Cultural issues in clinical use of the WISC-IV

Jacques Grégoire
Université Catholique de Louvain - UCLouvain
99 PUBLICATIONS **1,476** CITATIONS

Don Saklofske
The University of Western Ontario
324 PUBLICATIONS **5,419** CITATIONS

Fons Van de Vijver
Tilburg University
681 PUBLICATIONS **13,713** CITATIONS

13

Cultural Issues in Clinical Use of the WISC-IV c00013

Jacques Grégoire[1], James Georgas[2],
Donald H. Saklofske[3], Fons van de Vijver[4],
Claudine Wierzbicki[5], Lawrence G. Weiss[6]
and Jianjun Zhu AUQ1

[1]*Laboratory of Educational and Developmental Psychology, Belgium* AUQ2
[2]*University of Athens, Athens, Greece*
[3]*Division of Applied Psychology, University of Calgary, Calgary, Canada*
[4]*Faculteit Sociale Wetenschappen, Psychologie en Maatschappij, Tilburg, The Netherlands*
[5]*ECPA, France*
[6]*Dorsetshire, San Antonio, USA*

Cultural and linguistic diversity is an important consideration in test development and especially in the administration and interpretation of intelligence tests. These factors are among the very reasons that test scores cannot be viewed in isolation from the "world" of the child, and by themselves, can neither diagnose nor prescribe interventions. Culturally sensitive assessment requires an understanding of how children from different cultures typically approach and respond to standardized testing, a knowledge of the child's cultural background and how this can impact test scores and interpretation, and in many cases an examiner who is familiar with the child's language, dialect and culture. p0200

It is well known that there are differences in mean IQ scores between socioeconomic and ethnic groups. Less well known is "why" these differences exist, and how much of the difference is explained by environment. In discussing differences in IQ test scores among socioeconomic groups more than three decades ago, Wechsler stated, "The cause is elsewhere and the remedy is not in p0210 denigrating or banishing the IQ but in attacking and removing the social causes AUQ3 that impair it" (Wechsler, 1971). Despite this admonishment, early attempts to understand these socioeconomic and ethnic differences in IQ test scores focused mainly on identifying biased test items. Later research to examine differential prediction of achievement from IQ found none (Weiss et al., 1993; Weiss & Prifitera, 1995). Other efforts focused on describing culturally different approaches to the demands of the testing environment and the interaction between examiners and examinees. More recent studies have been directed at understanding those environmental and cultural factors that enhance or diminish cognitive potential during the child's developmental years.

p0220 The assessment of intelligence and cognitive processes of individuals from different ethnic, cultural and linguistic and even socio-economic backgrounds has been an area of controversy for psychologists following the introduction and widespread use of intelligence tests (see Tulsky, et al. 2003). With the ever increasing immigration of ethnic groups to multicultural countries, the search for psychometrically sound strategies for accurately and meaningfully assessing and evaluating the cognitive skills of children from different backgrounds is imperative. However, the assessment of cognitive processes of ethnic children also requires knowledge about the relationships between these background factors and cognitive processes.

p0230 This chapter will present issues related to the culturally sensitive assessment of culturally diverse children with particular focus on the *Wechsler Intelligence Scale for Children* – Fourth Edition (WISC-IV; Wechsler, 2003). The first section focuses on the various kinds of bias that must be considered by both test developers and users, followed by an overview of the various adaptations of the most recent Wechsler scales for children. A review is presented of the findings from a cross-cultural study of the WISC-III as well as a summary of findings of WISC-IV comparisons of the three largest ethnic groups in the United States. The chapter ends with suggestions that can guide the psychologist when assessing culturally diverse children using the WISC-IV.

s0010 CULTURAL BIAS IN INTELLIGENCE TESTING

p0240 Bias refers to the presence of nuisance factors that challenge the comparability of scores from intelligence and other tests and measures across cultural groups (van de Vijver & Leung, 1997). If scores are biased, their psychological meaning is culture/group dependent and group differences in assessment outcome are to be accounted for, at least to some extent, by auxiliary psychological constructs or measurement artifacts. Bias arises in the application of an instrument in at least two cultural groups and the ensuing comparison of scores, patterns or item values. The need for cross-cultural validation and verification should not be interpreted as blind empiricism. Nor should we simply concede that cultural differences make the comparison of such latent traits as intelligence impossible; rather we should use our psychometric skills and psychological knowledge to make every effort to minimize bias and maximize equivalence. On the contrary, not all instruments are equally susceptible to bias. For example, structured test administrations are less prone to bias influences. Analogously, comparisons of closely related groups will be less susceptible to bias than comparisons of groups with a widely different cultural background.

In order to detect and/or prevent bias, we need to recognize factors that can induce bias. Table 13.1 provides an overview of sources of bias, based on a classification by van de Vijver and Tanzer (2004) and van de Vijver and Poortinga (1997). Sources of bias are numerous, thus the overview is necessarily limited. p0250

TABLE 13.1 Sources of Bias in Cross-Cultural Assessment (van de Vijver, 2003)

t0010

Type of bias	Source of bias
Construct bias	u0010 • Incomplete overlap (or complete non-overlap) in the definition of intelligence across cultures
	u0020 • Differential appropriateness of items associated with the construct (e.g., skills do not belong to the repertoire of one of the cultural groups)
	u0030 • Poor sampling of all relevant behaviors (e.g., short instruments)
	u0040 • Incomplete coverage of all relevant aspects/facets of the construct (e.g., not all relevant domains are sampled)
Method bias	u0050 • Sample bias
	u0060 • Incomparability of samples (e.g., caused by differences in educational background or motivation)
	u0070 • Administration bias
	u0080 • Differences in environmental administration conditions, such as ambient noise in the classroom
	u0090 • Ambiguous instructions for pupils and/or guidelines for administrators
	u00100 • Differential expertise of test administrators
	u00110 • Differential usage of norms/instructions Tester/interviewer/observer effects (e.g., halo effects)
	u00120 • Communication problems between respondent and pupil
	u00130 • Instrument bias
	u00140 • Differential familiarity with stimulus material
	u00150 • Differential familiarity with response procedures
	u00160 • Differential response styles (e.g., social desirability, extremity scoring, acquiescence)
Item bias (Differential item functioning)	u00170 • Poor translation and/or ambiguous items
	u00180 • Nuisance factors (e.g., item may invoke additional traits or abilities)
	u00190 • Cultural specifics (e.g., incidental differences in connotative meaning and/or appropriateness of the item content)

s0020 CONSTRUCT BIAS

p0260 The first kind of bias, construct bias, is found when the construct measured is not identical across groups. Construct bias precludes the cross-cultural measurement of a construct with the same/identical measure. Construct bias can be a consequence of differential appropriateness of the behaviors associated with the construct in the different cultures. An example comes from studies on cross-cultural differences in everyday definitions of intelligence. Western intelligence tests tend to focus on reasoning and logical thinking (such as the Raven's Progressive Matrices) while tests of acquired knowledge have typically been added in large batteries (such as Vocabulary scales of the Wechsler scales). When Western individuals are asked which characteristics they associate with an intelligent person, skilled reasoning and knowing much are frequently mentioned. In addition, social aspects of intelligence are mentioned. The latter aspects are even more prominent in everyday conceptions of intelligence in non-Western groups. For, example, Kokwet mothers (Kenya) say that an intelligent child knows its place in the family and behaviors associated with it, like proper ways of addressing other people. Studies in various non-Western countries (Azuma & Kashiwagi, 1987; Serpell, 1993; Grigorenko et al., 2001) also show that descriptions of an intelligent person go beyond the school-oriented domain and involve social aspects and even obedience. Yan and Saklofske (2004) described the five basic human attributes found in ancient China: humility, loyalty, courtesy, intelligence and trustworthiness. Most Western psychologists would view these qualities is a mixture of intelligence, personality and conative characteristics. Until recently, the domain covered by most intelligence tests was usually restricted to scholastic intelligence but that has very much changed as reflected in the newest Wechsler scales. More recently Ackerman (2007) has called for a redefinintion of adult intelligence and proposed a four component model that includes intelligence-as-process, personality, interests and motivation, and intelligence-as-knowledge.

s0030 METHOD BIAS

p0270 The second kind of bias, called method bias, can result from such factors as sample incomparability, instrument differences, tester effects and administration mode. Method bias is used here as a label for all sources of bias emanating from factors often described in the methods section of empirical papers or study documentations. They range from differential stimulus familiarity in mental testing to differential social desirability in personality and survey research. Identification of method bias requires detailed and explicit documentation of all the procedural steps in a study. As an example of method bias, Deregowski and Serpell (1971) asked Scottish and Zambian children in one condition to sort miniature models of animals and motor vehicles and in another condition to sort photographs of these models. Although no cross-cultural differences were found for the actual models, the Scottish children obtained higher scores than the Zambian children when photographs were sorted.

p0280 Among the various types of method bias, sample bias is more likely to jeopardize cross-cultural comparisons when the cultures examined differ in more respects. Such a larger cultural distance will often increase the number of alternative explanations for cross-cultural differences to be considered. Recurrent rival explanations are cross-cultural differences in social desirability and stimulus familiarity (test-wiseness). The main problem with test-wiseness is their relationship with country affluence; more affluent countries can be expected to be more acquainted with psychological testing. Subject recruitment procedures are another source of sample bias in cognitive tests. For instance, the motivation to display one's attitudes or abilities may depend on the amount of previous exposure to psychological tests, the freedom to participate or not, and other sources that may show cross-cultural variation.

p0290 Administration method bias can be caused by differences in the procedures or mode used to administer an instrument. For example, when interviews are held in respondents homes, physical conditions (e.g., ambient noise, presence of others) are difficult to control. Respondents are more prepared to answer sensitive questions in self-completion contexts than in the shared discourse of an interview. Examples of social environmental conditions are individual (versus group) administration, the physical space between respondents (in group testing), or class size (in educational settings). Other sources of administration that can lead to method bias are ambiguity in the questionnaire instructions and/or guidelines or a differential application of these instructions (e.g., which answers to open questions are considered to be ambiguous and require follow-up questions). The effect of test administrator on measurement outcomes has been empirically studied; regrettably, various studies apply inadequate designs and do not cross the cultures of testers and pupils. The presence of the tester is usually not very obtrusive if the test administration takes place under standardized conditions (Jensen 1980). A final source of administration bias is constituted by communication problems between the pupil and the tester. The almost unavoidable obtrusiveness of interpreters is another example. Communication problems are not restricted to working with translators. Language problems may be a potent source of bias when an interview or test is administered in the second or third language of interviewers or respondents.

p0300 Instrument bias is a common source of bias in cognitive tests. A Raven-like figural inductive reasoning test was administered to high-school students in Austria, Nigeria and Togo (educated in Arabic) (Broer, 1996). The most striking findings were cross-cultural differences in item difficulties related to identifying and applying rules in a horizontal direction (i.e., left to right), which was interpreted by the authors as bias in terms of the different directions in writing Latin as opposed to Arabic.

p0310 The presence of method bias can be easily overlooked. When a single test is administered at a single occasion, it is not always easy to estimate the influence of method bias. Evidence on the presence of method bias can also be collected from applications of test–retest, training and intervention studies.

Nkaya et al. (1994) administered Raven's Standard Matrices three times to sixth-graders in France and Congo. Under untimed conditions score improvements were similar for both groups, but under timed conditions the Congolese pupils progressed more from the second to the third session than did the French pupils. Ombredane et al. (1956) have shown that in some groups repeated test administrations can also affect the relationship with external measures. The predictive validity of the Raven test increased after repeated administration in a group of illiterate, Congolese mine workers. It is likely that the results of both studies are due to learning processes that took place during the testing, such as a better task comprehension and more acquaintance with the test and the testing procedure. In this line of reasoning, the validity of the first test administration is challenged by sources of method bias.

s0040 ITEM BIAS (DIF)

p0320 The third type of bias distinguished here refers to anomalies at the item level and is called item bias or differential item functioning (DIF). According to a definition that is widely used in education and psychology, an item is biased if respondents with the same standing on the underlying construct (e.g., they are equally intelligent), but who come from different cultures, do not have the same mean score on the item. The score on the construct is usually derived from the total test score. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Holland & Wainer, 1993; Camilli & Shepard, 1994; van de Vijver & Leung, 1997).

p0330 Although item bias can arise in various ways, poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, and the influence of cultural specifics such as nuisance factors or connotations associated with the item wording are the most common sources. For instance, if a geography test administered to pupils in Poland and Japan contains the item "*What is the capital of Poland?*", Polish pupils can be expected to show higher scores on the item than Japanese students, even if pupils with the same total test score were compared. The item is biased because it favors one cultural group across all test score levels.

p0340 Even translations which seem to be correct can produce problems. A good example is the test item "*Where is a bird with webbed feet most likely to live?*", which was part of a large international study of educational achievement (*cf.* Hambleton 1994). Compared to the overall pattern, the item turned out to be unexpectedly easy in Sweden. An inspection of the translation revealed why: the Swedish translation of the English was "bird with swimming feet" which gives a strong clue to the solution not present in the English original.

p0350 This brief discussion on bias reminds us that bias can affect all stages of the testing and assessment enterprise from the theories and models that in turn guide the development of the test itself, to the standardization features or the test (norms, administration and scoring) and finally to the actual clinical use and interpretation. Thus minimizing bias is not an exclusive concern of only test developers, but of all who use tests in both research and clinical practice. Since bias can challenge all stages of a project, ensuring quality is a matter of combining good theory, questionnaire design, administration and clinical interpretation.

WHAT DID WE LEARN FROM THE WISC ADAPTATIONS ACROSS CULTURES? s0050

p0360 Following from the recognition that psychological tests may travel a "smooth or bumpy road" when they move from their place of origin to another country or culture, we now turn to a more focused discussion of how the Wechsler tests have addressed bias issues.

THE WISC ADAPTATIONS s0060

p0370 The term "test adaptation" should be preferred to "test translation," the first one being broader and more reflective of what happens when a test is developed in a culture based on a test previously developed for use in another culture. "Test translation is only one of the steps in the process of test adaptation and even at this step, adaptation is often a more suitable term than translation to describe the actual process that takes place" (Hambleton, 2005, p.4). Often, the translation of an item or an instruction is not straightforward. The translator has to find words, concepts or expressions that are equivalent in the source and the target languages. Finding such equivalences goes far beyond a literal translation that could be misleading. For example, a literal translation of the French expression "*J'ai le cafard*" would be in English "*I have the cockroach*", while the best equivalence is "*I have the blues*". Similarly, a literal translation of the English expression "*I've got butterflies in the stomach*" would be meaningless in French, the correct equivalence being "*J'ai le trac*" *(I get nervous).*

p0380 Test adaptation first includes the appraisal that the construct could be measured in a different culture. This is followed by the complex process of selecting the translators, decisions related to accommodating the directions, formats, contents and scoring rules to another culture and the assessment of the equivalence between the original or source measure and the adapted test. So, the whole process of test adaptation is most complex and time consuming. The psychometric qualities of the adapted test are closely related to the quality of this procedure. The International Test Commission has published a set of validated guidelines for test adaptation across languages and cultures (Hambleton, 1994, 2005). The 22 guidelines address all the facets of the test development procedure. They are currently used as a reference for test adaptation throughout the world.

p0390 The Wechsler scales are among the most adapted tests in the world. Despite the large number of these adaptations since the publication of the Wechsler–Bellevue in 1939, few comparative studies across languages and cultures have been conducted. Most of these studies were only comparisons between the factor structure of a single adapted version and the original test structure of the US version. The first broad study comparing simultaneously several adaptations of the Wechsler scales was published by Georgas et al. (2003). This study was based on the data of 12 adaptations of the WISC-III: (1) The United States, (2) Canada, (3) United Kingdom, (4) Austria, Germany and Switzerland (German adaptation), (5) France and French-Speaking Belgium, (6) The Netherlands and Flemish-Speaking Belgium, (7) Greece, (8) Sweden, (9) Slovenia, (10) Lithuania, (11) Japan, (12) South Korea and Taiwan.

p0400 Published in 2003 in the United States, the WISC-IV is currently adapted and standardized in Canada (both English and French versions), United Kingdom, Australia, Germany and France. Swedish and Chinese adaptations are pending. We can expect that a similar broad cross-cultural and cross-linguistic comparison, similar to the Georgas et al. (2003) study, will be possible in the near future. These comparisons are very useful because they help in identifying where are the most important cultural differences in the test and how large is their impact on the scores. In one of the following sections, the main observations done in Georgas et al. (2003) study will be discussed. But we will first discuss adaptations of the Wechsler verbal subtests as these are the most frequently modified of the WISC subtests across languages and cultures. In contrast, the performance subtests remain unmodified in the majority of WISC adaptations. Support for the robustness of the original performance subtests will be presented and discussed within the framework of cultural and country similarities in non-verbal tasks. Following this will be a summary of the empirical comparisons between the scores of the standardization samples of 12 adaptations of the WISC-III as presented by Georgas et al. (2003). The similarity of the test structure across the adaptations will be emphasized and the subtest scores differences between countries will be discussed. Finally, comparisons from the WISC-IV will be done, based on the limited empirical data available, mainly from the French adaptation.

s0070 ADAPTATION OF THE VERBAL SUBTESTS

p0410 The necessity to translate and, often, to adapt verbal subtests to each language/culture seems obvious. In the WISC-IV, these subtests are included in the verbal comprehension scale and the working memory scale. In the WISC-III, most of the verbal subtests (included in the verbal scale) required at least some modifications when adapted in another country (Georgas et al., 2003). The most frequent modifications were found in the Vocabulary subtest. In Slovenia, Korea and Lithuania only 23% of the vocabulary items were modified. But in Japan, 93% of these items were modified. In non-English-speaking countries, information was the second most modified WISC-III subtests (from 10% of the items in Taiwan to 57% of the items in Japan), followed by comprehension (0–56% of the items), similarities (0–37% of the items) and arithmetic (0–42%). Digit span stayed unchanged across adaptations. It should be noted that, even in another English-speaking country, the United Kingdom, some items were modified in the comprehension and the information subtests. This example shows us that test adaptation is not only a linguistic issue, but a broader cultural one.

Some items cannot be retained in the adapted version because, when translated, the answer is included in the question itself. One of the most typical example is an information item: "*How many things make a dozen?*". Literally translated, the English word "*dozen*" is "*douzaine*" in French, and the correct answer "*douze*" is essentially given in the question. Consequently, the original item, even correctly translated, could not be used in the French adaptation and had to be replaced by an equivalent one (i.e., an item having the same difficulty level). p0420

Some items have to be modified because of scoring issues. For example, the information item "*Name two kinds of coins*" cannot be scored in the same way in Germany or Korea compared to the United States because these countries have different categories of coins. Another example is the comprehension item "*Why doctors take additional classes after practicing medicine for a while?*". Such an item requires an adaptation of the scoring rules in all the countries where "additional classes" are not required by the state board for keeping the medical license. p0430

The most difficult issue when adapting verbal items is to find questions having a similar difficulty level in the source and the target culture. An item correctly translated can have a very different difficulty level in the target language than in the source language. For example, the question "*What is the Koran?*" will be easy in a dominant Muslim country, as Turkey, but more difficult in a mainly Christian country, as the United States. Similarly, a word such as a noun can be more common in one language than in another one, and will consequently be easier in the first one than in the second one. For example, "*What is a cranberry?*" will be easier for an American than for a Belgian child because this kind of berry is scarce in Belgium, but very common in the United States. In this case, a correct translation is not the solution. An equivalent word, having the same difficulty level, has to be identified. Finding such an equivalent item is not easy, and a selection only based on a subjective judgment could be problematic (e.g., should we select "raspberry" or "blueberry" as equivalent to "cranberry"?). An empirical assessment of the difficulty level of the word is often required before taking a decision. For this reason, test translators tend to create more verbal items than needed, and to select the best ones after the empirical assessment of their difficulty level. Such an empirical procedure is very helpful to develop a scale having equivalent graduations (i.e., items ranked according to their difficulty levels) in the source and the target languages. p0440

The adaptation of the arithmetic items raises specific problems. When questions refer to dollars or US measures, as miles or F°, a literal translation is often misleading. For example, a literal translation in Japanese of the following item would be inappropriate: "*If 5 bottles of water cost 6 dollars, what is the price* p0450 *of 25 bottles of water?*". The price of a bottle of water could be US 1.20$, but certainly not 1.20 Japanese Yen (US $1.00 ≈ 120 Yen). If the price of the bottle of water is adapted to the Japanese currency (i.e., 148 Yen), the adapted item will be much more difficult than the original one. The best solution would be keeping the numbers and the calculation from the original item (i.e., (25/5) × 6), but adapting the verbal problem to these numbers and calculation. This will likely result in the greater probability of a match between the difficulty levels of the adapted item and the original one. Because of problems encountered in adapting some arithmetic items, the best suggestion would be to think of the potential adaptations when developing the original items, and avoiding the use of US currencies and measures in the arithmetic problems.

p0460 Among the verbal subtests, digit span and letter–number sequencing are the only ones that were not modified in the adaptations of the WISC-III and the WISC-IV. These subtests are always literally translated. Since they are only composed of digits and/or letters, they look identical across languages and cultures. However, as we will see, this apparent similarity doesn't prevent cultural influences on the performances on these subtests.

s0080

ADAPTATION OF THE NON-VERBAL SUBTESTS

p0470 The non-verbal subtests (i.e., the subtests included in the perceptual reasoning and the processing speed scales, have the reputation to be less culturally loaded than the verbal ones. Most often, they are not modified or adapted in other countries and only the instructions are translated. The cross-cultural analysis of the WISC-III across 12 cultures/language showed that Japan was the only country where some non-verbal subtests (i.e., performance subtests) were modified. From this observation, we might conclude that these subtests are culture fair and represent universal measures of intelligence. Such a conclusion would be naïve. Understanding non-verbal stimuli and reasoning with these stimuli are also culturally influenced cognitive procedures. Non-verbal items do not measure "genuine" intelligence, independently of any cultural and educational influences. Cultural experiences are always the framework through which we perceive, analyze and process all the non-verbal stimuli. Even the capacity to analyze geometrical figures (orientation, number of components and angles…) presented in the Raven's matrices or in the matrices subtest of the WISC-IV is developed during the school education. Consequently, children with limited education will be less efficient in these tasks than children having a regular school experience.

p0480 With the exception of the block design subtest, all the non-verbal subtests of the WISC-IV use pictures (picture concepts, matrix reasoning and picture completion) or symbols (coding and symbol search). Like words, pictures are a relationship between a signifier (the perceived picture) and a signified (what this picture represent, i.e., its meaning). Some pictures represent rather a universal signified, what they represent being easily identified by most of the people across the world (e.g., a square or a car). But the signified can also be specific to some cultures and not be recognized in other cultures, even using the best pictures (e.g., some exotic fruits could not be recognized by Europeans, or some electronic devices could not be recognized in developing countries). Even when the signified is universally known, some pictures can be rather weak representations. For example, a "pick-up" would not be the best picture to represent the concept "car" or a figure would not be the most appropriate signifier of the concept "fruit."

A majority of the pictures presented in the WISC-IV do not require adaptation in most of the cultures where this test has been adapted, because these pictures are typical and unambiguous representations of realities well known in the target cultures. However, the familiarity with these realities and their representations should always be appraised rather than simply taken for granted. For example, pictures related to baseball (glove, ball, bat…) are very familiar to American children: they know the real objects and the pictures are very typical representations of these objects. Many French or Spanish children could recognize these pictures, because they may have seen these objects in American movies or television shows. But most of them have never seen the real object or played with them as baseball is not a popular or well-known sport in either France or Spain. In these countries, we could also expect that the familiarity with these pictures varies according to social class. Consequently, to keep unchanged the difficulty level of the items and to avoid potential bias related to social class, these pictures should be changed in the adapted version of the test. Equivalent characteristics to the original ones should be found in the target culture: the baseball glove could become a boxing glove, the baseball ball a tennis ball and the baseball player a soccer player. In both cultures, these characteristics and the associated pictures refer to popular games all the children are familiar with. Even when the problematic pictures are not part of the correct answer, they should be replaced with more familiar ones because children could be distracted with these unusual and even meaningless pictures. p0490

The knowledge of the words corresponding to the pictures should also be assessed in each culture. These word–picture connections are essential for giving the correct answer. For example, Swedish and Australian children could easily recognize the picture of a kangaroo, but naming this picture (i.e., giving the correct word "kangaroo") could be easier for Australians because they are more familiar with this animal. Even when the use of the words corresponding to the pictures is not required to give the correct answer, these words can facilitate the cognitive processing of some problems. Some children verbalize (out loud or not) the pictures of the picture concepts and matrix reasoning subtests because they can more easily process the items in this way. If they do not know the words corresponding to some pictures, such verbal mediation will be more difficult or even impossible. p0500

Sometimes, the reality represented by the picture is familiar in both the source and the target cultures, but its representation varies across cultures. There p0510 are fire hydrants in United States, France and Switzerland, but their appearance is different in each country. In this case, the US picture of a fire hydrant cannot be used in France and Switzerland and should be redrawn. Unfortunately, the fire hydrants are also different in France and Switzerland so the redrawn picture of the French adaptation of the WISC-IV is very familiar to the French children, but meaningless for the Swiss children (several Swiss children identified this picture as a ketchup bottle!). Mailboxes and ambulances raise a similar adaptation problem across countries, even culturally very close countries such as France and Switzerland. Thus when an original picture is retained in the adapted subtest, many details have to be changed to avoid distracting the children with inappropriate information. For example, the English texts appearing on the objects (newspaper, fire extinguisher, license plate…) have to be erased or modified.

s0090 INVARIANT STRUCTURE ACROSS CULTURES

p0520 The cross-cultural study of the WISC-III conducted by Georgas et al. (2003) provides important information about cultural differences on both the subtest scores and on the composite scores. Because the WISC-III and the WISC-IV share 11 subtests, the findings from this study can be very relevant to the clinical use of this latest version of the Wechsler scales. A strength of this study was that the analyses were based on representative samples of 6- to16-year-old children (*n* = 15,999) from 12 countries, in terms of geographical areas, parental education and occupation, gender, ethnic groups and other variables (*n* = 15,999) of relevance to describing the demographics of these countries. There are few examples in the literature on intelligence and cognitive processes with samples so carefully selected and representative of the social structural variables in each country.

p0530 The countries shared similar socioeconomic features. Except for Lithuania and Slovenia, they are among the most affluent countries in the world, with developed economies and educational systems. Their governments and their people place high value on the role of education for occupational success and a better life. All these countries have highly invested in information technology. However, these countries are from the northern hemisphere and are not representative of a wide range of cultures, economic levels and educational systems of the poor countries of Africa, South America or East and West Asia.

p0540 In each country, indigenous researchers adapted the items to avoid cultural bias. Although some degree of item bias due to cultural factors is still possible, children in all these countries were familiar with the tasks: verbal stimuli, pictures, blocks, puzzles, etc. Consequently, the different culturally adapted versions of the WISC-III reported here can be considered as conceptually very similar.

p0550 The main question in this cross-cultural study was the structural equivalence of the WISC-III across 12 countries. That is, are the same cognitive structures measured by the WISC-III found in each of these countries? If this structural equivalence is found, it would provide some support for the universality of the kind of intelligence measured by the WISC-III, and consequently also to a large extent, by the WISC-IV since the two tests are based on a similar model of intelligence and share similar tasks.

In order to assess the similarity of the intelligence construct measured by the WISC-III across the 12 countries, an analysis of the structural equivalence was conducted following van de Vijver and Leung (1997). The "one-to-one" procedure employed to determine structural equivalence consisted of determining the factorial agreement of all pairs of countries. The first step was to conduct an exploratory factor analysis of the standard scores of the 12 subtests, for each of the 12 data sets. Three and four factors were extracted. The second step was to compare the factor structure of each country with the data sets of all the other country. One country was arbitrarily designated as the target and the factor loadings of the second country were rotated so as to maximize their similarity with the target country. The similarity of the factor structures was assessed by comparing the factor loadings of all pairs of countries for each of the four factors with Tucker's phi, a factorial coefficient of agreement. This resulted in a country-by-country matrix of 66 (= (12 × 11)/ 2) Tucker's phi coefficients. A phi value ≥1 is considered as an indication that two factors are identical. All the coefficients were larger than 0.90 for the factors verbal comprehension, perceptual organization and processing speed, indicating factorial stability in the factor equivalence across all the 12 data sets on these factors. Many of the phi coefficients are at the level of 0.99. p0560

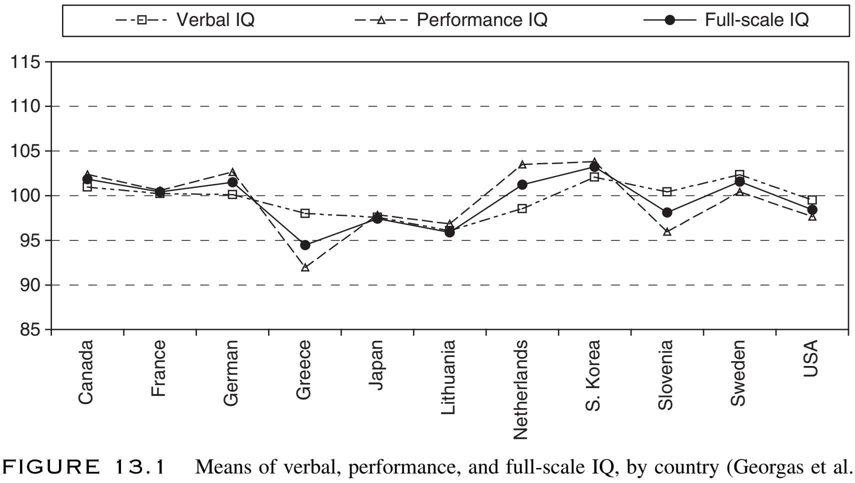
However, this was not the case with the factor freedom from distractibility, some phi coefficients being below 0.90. A close inspection of the factor structure suggested that the arithmetic subtest was the main source of distortion. In most countries, the loadings of arithmetic are split between verbal comprehension and freedom from distractibility. In some countries, the higher loading of arithmetic was on freedom from distractibility, while it was the reverse in other countries. This instability of the Arithmetic factor loadings across countries is one of the reasons why arithmetic is no longer a regular subtest in the WISC-IV, where it is now replaced by letter–number sequencing for the usual calculation of the working memory index. p0570

The main finding of Georgas et al. (2003) study was that there is clearly a structural equivalence across the 12 data sets of a three-factor structure. A four-factor solution was very stable for the first three factors, but less stable for the fourth one. The median value of the coefficient of factorial agreement for the fourth factor was 0.95, which is well above the minimum threshold value of 0.90. The conclusion is that there is a remarkable similarity of the structure across these countries underlying the WISC-III. This finding provided evidence of universal cognitive processes across these cultures and also indicated that the WISC-III can be validly administered with the same interpretations regarding its cognitive structure across these countries. A similar study will need to be conducted with the WISC-IV adaptations. Because of the overlap between the both versions of the WISC, p0580 we can expect that a robust factor structure across cultures will also be observed with the WISC-IV. For example, a recent study comparing the United States and the Canadian WAIS-III (Bowden et al., in press) showed that the measurement model, involving four latent variables reflecting VCI, PRI, WMI and PSI, satisfied the assumption of invariance across samples. The subtest scores also showed similar reliability in both samples, although slightly higher latent variable means were found in the Canadian normative sample.

s0100 CROSS-CULTURAL COMPARISONS OF SUBTEST AND COMPOSITE SCORES

p0590 An important question in Georgas et al. (2003) study concerned the comparative analysis of the WISC-III scores on the subtests, and the IQ scores. Do countries differ in level of intelligence, as measured with the WISC-III? These analyses were conducted with the raw scores from each country and not the scaled scores, nor the actual verbal, performance and full-scale IQs (FSIQs), because comparisons of means of scaled scores across countries would result in zero differences.

p0600 In order to test the potential cross-cultural differences in mean scores on the WISC-III, analyses of variance were conducted with the full-scale, verbal and performance IQ raw scores, and with the subtest raw scores, using country as the independent variable. Although, as described above, these analyses were conducted with the raw scores, in order for the results to be meaningful to the reader, the results are presented in Figure 13.1 as IQ scores (mean = 100 and SD = 15).



f0010 FIGURE 13.1 Means of verbal, performance, and full-scale IQ, by country (Georgas et al., 2003).

p0610 The main conclusion is that cross-cultural differences in composite scores (IQ scores) of the 12 data sets were very small (Figure 13.1). There were no major mean score differences on the WISC-III across the countries; neither the FSIQ, the verbal IQ or the performance IQ showed large cross-cultural score differences.

p0620 Because of the large sample (n = 15,999), the size of differences was not evaluated through the traditional level of statistical significance, but by the η^2 values, which estimate the proportion of variance accounted for by country in the explanation of the dependent variable and not by level of significance. The η^2 values were low, the mean value being 0.033 (range from 0.003 to 0.103). This means that the influence of the countries on the observed differences is low. For each difference between a country mean and the global mean, the effect size was calculated dividing each difference by the average standard deviation. According to Cohen (1988), an effect size of 0.20 is considered as small, 0.50 as medium and 0.80 as high. According to Cohen's criteria, all the effect sizes were small.

p0630 Table 13.2 shows the proportion of variance (η^2) accounted for by country in the differences observed on the six verbal scores. This proportion of variance is always low, with the exception of the digit span subtest where larger differences between countries were observed. The larger effect sizes were observed for Lithuania (−0.45) and South Korea (0.69). As the subtest was literally translated in both countries, without any modification, these differences cannot be explained by the non-equivalent difficulty level of modified items (as they could

t0020 TABLE 13.2 Effect Size for the Six Verbal Scores, by Country (Georgas et al., 2003)

	Inf.	Sim.	Ari.	Voc.	Com.	DS
Canada	−0.03	0.41	−0.21	−0.09	0.12	0.04
France	0.13	−0.18	0.12		0.23	−0.27
Germany	−0.10	−0.09	−0.05	0.34	0.16	−0.23
Greece	−0.01	−0.22	−0.16	0.02	0.08	−0.25
Japan	−0.23	−0.04	0.09		−0.59	0.26
Lithuania	−0.25	−0.25	−0.17	0.17	−0.04	−0.45
The Netherlands	0.00	0.18	0.12		−0.27	−0.37
South Korea	−0.03	−0.31	0.34	−0.18	0.01	0.69
Slovenia	0.02	−0.33	−0.04	0.22	0.07	0.16
Sweden	0.44	0.33	−0.04		0.10	−0.33
U.S.A.	−0.01	0.33	−0.23	−0.17	0.01	−0.07
η^2	0.006	0.034	0.017	0.022	0.024	0.091

Note: For Vocabulary, the information missing for several countries because too many items were modified and comparisons were therefore not possible.

for the high score on information in Sweden or the low score in comprehension in Japan).

p0640 The most plausible explanation is related to Baddeley's phonological loop hypothesis. According to Baddeley (1986), we can store more short words than long words in our short-term memory because speech-based information is held in memory through an articulatory control process based on inner speech. As various East Asian languages, such as South Korea, have short words for digits, people speaking these languages can store more digits in their short-term memory. On the other hand, Lithuanians have longer words for digits and, consequently, people speaking Lithuanian can store fewer digits in their short-term memory. This observation is very important for a culturally based interpretation of the digit span subtest. Thus, Lithuanians do not have a weaker short-term memory than the South Koreans; rather the observed difference between Korea and Lithuania can be explained by the specificity of each language. It is not a consequence of a poor test adaptation, and there is no psychometric procedure to erase this difference. Therefore, translating the digit span subtest to assess a child speaking a different language is not a magical solution to eliminate cultural bias.

p0650 Table 13.3 shows the proportion of variance (η^2) accounted for by country in the differences observed on the six performance scores. This proportion of variance is very low, with the exception of the symbol search subtest where larger differences between countries were observed. The larger effect sizes were

t0030 TABLE 13.3 Effect Size for the Six Performance Scores, by Country (Georgas et al., 2003)

	PC	CD	PA	BD	OA	SS
Canada	0.19	0.22	0.07	0.00	0.13	−0.05
France	−0.03	0.22	0.02	0.13	0.10	−0.29
Germany	0.00	0.03	0.07	0.09	0.46	−0.02
Greece	−0.40	0.06	−0.35	−0.10	−0.51	−0.52
Japan	0.41	−0.40	−0.04	−0.04	−0.10	−0.28
Lithuania	−0.22	−0.02	−0.21	0.07	−0.03	−0.29
The Netherlands	0.08	0.18	0.12	0.04	0.24	0.16
South Korea	−0.05		0.03	−0.07	−0.04	0.87
Slovenia	−0.02	−0.37	−0.15	0.09	−0.20	−0.24
Sweden	0.12	−0.15	0.14	0.03	0.10	−0.11
U.S.A.	−0.10	0.07	0.03	−0.10	−0.21	−0.20
η^2	0.014	0.021	0.011	0.009	0.034	0.103

Note: For Korea, the information was missing for the Coding subtest. PC = picture completion, CD = coding, PA = picture arrangements, BD = Block Design, OA = object assembly and SS = symbol search.

observed for Greece (−0.52) and South Korea (0.87). Georgas et al. (2003) suggested that South Koreans' high score on symbol search may have reflected a strong motivation because other studies and international comparisons of educational achievement suggest that South Koreans often show high motivation in education-related matters. However, no suggestion was offered for explaining the lower score of children from Greece.

Another interpretation of the high score for South Korean children could be related to Korean writing. Korean is written using an alphabetic system (called Hangeul) and a system of characters close to the Chinese one (called Hanja). Learning to read and write using both systems could improve the ability to analyze and recognize symbols such as those presented in the symbol search subtest. If the mastery of the Korean writing system improves the performances on the symbol search subtest, it should also improve the performances on the Coding subtest since this subtest also requires children to analyze, memorize and write symbols. Unfortunately, this information was missing for Korea in the Georgas et al. (2003) research. p0660

However, a study conducted by Chen and Zhu (2004) on Taiwanese children using Chinese letters provided a strong support to this hypothesis. These authors postulated that children who routinely read and write Chinese may perform faster on both the coding and symbol search tasks than peers who primarily read and write English because the symbols used in the subtests are more similar to Chinese characters than to English alphabet. To test this hypothesis, a total of 1,003 cases from the normative sample data of the Taiwan WISC-III and the US WISC-III were matched by age, gender and parent educational level. To facilitate this comparison because Taiwan is a more homogeneous society than the United States, minority children of the US normative sample were excluded from the study. Except for the language difference, the items, instructions, administration procedures and scoring rules were the same for the Taiwan and the US editions of the non-verbal subtests. Mean raw total scores of the matched Taiwan and US samples on block design, object assembly, picture completion, picture arrangement, mazes, coding and symbol search subtests were compared directly. The effect sizes of the mean score difference were calculated using Cohen's *d*. p0670

The results revealed that on the picture completion, picture arrangement and object assembly subtests, Taiwanese children had lower scores than their US peers. But the average effect sizes across the 11 age groups were rather small (−0.11, −0.26 and −0.09, for picture completion, picture arrangement and object assembly, respectively). However, Taiwanese children did significantly better than their US peers on coding, symbol search, block design and mazes subtests. The average effect sizes across the 11 age groups were 1.20, 0.89, 0.67 and 0.55, respectively. The effect sizes were particularly large for coding and symbol search. These results supported the hypothesis that reading and writing Chinese on daily bases facilitates children's performance on processing speed tasks that utilize symbols. However, further research is needed before drawing a strong conclusion on this issue. This research should particularly address the p0680 relatively poor performances of Japan on the coding subtest and of Greece on the symbol search subtest. These differences could be related to cultural specificity in early education, but this interpretation is not currently supported by empirical data.

s0110 WITHIN COUNTRY DIFFERENCES: THE US EXPERIENCE

p0690 Before examining the ways of managing the possible cultural bias in the administration, scoring and interpretation of the WISC-IV, it is important that users of this and all other intelligence tests understand some of the other causes of bias that may creep into our efforts to assess the cognitive abilities of children. In the cross-cultural research described above, IQ test scores in the 12 different countries were also examined as a function of national indicators of affluence and education (Georgas et al., 2003). The factors which influence mean IQ test scores between nations (i.e., gross national product and percent of gross national product spent on education) are essentially group level counterparts of the individual difference variables known to moderate IQ tests scores between ethnic groups within the United States (i.e., parent education and income).

p0700 More recently, data from the standardization studies of the WISC-IV have been used to explore the environmental factors that enhance or negatively impact cognitive potential during the child's developmental years (Prifitera et al., 2005; Weiss et al., 2006). While genetics is a significant determiner of intellectual ability, children are not borne with a genetically predetermined, fixed IQ score, but with a range of intellectual potential. Thus we are not contesting the large contribution of genetics to measured intelligence but rather reinforcing that intellectual potential may be fully or partially actualized depending on qualities of the environment during the critical developmental years. In brief, cognitively and linguistically stimulating environments enhance intellectual growth while cognitively and linguistically impoverished environments negatively impact intellectual potential – and much of this may be related to parental behaviors that occur within the home during the child's developmental years. The United States provides a natural laboratory for this research. It is comprised of African American, Hispanic and Caucasian majorities although these demographics are showing change with the influx of people from other countries. Further, in spite of the "melting pot" philosophy in the United States, there is also evidence that income, education and other critical social–economic factors differ across these major groups. Of course the WISC-IV was developed and carefully standardized in the United States such that data from all three groups could be analyzed in a comparative manner.

p0710 Weiss et al. (2006) reported that ethnicity explained 1.4% of the variability in IQ scores between Hispanic and White children, and that this difference was eliminated after controlling for the educational level and income of the parents. With respect to African American – White IQ score differences, Weiss et al. found that race accounted for 4.7% of the variance in IQ scores. After controlling for parental education and income, race explained only 1.6% of the remaining variance. It is not known why the African American – White difference was not completely eliminated by controlling for parent education and income as it was in the Hispanic – White analyses. However, Weiss et al. suggested that differences in quality of education and historical discrimination in employment practices among the generations that are now parents of children and adolescents may have resulted in these variables (parent education level and income) not having the same meaning and relevance for African Americans as Whites.

p0720 Next, the role of parent's expectations of children's academic success was examined. Parent expectations explained 30.7% of the variance in IQ scores across all children – far greater than that explained by race, ethnicity, parent education or income. Finally, parental expectations as a function of parent education and income was investigated, and after controlling for parent education and income, parental expectations continued to account for 15.9% of the variance in children's IQ test scores – still large. This means that parent expectations are not fully overlapping with socio-economic status. Weiss et al. conjectured that the expectations which parents hold for their children's academic success motivate parenting behaviors that enhance the intellectual development of their children toward the upper end of the range of their potential. The specific parental behaviors and the general characteristics of cognitively enriching home environments need to be better understood by researchers and clinicians alike, but may include an increased amount and variety of linguistic and motoric stimulation in the early developmental years followed by parental monitoring of homework and leisure activities during later childhood and adolescence. One of the most obvious take away points from this analysis is that culturally sensitive assessment practices should include understanding the unique aspects of each child's home environment without making generalizations based on race, ethnicity or socio-economic status.

HOW TO ADDRESS CULTURAL BIAS WHEN ASSESSING COGNITIVE ABILITY? s0120

p0730 Clinicians frequently are requested to assess the intelligence of children with different cultural and linguistic backgrounds. In such situations, their main concern is to avoid bias and conduct a fair assessment. With this aim, they can act in different ways. Some of their actions can be efficient, but others could be inappropriate. The most common actions to overcome possible cultural bias in cognitive ability tests are reviewed in the next sections, and their advantages and limitations are pointed out.

ASSESSING ACCULTURATION s0130

p0740 The first step when assessing a child with a different cultural and linguistic background is to assess his degree of acculturation to the culture in which he is tested and he is living. To what extent can the child be considered as being a member of the population in which the test was developed and standardized?

To answer to this question, several acculturation scales were developed. Acculturation is not a dichotomous phenomenon: being or not being member of a specific culture. It is continuous and often long process of incorporation of a different culture. Berry (1996) suggested appraising the degree of acculturation referring to two dimensions: the immersion in the culture of the native society and the immersion in the culture of the immigration society. Comparing the levels of immersion in each culture corresponds to four acculturation processes: assimilation (moving away from the native culture and immersing fully in the immigration culture), integration (equal immersion in both cultures), separation (complete immersion in the native culture and withdrawal from the immigration culture) and marginalization (lacking of meaningful immersion in both cultures).

p0750 The *Stephenson Multigroup Acculturation Scale* (Stephenson, 2000) is a good example of a scale assessing acculturation. It includes 32 items reflecting acculturation in the domains of language, interaction, media and food. For example, acculturation to the native culture is assessed with items such as: "I like to speak my native language" or "I regularly read magazines of my ethnic groups." Acculturation to the immigration culture is assessed with items as: "I like to eat American food" or "I speak English at home."

p0760 Acculturations scales can be useful if a test is appropriate for assessing a child with a different cultural background. However the results cannot be regarded alone or in isolation. They should take into account other relevant information regarding linguistic and school education. If the child's acculturation position in Berry's categories is assimilation or the integration, the test developed in the dominant culture may be used without major problem. On the other hand, if the child's acculturation position is segregation or marginalization, the regular test for the dominant culture cannot be used and another instrument or procedure should be identified.

s0140 DEVELOPING SPECIFIC NORMS

p0770 Most of the international languages include several dialects and variants across the countries where these languages are spoken. For example, different variations of French are spoken in France, Belgium, Switzerland and Canada. Differences between the standard language (usually spoken by the main group of the population or in the larger country where this language is spoken) and its local variants can be small, but sometimes important. Typically, when developing or adapting a test, only the standard language is taken into account. The test is then considered as the reference for all the variants of this language. But not taking variants and dialects into account when developing or adapting a test can lead to assessment bias and unfair decisions based on the scores in the adapted instrument.

p0780 One solution to this problem is to develop a local adaptation and norms. For example, the US WISC-IV is not the only English version. Norms were developed for the same test in Canada, United Kingdom and Australia. However, the content of the test was not modified, with minor exceptions. Why develop local norms if the test is similar to the original one? Why not use the US norms? In fact this is an empirical question that can best be addressed when sufficient data are available. While the United States and Canada share a common border and English is the majority language spoken, the question of how well the Wechsler scales travel from the United States, where the tests were developed, standardized and normed, to Canada has been addressed. It was first shown that with only a very few word changes, the items from these tests work very well in both countries (DIF analysis, subtest reliabilities, etc.) but in the case of the WISC-III and WAIS-III, there were some very noticeable differences in the raw scores and score distributions on particular subtests and index scores. For example, the PRI difference between the United States and Canada was about five-scaled score points thereby necessitating the development of Canadian norms. More recently with the addition of more "fluid" intelligence to the WISC-IV, it was observed that not only did the items fit very well but that the FSIQ difference between countries was only about two points. In the US, the Spanish version of the WISC-IV not only reflects item translation but also the generation of norms for specific use with Spanish-speaking children. This will be further discussed later (see Harris & Antolin, this book).

Developing local norms is not always possible because it is too expensive in small populations. For example, there is only one French version of the WISC-IV for France and French-speaking Belgium, and the standardization sample includes only children from France. While the same language is spoken in both countries, there are some dialectal differences (vocabulary, syntax, expressions…) that could create bias in the verbal items and, consequently, lead to unfair assessment for the Belgian children. However, before the standardization of the French adaptation of the WISC-IV, a try-out of all the verbal items was conducted simultaneously in France and Belgium with an analysis of DIF. Biased items, showing a differential functioning between the French and Belgium groups, were flagged and deleted. Although the norms for the French WISC-IV did not includ any Belgian children, the DIF analysis conducted during the item try-out helped to make possible a more fair measure for assessing these children. Unfortunately, DIF analysis conducted between linguistic subgroups of the population for which a test is being developed or adapted is uncommon. This methodology is well known in educational measurement (Holland & Wainer, 1993; Camilli & Shepard, 1994), but should be used more often for developing and adapting psychological tests across linguistic groups. Use of DIF would improve the fairness of the common linguistic version of a test. p0790

SCORE ADJUSTMENT s0150

Based on the idea that we can only compare what is comparable, some authors proposed to adjust the scaled scores according to the child's socio-cultural characteristics in order to "correct" their test performance bias. p0800 p0810 In the United States, such a procedure was proposed by Mercer (1979) with her "*System of Multiculticultural and Pluralistic Assessment*" (SOMPA). Mercer considered that each child should be compared to individuals living in similar social and cultural conditions. Consequently, practitioners should always appraise the socio-cultural characteristics of the children they are testing including such variables as the size and structure of their family, the socio-economic status of their parents and their urban acculturation. This information should be quantified according to their ethnic group (White, African American and Hispanic). The child's "socio-cultural" score is used to adjust his/her IQ calculated with the regular norms, producing an estimated potential learning IQ (EPL IQ). The EPL IQ is supposed to be a more precise measure of the child's true learning potential. According to the multiple norms proposed by Mercer, the more a child is marginal regarding the dominant culture, the higher will be her/his EPL IQ compared to her/his regular IQ. The adjusted IQs are often higher than the traditional IQ score, resulting in a reduction in the number of children identified as disabled.

p0820 Mercer's score adjustment procedure was criticized from an empirical and an epistemological viewpoint. Johnson and Danley (1981) conducted research on the predictive validity of the EPL IQ. They compared two groups of children with similar IQs. The children of one group came from disadvantaged sections of the population, according to the SOMPA socio-cultural scales. Consequently, the EPL IQ of these children was higher than their regular IQ, while the EPL IQ of the children from the other group was equal to their regular IQ. The children of both groups received two learning tasks, selected as being less influenced by the dominant culture. Johnson and Danley observed that the EPL IQ and the regular IQ were both weak predictors of the learning performance. The correlations between both IQs and the learning tasks were moderate and very similar. Johnson and Danley concluded that EPL IQ is not a better predictor of learning outcomes than the regular IQ. Consequently, it should not be considered as a better estimate of the child's learning potential than the traditional IQ.

p0830 The other critique addressed to Mercer was epistemological and related to the status of knowledge underlying the SOMPA. Mercer considered the specific knowledge of all cultural groups as equal, defending therefore a relativistic view on knowledge. However, several authors emphasized that different educational environments do not develop equally efficient cognitive aptitudes. Some environments develop stronger aptitudes useful for adaptation in a broad society, while other environments develop only limited aptitudes useful in a very narrow society. Not recognizing these differences may lead to rejection of needed educational supports resulting in an increase in the cognitive differences between individuals living in the same society. Consequently, SOMPA seems to be a wrong answer to a good question. As Jirsa (1983, p.19) emphasized: "The statistical manipulation of current performance (WISC-R IQ) may succeed in eliminating certain children from special education programming, but that in no sense changes the child in terms of his or her current functioning." The goal of intellectual assessment is not to reflect the desired picture of oneself, but to collect useful information to help psychologists make the most appropriate and beneficial choices to help solve real-life problems. The messenger who carries bad news should not to be killed. On the contrary, their news should be taken into account to defining and addressing the issues at hand.

ADAPTING ADMINISTRATION RULES AND ITEMS s0160

When the child's degree of acculturation is insufficient, some clinicians use to accommodate the testing procedures, believing they could eliminate some bias in this way. However, these modifications of the standard instructions and procedures entail negative consequences. The norms of a test were collected according to standard conditions: item format, material, instructions and scoring rules being similar for all the individuals. To compare a child's scaled score to the norms of the standardization sample, the testing should be conducted according to the standard application rules. If some conditions were modified during the testing (e.g., the wording of several questions or the demonstration of items), such a comparison may be no longer valid. In this case, using the standard norms could lead to an overestimation of the child's intelligence because the modification of the testing rules could have been too helpful for this child. p0840

The best solution would be standardizing tests with item formats and instructions reducing the problems associated with cultural differences. For example, most of the WISC-IV subtests use demonstration items, limiting potential bias due to verbal instructions. Some subtests, such as picture concept and matrix reasoning, allow pointing at the correct answer instead of giving a verbal answer that could be more difficult to produce for a child with limited vocabulary. However, this last procedure has some limitations. For example, how could we test verbal memory without using a specific language, or even a selection of syllables and phonemes that are always specific to each language? p0850

Translating instructions and items to the native language of the child could appear to be a good solution to avoid cultural bias. However, we have seen in the section devoted to the adaptations of the WISC that a literally translated item can be easier than the original one because translated words could give some clues to the correct answer. It can also be more difficult when the translated words are less common in the target language than in the source language. A translated item, even literally, is not *ipso facto* equivalent to the original item. We have seen, for example, that the length of the words representing the digits explained the huge difference between South Koreans' and Lithuanians' performances. Consequently, when a clinician has to assess a child who is not fluent with the language used in the WISC-IV, a literal translation of the instructions and the items in the native language of the child should not be seen as the best solution. There is always a risk of bias when using a modified testing procedure. Translated instructions should be used very cautiously and, when they p0860 are necessary, the practitioner should always document the nature of the adapted procedures used in the clinical report.

s0170 SELECTING SUBTESTS

p0870 When a child's acculturation and language knowledge are too weak to allow testing with the version of WISC-IV typically used in the country where he or she is being tested (i.e., the local version), the best option would be to use an adaptation of the test in the child's native language. However, this solution is not always an option when there is no test adaptation in the child's native language, or when the clinician is unable to use the existing adaptation because of a too limited knowledge of the child's native language. In this case, clinicians often chose to present the child with only the non-verbal subtests of the WISC-IV, usually the subtests of perceptual reasoning scale, and occasionally the subtests of processing speed scale. The perceptual reasoning index (PRI) is then used as an estimate of the FSIQ. For several reasons, such a procedure should be used cautiously, and only when no other option is available.

p0880 First, the non-verbal subtests are not culture free. Removing the verbal component from testing doesn't mean there is no longer any cultural influence on test performance. As discussed above, the pictures and the geometrical figures presented in these subtests can be more or less familiar to children according to their cultural background. Moreover, modifications of the instructions are often needed for explaining the tasks to children with limited knowledge of the local language. These modifications could have an influence on the children's performances. The task is to assess cognitive ability and not some artifact that has been confounded because of language or cultural differences.

p0890 Another reason to be cautious is the imperfect correlation between the PRI and the FSIQ. The US manual reports a correlation of 0.82 (Wechsler, 2003), which is rather high, but only allows a rough estimation of the FSIQ on the basis of the PRI score. Moreover, index scores showed relatively important scatter in several WISC-IV standardization samples (Longman, 2005; Grégoire & Wierzbicki, 2007). For example, in the French standardization sample, 50% of the children showed a significant difference of 7 or more points between their PRI score and their average index score, and 25% showed a significant difference of 11 or more points. That means the PRI is not always representative of the child's general ability. Sometimes, it can underestimate this general ability, and sometimes, it can overestimate it.

p0900 The final reason to be cautious using the PRI to estimate the child's general ability is that the correlation between the PRI and measures of school achievement are lower than the correlations observed with the FSIQ, or with the verbal comprehension index (VCI). The US manual (Wechsler, 2003) reports the following correlations with the total achievement score on the WAIT-II: FSIQ = .87, VCI = .80 and PRI = .71. The PRI score is clearly a weaker predictor of school achievement than the FSIQ or the VCI. Therefore, it should be used cautiously as an indicator of the child's learning potential.

INTERPRETING THE SCORES s0180

p0910 Even when an immigrant child's acculturation is sufficient to allow the use of the local version of the WISC-IV (i.e., the version adapted and normed in the country where the child is currently living and being tested), cultural bias should be controlled throughout the testing procedure. In a WISC-IV protocol, we commonly observe a very small number of biased items and children having no culturally related difficulties with most of the WISC-IV subtests. For example, some children from Morocco (Muslim country), but living in Belgium for a long time and being acculturate to this society, told us that the important part missing from the woman's face (picture completion subtest) was her veil, while the correct answer was her eyelashes. In this case, the item should be scored according to the rules used for the test standardization, and the children should receive no credit for this item. Some clinicians may consider such a decision as too strict and would advocate giving credit for this answer because it was meaningful. However, practitioners should stick to the scoring rules if they want to use the norms collected according to standard conditions. Any departure from the testing rules could invalidate the observed scores. Instead of modifying instructions or scoring rules, it seems more appropriate to always add an interpretation to the scaled scores in the psychological report. This interpretation should include clinical observations related to the impact of culture on the testing (misunderstanding of instructions, lack of knowledge of words or pictures, unusual answers…) that could moderate some scores or explain some differences between scores.

p0920 Isolated instances of culturally biased items in a WISC-IV protocol will usually have a limited impact on the subtest and the composite scaled scores. For example, one biased item in the Vocabulary subtest would only be 1/35 of the whole subtest, and much less of the whole test. However, when the number of biased items in a subtest is large, it could invalidate the subtest. When the proportion of inappropriate items in a subtest becomes too large, the best option is to invalidate the subtest and use an alternative subtest when possible. When no alternative subtest is available, the calculation of the composite scores should be prorated, i.e. multiplying the sum of the valid scaled scores per a fraction where the denominator is the regular number of scores (e.g., 5) and the numerator is the number of valid scores (e.g., 4).

CONCLUSION s0190

p0930 There is no way to assess intelligence independently of any cultural influence. There is no culture free test and no culture fair test. Bruner (1974, p.364) emphasized that: "Culture free means intelligence free." Even if the roots of intelligence are in our genetic inheritance, our intellectual behaviors are always shaped by culture. We cannot think in a vacuum independently of any content. The contents are cultural knowledge, learned through education and experience.

It is simplistic to think that only verbal knowledge is influenced by culture. Nonverbal knowledge is equally influenced by culture.

p0940 As every intelligence test, the WISC-IV is a cultural product. Its items reflect the society in which it was developed. However, Georgas et al. (2003) study conducted on the standardization sample data of 12 adaptations of the WISC-III across the world showed that its factor structure was universal. Even if a lot of items were modified in these adaptations, their common factor structures allow cross-cultural and cross-linguistic comparisons. Such an international comparison has not yet been conducted with the WISC-IV. Because of the substantial content and structural overlap between the WISC-III factor based index scores and the WISC-IV, the observation of a similar universal factor structure can be expected.

p0950 When a clinician has to assess the intelligence of a child having a different cultural and linguistic background, his or her main concern is to avoid bias and conduct a fair assessment. Before applying the WISC-IV, he or she should assess his/her acculturation to the local culture in which the test was developed. When the degree of acculturation is insufficient, the first option is to apply a version of the WISC-IV adapted to the child's native culture. When this option is not possible, the clinicians should avoid modifying the local version of the WISC-IV to fit the child's characteristics. However, some clinicians do personal adaptation of the testing rules. The most common modifications are: the translation of instructions or items, the non-standard demonstrations of the subtests and modifications of the scoring rules. The main consequence of these modifications of the testing rules is that the comparison of the child's performances to the local norms is no longer valid. A better option is to select the most appropriate subtests to the child's characteristics, usually the subtests of the PRI, and calculate an estimate of the FSIQ on the basis of this selection of subtests applied according to the standard rules. However, such a procedure should be used cautiously because it may under-represent the construct of intelligence as defined in the WISC-IV, and provides only a rough estimate of the FSIQ. Moreover the predictive validity of the estimated FSIQ is usually lower than the predictive validity of the traditional FSIQ.

p0960 When an immigrant child's degree of acculturation is high, the local version of the WISC-IV may be used. Minor biases could nevertheless be observed during the testing. These biases should not lead to any modification of the scoring rules. It is always appropriate, in fact necessary, to mention in the psychological report the possible impact of culture on the child's performance. This information should be used to moderate and guide the interpretation of scores and score differences.

REFERENCES

Ackerman, P. L. (2007). *Adult intellectual development: An integrated cognitive, affective and conative framework.* Paper presented at the 19th Annual convention of the Association for Psychological Science, May, 2007 AUQ4

Azuma, H., & Kashiwagi, K. (1987). Descriptors for an intelligent person: A Japanese study. *Japanese Psychological Research*, 29, 17–26.

Baddeley, A. (1986). *Working Memory*. Oxford: Oxford University Press.

Berry, J. W. (1996). Immigration, acculturation and adaptation. *Applied Psychology*, 46, 5–34.

Bowden, S. C., Lange, R. T., Weiss, L. G., & Saklofske, D. H. (in press). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale-III in the United States and Canada. *Educational and Psychological Measurement*. AUQ5

Broer, M. (1996). Rasch-homogene Leistungstests (3DW, WMT) im Kulturvergleich Chile-Österreich. Erstellung einer spanischen Version einer Testbatterie und deren interkulturelle Validierung in Chile [Rasch-homogeneous performance tests in cross-cultural comparison Chile-Austria. Making a Spanish version of a test battery and its intercultural validation in Chile]. Unpublished thesis, University of Vienna, Austria.

Bruner, J. S. (1974). *Beyond the Information Given*. London: Allen & Unwin.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage Publications.

Chen, H., & Zhu, J. (2004). *Developmental trend on children's digit symbol coding and symbol search abilities: U.S.A. and Taiwan WISC-III norms compared*. Paper presented at the 28th International Congress of Psychology, Beijing.

Cohen, J. (1988). *Statistical Power Analysis for Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

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

Georgas, J., Weiss, L. G., van de Vijver, F. J. R., & Saklofske, D. H. (Eds.) (2003). *Culture and Children's Intelligence: Cross-Cultural Analysis of the WISC-III*. San Diego, CA: Academic Press.

Grégoire, J., & Wierzbicki, C. (2007). Analyse de la dispersion des Indices du WISC-IV en utilisant l'écart significatif par rapport à la moyenne des quatre Indices [Analysis of the WISC-IV Index score scatter using the significant deviation from the mean of the four Index scores]. *Revue Européenne de Psychologie Appliquée*, 57, 101–106.

Grigorenko, E. L., Geissler, P. W., Prince, R., Okatcha, F., Nokes, C., Kenny, D. A., Bundy, D. A., & Sternberg, R. J. (2001). The organisation of Luo conceptions of intelligence: A study of implicit theories in a Kenyan village. *International Journal of Behavioral Development*, 25, 367–378.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229–244.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In: R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 3–38). Mahwah, NJ: Lawrence Erlbaum.

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

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

Jirsa, J. E. (1983). The SOMPA: a brief examination of technical considerations, philosophical rationale, and implications for practice. *Journal of School Psychology*, 21, 13–21.

Johnson, D. L., & Danley, W. (1981). Validity: Comparison of WISC-R and SOMPA estimated learning potential scores. *Psychological Reports*, 49, 123–131.

Longman, R. S. (2005). Tables to compare WISC-IV index scores against overall means. In: A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV. Clinical Use and Interpretation* (pp. 66–69). San Diego, CA: Elsevier Academic Press.

Mercer, J. R. (1979). *System of Multicultural Pluralistic Assessment. Technical manual*. San Diego: The Psychological Corporation.

Nkaya, H. N., Huteau, M., & Bonnet, J. (1994). Retest effect on cognitive performance on the Raven-38 Matrices in France and in the Congo. *Perceptual and Motor Skills*, 78, 503–510.

Ombredane, A., Robaye, F., & Plumail, H. (1956). Résultats d'une application répétée du matrix-couleur à une population de Noirs Congolais [Results of a repeated application of the colored matrices to a population of Black Congolese]. *Bulletin, Centre d'Etudes et Recherches Psychotechniques*, 6, 129–147.

Prifitera, A., Saklofske, D. H., & Weiss, L. G. (2005). *WISC-IV. Clinical Use and Interpretation*. San Diego, CA: Elsevier Academic Press.

Serpell, R. (1993). *The Significance of Schooling. Life-Journeys In an African Society*. Cambridge, UK: Cambridge University Press.

Stephenson, M. (2000). Development and validation of the Stephenson Multigroup Acculturation Scale (SMAS). *Psychological Assessment*, 12, 77–88.

Tulsky, D., Saklofske, D. H., Chelune, G. J., Heaton, R. K., Ivnik, R. J., Bornstein, R., Prifitera, A., & Ledbetter, M. F. (Eds.) (2003). *Clinical Interpretation of the WAIS-III and WMS-III*. San Diego: Academic Press.

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. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13, 29–37.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, 54, 119–135.

van de Vijver, F. J. R., & van de, E. (2003). Test adaption/translation methods. In: R. Fernandez-Ballesteros (Ed.), *Encyclopedia of Psychological Assessment* (pp. 960–964). Thousand Oaks: Sage.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children-IV*. San Antonio, TX: Psychological Corporation.

Weiss, L. G., & Prifitera, A. (1995). An evaluation of differential prediction of WAIT achievement scores from WISC-III FSIQ across ethnic and gender groups. *Journal of School Psychology*, 33, 297–304.

Weiss, L. G., Prifitera, A., & Roid, G. (1993). The WISC-III and the fairness of predicting achievement across ethnic and gender groups. *Journal of Psychoeducational Assessment, Monograph Series*, 15(2).

Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (2006). *WISC-IV Advanced Clinical Interpretation*. San Diego, CA: Academic Press.

Yan, G., & Saklofske, D.H. (2004). *Intelligence: Views of Chinese psychologists*. 28th International Congress of Psychology, August. Beijing, China.

Author Queries

{AQ1} Please provide affiliation for the author Jianjun Zhu.
{AQ2} Please provide more details in the affiliation for all the authors.
{AU3} please note that the cross reference is not given in the reference list.
{AU4} Please provide place.
{AU5} Please update.