraclass correlation estimating the consistency across subtests (with countries as replications) obtained a significant value of 0.83 ($p < 0.001$). Both values indicate that there is a remarkable consistency across the various teams who developed the local versions of the WISC-III about which subtests travel better.

In sum, the judgmental bias analysis shows a clear result: there is considerable agreement across countries about the question which subtests are more susceptible to bias. However, as indicated by the large country differences in an average number of adapted items, the local teams did not agree on the absolute proportions that required adaptation.

CROSS-CULTURAL DATA ANALYSES

The cross-cultural data analysis consists of three parts. The first part addresses the structural equivalence of the subtests. To what extent is the cognitive structure identical across cultures that presumably underlies test behavior? It was stated in Chapter 17 that much cross-cultural work, dealing with the different versions of the WISC (usually the WISC-R) as well as other intelligence tests, has shown the stability of the cognitive structure underlying intelligence tests (e.g., Kush *et al.*, 2001; Naglieri & Jensen, 1987; Reschly, 1978; Sandoval, 1982; Taylor & Ziegler, 1987; Valencia, Rankin, & Oakland, 1997). However, as usual, the proof of the pudding is in the eating: We need to demonstrate that this also holds for the WISC-III subtests in the twelve data sets of our study. The WISC-III was adapted to 16 countries: the U.S., Canada, UK, Austria, Germany, and Switzerland, France and French-speaking Belgium, The Netherlands and Dutch-speaking Belgium, Greece, Sweden, Slovenia, Lithuania, Japan, South Korea, and Taiwan. However, there are 12 data sets: the U.S., Canada, The Netherlands and Flanders (Dutch language area), France and French-speaking Belgium, Germany, Austria, and Switzerland (German language area), Greece, Japan, South Korea, Lithuania, Slovenia, Sweden, and Taiwan. Data are not available from the UK. In the unlikely event that we would not find evidence to support the structural equivalence of the subtests, a further analysis of the nature of the differences would be needed.

If we find evidence to support the structural equivalence of the subtests, we can proceed with the next step. The second type of statistical analysis addresses the metric equivalence of the data (measurement-unit and full-score equivalence).

As argued before, we treat these together in order to avoid the theoretically complex and ideology-laden discussion whether country differences in raw scores on WISC-III subtests are due to valid differences or to method bias (e.g., differential stimulus and test familiarity).

The third stage of the analysis further addresses the nature of the cross-cultural differences in scores in more detail. The importance of the distinction between measurement-unit and full-score equivalence is seen here as less important than often suggested in the literature. We do not claim that the present data set can settle an issue that has been around for so long: the nature of the cross-cultural differences in scores on intelligence tests (e.g., Bruner, 1914; Burks, 1928; Herrnstein & Murray, 1994; Jensen, 1980; Leahy, 1935; Vernon, 1969, 1979). We argue in favor of a more pragmatic perspective, in which we examine the nature of the cross-cultural score differences by relating these to various country-level indicators, such as educational expenditure (per capita). Thus, Van de Vijver (1997) found that educational expenditure could partly explain cross-cultural differences in cognitive test scores. Correlations between these country-level indicators and observed cross-cultural score differences on raw test scores on the WISC-III subtests can provide important information about the nature of these differences.

In the following section each of the statistical analyses are described in more detail.

ANALYSIS OF STRUCTURAL EQUIVALENCE

Various statistical techniques can be employed to examine structural equivalence, such as factor analysis (exploratory or confirmatory), multidimensional scaling, and cluster analysis (Van de Vijver & Leung, 1997). In the context of intelligence tests exploratory factor analysis has been used most frequently. This technique is also used here for different reasons. First and foremost, factor analysis is a tried and tested procedure for examining the similarity of the cognitive structure underlying intelligence tests. Second, in the initial stage of the analysis the interest is not in the measurement unit of the subtests (in which confirmatory factor analysis would be the preferred choice) but in the factors underlying the battery. As a consequence, an analysis is needed that is based on correlations (rather than covariances). Finally, the usage of exploratory factor analysis allows us to relate our findings to common findings in the literature (notably the discussion on the dimensionality of the WISC-III; cf. Allen & Thorndike, 1995; Blaha & Wallbrown, 1996; Bracken & McCallum, 1993; Kush *et al.*, 2001; Prifitera & Saklofske, 1998; Reynolds & Ford, 1994).

An examination of the structural equivalence of an intelligence battery could start with an analysis of the similarity of the structure of the subtests, followed by an analysis of the subtest scores. The first step scrutinizes item scores and the second subtest scores (raw or standard scores). Unfortunately, the opportunity to analyze the subtests separately and to examine the structural equivalence of the subtests is very limited. For the speeded tests item responses are typically not

recorded and if they would be recorded, they would not yield meaningful intercorrelations because of the small number of errors per item. Furthermore, for the adapted tests (i.e., subtests containing items that were not closely translated but substituted) only the items that are common can be examined for structural equivalence. As will become clearer in the second part of this chapter, this number varies across the subtests and countries involved; yet, for some subtests the number of items shared with the American subtests (from which the other language versions were developed) is very small, thereby precluding a meaningful analysis. Another problem in the item-level factor analyses is the variable number of respondents across items. More difficult items are answered by fewer children. The factor analysis would have to accommodate this difference in number of respondents. Finally, we share the preference in the literature to focus on subtest scores in the study of structural equivalence. If we are interested in comparing cognitive structures across cultures, subtest scores provide a more comprehensive picture than item-level scores. In sum, without belittling the value of item-level studies when addressing structural equivalence, we focus here on the analysis of subtest scores.

The procedure to use exploratory analysis in the study of structural equivalence has two stages. In the first stage the factor analyses are carried out (suppose that we want to compare the American and British data). The American and British data are analyzed separately. The factor loadings (suppose that four factors were extracted) are then compared to each other. Rotations are an important part of factor analysis, aimed at increasing the interpretability of the solution. However, rotation methods have some arbitrariness. This arbitrariness needs to be resolved before the similarity of the factors can be evaluated. One of the countries is designated as target, (e.g., the American factor loadings. Usually the choice is immaterial; however, in this study, because the WISC tests were originally developed in the U.S., this country is the obvious target. Thus, the Greek factor loadings are rotated so as to maximize their similarity with the American factor loadings. It should be noted that this is not done to artificially boost the similarity, but the need for target rotations is a consequence of the arbitrariness of the rotations in the two country solutions.

The most frequently employed agreement (or congruence) coefficient is Tucker's phi (originally due to Burt). The coefficient is insensitive for positive constants with which loadings are multiplied (Tucker's phi between the loadings the vectors {0.1, 0.2, 0.3} and {0.2, 0.4, 0.6} is perfect as the values of the second vector are exactly twice as large as the value of the first vector). However, the coefficient is sensitive to differences in means (Tucker's phi between the loadings of vectors {0.1, 0.2, 0.3} and {0.2, 0.3, 0.4} is not perfect). The insensitivity to differences in (positive) multiplicative constants is deliberately chosen; it makes the coefficient invariant for differences in eigenvalues of the factors. The underlying idea is that differences in the reliability of factors across cultures do not challenge the structural equivalence of the factors. Values of Tucker's phi larger than 0.90 are often taken to indicate equivalent factors. More recently, this value has been challenged by Van de Vijver and Poortinga (1994), who showed in a simulation

study that values substantially higher than 0.90 can be obtained even when one or two items show markedly different loadings on factors with high eigenvalues (factors with many items with high loadings). The major problem with congruence coefficients is that statistical tests cannot be performed on these coefficients as their sampling distribution is unknown. To address this problem, Chan, Ho, Leung, Chan, and Yung (1999) propose using a bootstrap procedure to estimate the standard error of Tucker's phi. A visual inspection of differences in loadings of the target matrix and rotated source matrix may also help to identify anomalous items.

In our description of the factor analytic procedure we conveniently simplified the analysis to two groups. However, we employ 12 data sets, which means that we have 66 (= $(12 \times 11)/2$) comparisons per factor. The obvious question to address is how it is possible to deal with such a multitude of comparisons. Two types of different approaches can be envisaged to deal with comparisons involving a larger number of countries (see Welkenhuijsen-Gybels & Van de Vijver, 2002, for a more elaborate description). The first is a "one-to-all" approach. It amounts to pooling all data in a single data matrix (controlling for confounding differences in mean country-levels on the variables studied). Statistical details have been described by Van de Vijver and Poortinga (2002; see also Muthén, 1994). The factor analysis of the pooled data yields the target solution with which the factor solutions of each country are compared. The agreement is evaluated by means of Tucker's phi. If all countries show values larger than some critical value, evidence for structural equivalence is obtained. If some values show lower values, structural equivalence is not supported and reasons for the lower values need to be identified.

The second is a "one-to-one" procedure, in which all country comparisons are computed. The advantage of this procedure is its level of detail; however, from a computational perspective the procedure is cumbersome. The output of the procedure is a country-by-country matrix with values of Tucker's phi in the cells (there is one matrix per factor). Each cell indicates the level of similarity between two factors obtained in the two countries. Cluster analysis or multidimensional scaling can be employed to identify homogenous groups of countries with a high internal factorial for each cluster.

In principle, both approaches could be applied here. An attractive feature of the first approach is that it starts from the overall pooled solution that is based on data from 12 data sets. If the factor solutions in all countries show a good agreement with this overall structure, the latter is the global structure which has a wide applicability and is based on an impressive sample size. However, when there are country differences in structure, a comparison to a pooled solution may mask the patterning of the differences. In the latter case a one-to-one comparison would be more informative.

In conclusion, the analysis of structural equivalence consists of three parts. The first is a factor analysis per country of the subtest scores; four factors are extracted. In order to control for age effects standard scores are used. In the second

stage the factorial agreement for all pairs of countries is computed. Four matrices, indicating the factorial agreement for each pair of countries, form the output of this stage. Finally, per factor a cluster analysis is employed to identify sets of countries with a high factorial agreement.

METRIC EQUIVALENCE

Assuming an affirmative answer to the question of the structural equivalence of the WISC-III, possibly amended for the fourth factor, we can continue with an analysis of metric equivalence. The adapted items may seem to preclude a direct comparison of raw scores across countries. The first and easiest solution would be to disregard the adapted items and to restrict the comparison to the items, used in all countries. A quick scan of the tables in the Appendix quickly reveals the unattractiveness of this option. If we would be forced to restrict the comparison to the common items, the comparisons would be based on very small item numbers, which would challenge the validity and replicability of the comparison. So, if we were to restrict the comparison to common items, some subtest comparisons not involving the U.S. may be based on small item sets.

The problem can be tackled using Item Response Theory (for introductions, see Fischer & Molenaar, 1995; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). Item Response Theory assumes that a person has a certain ability, which is statistically estimated on the basis of his or her responses on the subtest items; analogously, each item has a certain difficulty level, which is estimated on the basis on the number of correct responses to the item. The theory assumes that all items of a subtest measure the same underlying trait in all countries involved; in the case of cross-cultural data we also need to assume that items have the same item parameters in all countries involved (items should not be biased in favor of or against any country). Furthermore, Item Response Theory assumes local independence, which formally means that within groups of equal ability responses are statistically independent. In more informal terms, the assumption means that each subject answers each item independently and that there are no carry-over effects across items (e.g., effects due to memory or fatigue). If these assumptions are met, Item Response Theory allows for the estimation of a person's ability even if not all items are identical in all groups. As long as there are "anchors" (a set of items that are common to the countries to be compared), Item Response Theory can estimate the ability level of each person (and, by implication, the mean ability level of the country).

What does the application of Item Response Theory amount to in the present case? Let us take Vocabulary as an example. Suppose that two countries have 10 common and 20 country-specific items. We need to assume that in the two countries each of the 30 items measures the same underlying trait, say vocabulary knowledge (and nothing else than this trait). We also assume that the difficulty level of the common items is identical. The occurrence of an item that is much

more difficult in one country, compared to the difficulty of the other items, would invalidate the applicability of Item Response Theory for that item; an item that appears very early in the list of words in one country should not appear at the end of the subtest in another country. Finally, we assume the absence of carry-over effects. If this set of (restrictive) assumptions is met, we can use both the 10 common items (as anchor) and the 20 country-specific items (as helping to improve the estimate of the ability of each person). The average of these person's abilities gives us a country mean for that subtest, in which all responses of all children have been used.

NATURE OF THE CROSS-CULTURAL SCORE DIFFERENCES

The distinction between measurement unit and full score equivalence is difficult to make in the present data set. Full score equivalence requires the absence of any form of bias; when dealing with a large number of cultures it is difficult to demonstrate the absence of bias. Therefore, a more modest approach is chosen here in which the nature of the cross-cultural score differences observed is examined by correlating these differences with country indicators. In other words, we validate the cross-cultural score differences by examining their nomological network (Cronbach & Meehl, 1955). In this analysis we shift the focus from the individual to the country level. Each country is represented by its average score on each subtest.

The study of psychological indicators at country level is gaining popularity in cross-cultural psychology (Georgas & Berry, 1995; Georgas, Van de Vijver, & Berry, 2003; Lynn & Martin, 1995; McCrae & Allik, 2002; Poortinga, Van de Vijver, & Van Hemert, 2002; Van de Vijver, 1997; Van Hemert, Van de Vijver, Poortinga, & Georgas, 2002). Building on Berry's Ecocultural Framework, Georgas *et al.* (2002) examined the patterning of cross-cultural differences in attitudes and values (Hofstede, 2001; Schwarz, 1992) and subjective well-being (Inglehart, 1997). Two sets of country indicators were found to be relevant: religion and affluence. Some religions, such as Islam and traditional beliefs, and countries with low levels of affluence were found to place more emphasis on interpersonal aspects, such as power, loyalty, and hierarchy. Protestant countries and more affluent countries placed more emphasis on intrapersonal values, such as individualism and well-being. Van de Vijver (1997) examined cross-cultural differences in scores on cognitive tests. He found affluence to be related to these differences. More specifically, educational expenditure (per capita) was a good predictor of cross-national score differences.

From a statistical perspective the country-level studies mentioned are usually simple and straightforward. Correlational and multiple regression analyses are employed to examine relationships between psychological variables and country indicators. In the present study correlations will be used. Correlations are computed between country indicators and subtest scores (either the mean raw scores for

the subtests that are identical across the countries or the means of the ability distributions based on Item Response Theory in the case of subtests with adapted items).

CONCLUSION

The quality of large cross-cultural projects such as described in this book is based on a combination of substantive and methodological factors. The current chapter focused on the latter. The approach adopted in this chapter starts with a careful analysis of the question to what extent the various subtests of the WISC-III measure the same underlying construct(s). A comparison of the factor structures obtained in each country to the structure of all countries combined is expected to yield valuable information about the similarity of the structure of intelligence underlying the WISC-III in the countries examined. Based on the literature and on the judgmental bias analysis, which showed a remarkable consistency across countries about which subtests needed more and which needed fewer adaptations, we expect to find a good cross-cultural agreement of the factor structure. In the next step mean scores are computed per subtest and country. Item Response Theory is used to deal with the problem that several subtests do not have exactly the same number of items. Finally, these mean scores are related to country-level variables (such as educational expenditure) in order to understand the nature of the differences in scores.

REFERENCES

Allen, S. R., & Thorndike, R. M. (1995). Stability of the WAIS-R and WISC-III factor structure using cross-validation of covariance structures. *Journal of Clinical Psychology, 51,* 648–657.

Blaha, J., & Wallbrown, F. H. (1996). Hierarchical factor structure of the Wechsler Intelligence Scale for Children-III. *Psychological Assessment, 8,* 214–218.

Bracken, B. A., & McCallum, R. S. (1993). *Wechsler Intelligence Scale for Children—Third Edition.* Brandon, VT: Clinical Psychology Publishing Company.

Bruner, F. G. (1914). Racial differences. *Psychological Bulletin, 11,* 384–386.

Burks, B. S. (1928). The relative influence of nature and nurture upon mental development: A comparative study of parent-foster child resemblance. *Twenty-Seventh Yearbook of the National Society for the Study of Education, 27,* 219–316.

Chan W., Ho, R. M., Leung, K., Chan, D. K.-S., & Yung, Y.-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods, 4,* 378–402.

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

Fischer, G. H., & Molenaar I. W. (1995). *Rasch models. Foundations, recent developments, and applications.* New York: Springer-Verlag.

Georgas, J., & Berry, J. W. (1995). An ecocultural taxonomy for cross-cultural psychology. *Cross-Cultural Research, 29,* 121–157.

Georgas, J., Van de Vijver, F. J. R., & Berry, J. W. (2003). The ecocultural framework, ecosocial indices and psychological variables in cross-cultural research. *Journal of Cross-Cultural Psychology*. In press.

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

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

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

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

Inglehart, R. (1997). *Modernization and postmodernization: Changing values and political styles in advanced industrial society*. Princeton, NJ: Princeton University Press.

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

Kush, J. C., Watkins, M. W., Ward, T. J., Ward, S. B., Canivez, G. L., & Worrell, F. C. (2001). Construct validity of the WISC-III for White and Black students from the WISC-III standardization sample and for Black students referred for psychological evaluation. *School Psychology Review, 30*, 70–88.

Leahy, A. M. (1935). Nature-nurture and intelligence. *Genetic Psychological Monographs, 17*, 237–308.

Lynn, R., & Martin, T. (1995). National differences for thirty-seven nations in extraversion, neuroticism, psychoticism and economic, demographic and other correlates. *Personality and Individual Differences, 19*, 403–406.

McCrae, R. R., & Allik, J. (2002). *The five-factor model across cultures*. New York: Kluwer Academic/Plenum Publishers.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*, 376–398.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison of Black-White differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence, 11*, 21–43.

Poortinga, Y. H., Van de Vijver, F. J. R., & Van Hemert, D. A. (2002). Cross-cultural equivalence of the big five: A tentative interpretation of the evidence. In A. J. Marsella (Series Ed.), R. R. McCrae, & J. Allik (Eds.), *The five-factor model across cultures* (pp. 271–292). New York: Kluwer Academic/Plenum Publishers.

Prifitera, A., & Saklofske, D. H. (1998). *WISC-III clinical use and interpretation: Scientist-practitioner perspectives*. San Diego, CA: Academic Press.

Reschly, D. (1978). WISC-R factor structures among Anglos, Blacks, Chicanos, and Native-American Papagos. *Journal of Consulting and Clinical Psychology, 46*, 417–422.

Reynolds, C. R., & Ford, L. (1994). Comparative three-factor solutions of the WISC-III and WISC-R at 11 age levels between 6-1/2 and 16-1/2 years. *Archives of Clinical Neuropsychology, 9*, 553–570.

Sandoval, J. (1982). The WISC-R factorial validity for minority groups and Spearman's hypothesis. *Journal of School Psychology, 20*, 198–204.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). Orlando, FL: Academic Press.

Taylor, R. L., & Ziegler, E. W. (1987). Comparison of the first principal factor on the WISC-R across ethnic groups. *Educational and Psychological Measurement, 47*, 691–694.

Valencia, R. R., Rankin, R. J., & Oakland, T. (1997). WISC-R factor structures among White, Mexican American, and African American children: A research note. *Psychology in the Schools, 34*, 11–16.

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., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1994). Methodological issues in cross-cultural studies on parental rearing behavior and psychopathology. In C. Perris, W. A. Arrindell, & M. Eisemann (Eds.), *Parental rearing and psychopathology* (pp. 173–197). Chicester: Wiley.

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

Van Hemert, D. D. A., Van de Vijver, F. J. R., Poortinga, Y. H., & Georgas, J. (2002). Structure and Score Levels of the Eysenck Personality Questionnaire across individuals and countries. *Personality and Individual Differences.*

Vernon, P. E. (1969). *Intelligence and cultural environment.* London: Methuen.

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

Welkenhuysen-Gybels, J., & Van de Vijver, F. J. R. (2002). *Methods for the evaluation of construct equivalence in studies involving many groups.* In review.