7

The Status of Intelligence as a Panhuman Construct in Cross-Cultural Psychology

Johnny R. J. Fontaine and Ype H. Poortinga

In early research with intelligence tests, it was found time and again that people of European descent outperformed others. Such score differences were widely interpreted in terms of innate differences in mental capacities. A strong reaction followed: comparison of intelligence test scores between populations was deemed inherently discriminatory and should be abandoned. As a consequence, research on intelligence was greatly reduced in cross-cultural psychology. The viewpoint of this chapter is that in a shrinking world, often equated with a global village, the notion of intelligence has to be either abandoned entirely or conceptualized and applied as a feature of human psychological functioning everywhere.

The first section of this chapter outlines the early history of research comparing intelligence test scores between populations as well as the

J. R. J. Fontaine
Ghent University, Ghent, Belgium
e-mail: Johnny.Fontaine@UGent.be

Y. H. Poortinga (✉)
Tilburg University, Tilburg, Netherlands
e-mail: Y.H.Poortinga@tilburguniversity.edu

R. J. Sternberg, D. D. Preiss (eds.), *Intelligence in Context*,
https://doi.org/10.1007/978-3-030-92798-1_7

reactions to it. Next to a few research traditions claiming support for these earlier findings, two dominant reactions have emerged: (i) rejection of the idea that intelligence can be a common construct for all humankind and (ii) critical examination of the psychometric comparability of intelligence assessment. The second section presents a broad conceptual and methodological framework from cross-cultural psychology about comparability of psychological constructs and assessment across populations. In the third section, this framework is applied to identify weaknesses and strengths of two different positions about population differences in intelligence and to explore the scope for transfer and adaptation of tests.

Some Historical Trends

There is a widespread tendency to evaluate one's own group as superior to other groups, and this clearly predates intelligence testing. The explanations of surmised differences have varied over time and place. In Europe during the Enlightenment, reference was made to the external condition of climate. Temperate climates, as found in Western Europe, were considered more conducive to the development of "*civilisation*" (high culture) than arctic or tropical regions. Somewhat later, Darwin's evolution theory provided a rationalization for the superiority of white people; allegedly, they had evolved further. In the twentieth century, psychologists obtained with intelligence tests a powerful tool that could help assess how smart people are. In study after study, evidence of superiority of Europeans emerged (see Mann, 1940 for a summary). A research program that can serve as an illustration is that of Porteus (e.g., Porteus, 1937), who administered a maze test to peoples across the world. He considered this paper-and-pencil test to reflect foresight and planning, which in his view was the essence of "intelligence", conceptualized as an inherited capacity. On the basis of the score distributions of small ad hoc samples, he provided a ranking of several populations, qualified as "races", in which the Bushmen, or San, of the Kalahari Desert gained the lowest position, followed by the Australian Aboriginals, and in which "Caucasians" held the top position. Of interest for the present discussion is Porteus' rejection of criticisms challenging his findings (e.g., Klineberg, 1935). Notably, he insisted that performance on tests with concrete materials, such as colored blocks or mazes, does not depend on prior experience, such as schooling; careful instruction was, according to Porteus, the only prerequisite for obtaining good estimates of inborn intellectual capacity. The interpretation of scores on Western tests as reflecting innate differences in intelligence has contributed to a history of discrimination and has made of intelligence a deeply and widely challenged concept.

Three broad traditions can be distinguished that can help explain contemporary views on intelligence as a psychological construct and the use and misuse of intelligence tests in international and cross-cultural settings. The first tradition is seeking further confirmation of population differences in intelligence as inherited. The second tradition is to reject intelligence as a panhuman concept. The third tradition, emphasized in this chapter, is to accept intelligence as a common psychological concept, but to examine critically the scope for cross-cultural comparison of scores obtained with tests.

Seeking Further Confirmation

Strong claims about both universality of intelligence and the validity of common assessment instruments are associated with the notion of “g”. The basis for these claims is the positive manifold of correlations between scores on very different subtests in intelligence batteries. In multivariate analysis with a variety of cognitive ability tests, a single higher-order factor emerges, which is called “g” (e.g., Carroll, 1993). Empirically, cognitively demanding tests (e.g., tests of abstract reasoning) tend to have high loadings on this latent variable, and the size of the loadings also tends to correlate with the size of population differences in score distributions.

From this constellation of findings, some authors have continued to infer that observed differences between populations (especially populations defined as “races”) must be due to genetic inheritance (Jensen, 1974; Lynn, 2006). However, there is explicit empirical evidence incompatible with this position. We mention three key findings. First, cognitively complex tasks with a high “g” loading have been found to be more context-dependent[1] than those with lower "g" loadings, uprooting claims of high heritability at population level for tasks with high "g" loadings (Helms-Lorenz et al., 2003; Kan et al., 2013). Second, differences in context almost invariably come with differences in the experiences needed to respond quickly and accurately to the items in an ability test. In cross-cultural research, effects of differences in stimulus familiarity have been demonstrated extensively for speeded tasks, such as Choice Reaction Time tasks (that are part of the "g" hierarchy according to Jensen and Lynn). In addition, it has been shown that a large amount of experience (training) is needed to reduce such effects (e.g., Sonke et al., 2008). In a set of tasks presenting figural codes and Roman letters to Iranian migrant students and Dutch students in the Netherlands, Sonke et al. (1999) found faster response times for the Dutch students. When (four) easily distinguishable Arabic letters were used as stimuli, the Iranian students responded faster than the Dutch students did. Training sessions for both samples with the letters from the less familiar alphabet did not change the differences substantially. Thorough familiarity requires extensive experience, as demonstrated for traditional Morse code telegraphers, who show slight increases in performance even after years of practice (e.g., Fitts & Posner, 1968). The third finding is that the mean level of performance on intelligence tests within a population can change dramatically over time. Increases of more than 1.0 standard deviation in the mean of test score distributions from one generation to the next have been observed in some Western populations. Such changes are referred to as the Flynn effect (Flynn, 1987), after the author who showed this effect in longitudinal data sets. There are debates over explanatory factors, including school education and economic prosperity (e.g., Pietschnig & Voracek, 2015). The point to note here is that substantial changes in score levels over one or two generations are a further argument that population means on intelligence tests to an important extent reflect context effects. In terms of genetic inheritance, we cannot be, on average, much more or much less intellectually gifted than our parents and grandparents.

[1] Context refers to the social and ecological environment in which humans function. The term has a more limited meaning than "culture" (see Poortinga, 2021).

In our opinion, such findings as mentioned make valid assessment of population differences in intelligence as an inherited capacity fictitious.[2] Therefore, this position is not further discussed in this chapter.

Rejecting Intelligence as a Universal Concept

An important tradition in cross-cultural psychology that emerged as a reaction against conceptualization and test use differentiating between human groups holds that intelligence needs to be conceptualized differently for many non-Western populations. Needless to add, that, as a consequence, operationalization used for assessment also has to be developed within each local context. For example, Mundy-Castle (1974) distinguished two aspects in traditional African "intelligence", a technological aspect and a social aspect. In his view, the Western world has emphasized the first at the expense of the second. The importance of the social aspect was linked by Mundy-Castle to the socialization of African children. Dasen et al. (1985) examined the concept of *n'glouèlê* among the Baoulé in Ivory Coast. They found both social and technological components, with the technological or cognitive dimension being subordinate to the social dimension. Grigorenko et al. (2001) identified in an ethnographic study among the Luo in Kenya four concepts on how locally qualities of individual children are assessed. Ratings for these concepts were obtained for a sample of children from others (peers, teachers, community elders) who had first-hand knowledge of these children. For each of the three sets of ratings, Principal Components Analysis led to the identification of two components: cognitive competence and social-emotional competence.[3]

[2] We do not argue that genetic underpinnings of population differences in intellectual functioning can be ruled out. However, before we can even consider to examine such underpinnings, the equivalence and validity of test score differences have to be demonstrated. Moreover, it requires the identification of genetic variations that directly causally affect intellectual functioning within and across populations. The only relevant empirical research to date are genome-wide association studies that explore correlations between the genetic variations and intelligence scores (GWAS studies; e.g., Lee et al., 2018). However, they offer no evidence for the equivalence and validity of the test score differences, nor can they identify direct causal effects.

[3] Grigorenko et al. administered also two Western intelligence tests to the children. Only the cognitive component showed some (moderate) correlations with scores on these tests.

Critically Examining Assessment

By the 1960s, most psychologists had realized that the interpretation of population differences in score distributions on psychometric tests is highly problematic. Notions that tests can be “culture-free” or “culture-fair” were challenged. In a commentary in the proceedings of a major conference on cross-cultural testing held in 1971, the editors noted:

> It is hazardous to interpret a test in its new setting as if it measured ‘the same thing’ as it did originally. Serious questions of comparability arise for translated performance tests as well as verbal tests. (Cronbach & Drenth, 1972, p. 470)

A plethora of approaches followed, addressing either the measurement or the conceptualization of intelligence, or both of these. Cattell (1963), for instance, formalized the idea that some ability tests are more dependent on specific knowledge and experience than other tests, with the distinction between crystallized and fluid intelligence. Crystallized intelligence reflects previously acquired knowledge and skills (e.g., tests of vocabulary). Fluid intelligence reflects cognitive processing, notably reasoning and problem-solving. Fluid tests tend to be seen as operationalizations of intelligence that can be used for assessment of intelligence across populations. Such tests are sometimes called “culture-reduced” tests. The best-known example are Raven’s Progressive Matrices tests (RPM; Raven et al., 2004). However, avoiding tests for crystalized intelligence does not solve the problem as experience and familiarity affect all test results.

Reuning and colleagues conducted extensive studies of the cognitive and perceptual abilities of the Bushmen (Reuning & Wortley, 1973). Addressing measurement issues, their focus was on the adaptation and transfer of existing assessment instruments. For example, with a three-dimensional device to present items from Porteus’ maze test in a more context appropriate manner, they found with a fairly large sample that the Bushmen performed rather well.[4] They thus demonstrated that the

[4] In a fairly recent monograph, Lynn (2006) continues to attribute subnormal intelligence to the Bushmen, referring to work by Reuning. This is a blatant misrepresentation of Reuning’s views. On the basis of field observations and of their performance on a range of tests, he considered the Bushmen to be “clever” (see Reuning & Wortley, 1973, for extensive evidence).

original Mazes test—a nonverbal test—used by Porteus, severely underestimated (the form of) intelligence assessed with solving mazes.

We can conclude that there is nowadays a broad tendency to treat comparisons of scores on intelligence tests with suspicion. We wish to state explicitly that there are valid reasons for this suspicion. In the next section, we make suggestions on how to move forward.

Conceptual and Methodological Framework for Comparability

A broad framework addressing both conceptual and psychometric issues of comparability has been developed in cross-cultural psychology. The conceptual issues center on the contrast between universalism and relativism, which differ on the question whether or not the psychological traits and processes that are underlying daily psychological functioning differ between groups of humans, labeled as cultural groups (e.g., Berry et al., 2011; Fontaine, 2011; Fontaine & Breugelmans, 2021). The methodological issues center on the analysis of bias and equivalence in psychological data (Fontaine, 2005, 2008; Poortinga, 1989; Poortinga & Van de Vijver, 1987; Van de Vijver & Poortinga, 1997; Van de Vijver & Leung, 2021; Van de Vijver et al., 2008). This framework can help to guide us with the theoretical questions to be asked and the choice of methodological and psychometric tools to be used in empirical analysis of intelligence. In this section, we present an integrated version of this framework (see Table 7.1), and in the next section we address its application to issues encountered in the intelligence domain.

In the framework, a distinction is made between four conceptual positions on the comparability of constructs across populations: full relativism, construct universalism, domain universalism, and full universalism. Each of these four positions is linked to the empirical requirements for comparability to justify that position. The required levels of equivalence as well as sources of bias to be excluded for each conceptual position are mentioned.

Table 7.1 Overview of the integrative framework for comparability of test scores across populations

	Full relativism	Construct universalism	Domain universalism	Full universalism
Theoretical claims				
There exists behavior within each of the populations that can be accounted for by the same theoretical variable	No	Yes	Yes	Yes
The domain of behavior accounted for by the theoretical variable is highly overlapping between populations	No	No	Yes	Yes
Populations can be compared quantitatively on the theoretical variable	No	No	No	Yes
Empirical conditions				
Required level of equivalence	No equivalence possible	Construct equivalence	Domain and structural equivalence	—Metric equivalence for comparisons of score differences between more than one measurement —Full score equivalence for direct comparison of scores
Sources of bias and threats to be disconfirmed	Lack of validity evidence within specific population	Construct bias	Partial domain nonoverlap leading to: – Construct underrepresentation – Construct irrelevance	Method bias Item bias

Four Conceptual Positions

The conceptual positions are based on three basic claims that can be made about a psychological construct (like intelligence) and the way it becomes manifest in observable behavior:

(i) There exists behavior within each of the populations examined that can be accounted for by the same theoretical variable. For example, in every population evidence of inductive and deductive reasoning can found.

(ii) The domain of behavior accounted for by the theoretical variable is highly overlapping between populations. For example, in populations that have Western-type schooling, pupils learn to apply inductive and deductive reasoning to both familiar and new, unfamiliar problems.

(iii) Populations can be compared quantitatively on the theoretical variable. For example, the validity can be demonstrated of population differences in average inductive and deductive reasoning ability as assessed with a common reasoning test.

These three claims are hierarchically ordered. When the first claim is rejected, the two other claims have to be rejected also. When the last claim is made, the first two claims are implied. Thus, based on the three claims, four positions are possible:

Full Relativism

When none of the three claims is made, we are dealing with the position of "full relativism". Population-specific processes and/or traits are needed to account for manifest behavior. The construct concerned does not cross population boundaries: it can only be studied within the context of a specific population.

Construct Universalism

When only the first claim is made, we are dealing with the position of "construct universalism". The same theoretical framework can be applied to account for behavior across populations, but without the behavior manifestations necessarily being the same in each of the populations. The fact that the behavioral repertoire differs between populations does not imply that different explanatory variables are needed to account for it. For instance, empirical evidence makes it plausible that logical reasoning follows Aristotelian principles everywhere (e.g., Scribner, 1979). The fact that non-Western populations have more difficulties solving deductive reasoning tasks with typical Western items does not prove that the processes to solve these tasks are a Western construction not applicable elsewhere. Rather, the type of problems to which Aristotelian reasoning principles are applied depends on the relevance for a specific context (e.g., van de Vijver & Willemsen, 1993).

Domain Universalism

When the first two claims are made, but not the third claim, we have the position of "domain universalism". It means that the domains of observable behavior accounted for by the same explanatory variable shows sizable overlap across populations. However, this position does not imply that direct quantitative comparisons of test scores between populations are valid. It is a fundamental insight in psychological assessment that concrete behavior has multiple determinants (e.g., Messick, 1989). Thus, differences between populations in nontargeted constructs may affect the observed behavior. For instance, populations differ in the extent to which speed versus accuracy are valued in solving cognitive tests, in Western populations speed being more valued than accuracy (e.g., Sternberg et al., 1981). When speeded cognitive tests are used, in which performance heavily depends on the trade-off between speed and accuracy, scores cannot be compared across populations that have different expectations about the optimal trade-off. Thus, within this position the same theoretical framework is used to account for behavior within populations, but generally one refrains from direct population comparisons.

Full Universalism

If all three claims are made, full universalism is specified. This implies that not only the same theory is assumed to account for the same observed behavioral repertoire, but that there is scope for valid quantitative comparison of scores and/or of score differences (e.g., across measurement occasions) between populations (see below).

Levels of Equivalence

The lead question is: Does a score of a test taker have the same meaning across certain populations in terms of the intended interpretation? In other words, are scores comparable; are they equivalent, are they unbiased? In samples of test takers, both item scores and test scores form variables of which equivalence can be analyzed. There is more to the answer than a simple yes or no; various levels of equivalence can be distinguished. Analysis of equivalence takes place mostly, though not exclusively, through examination of statistical conditions that are set in such a way that they are likely to be satisfied by equivalent sets of scores, but not by nonequivalent or biased scores.

In this chapter, four hierarchically ordered levels of equivalence are distinguished:

Construct Equivalence

Comparison of data always requires that there is construct equivalence, that is, the construct is identifiable in all populations in a study. For this type of equivalence, validity should be demonstrated of (possibly context-specific) instruments using the same theoretical framework.

Content and Structural Equivalence

For the same instrument to be used validly across contexts and populations, content and structural equivalence must be demonstrated. The content of the instrument should be relevant and representative within each of the populations, and the dimensions assessed with the instrument (internal structure) should be the same. Content and structural equivalence do not necessarily imply the same quantitative scale for all populations. The state of affairs is reminiscent of recordings of temperature made on the Celsius scale in one setting and on the Fahrenheit scale in another setting; temperature is assessed everywhere, but readings of the same temperature differ.

Metric Equivalence or Measurement Unit Equivalence

For a test, the measurement units on the scoring scale are the same across populations, but there may not be a common scale anchor (e.g., the same origin or zero point). A given difference between two scores can be interpreted in the same way, independent of the population in which it was found. Imagine that recordings of temperature are made on the Celsius scale in one setting and the Kelvin scale in another setting. Although no direct comparisons can be made, it is possible to compare directly the difference between recordings (e.g., between averages of summer and winter temperatures in various locations).

Scale Equivalence or Full Score Equivalence

A score of a given value can be interpreted in the same way independent of the population of a test taker. Imagine that recordings of temperature are made on the same scale, for example Celsius scale, in all settings. Only when this type of equivalence is achieved can direct comparisons be made between populations.

These four levels of equivalence can be linked to the three universalist positions.[5] Construct universalism requires construct equivalence. Domain universalism allows the construction of a common instrument that must satisfy both content and structural equivalence. The content of the common instrument is relevant and representative for the respective domains in the different populations (content equivalence), and it is

[5] Relativism precludes any form of equivalence as a construct does not cross borders.

assessing the same psychological dimensions across populations (structural equivalence). For full universalism, metric or full score equivalence is needed, depending on whether only score differences or scores are compared between populations.

Sources of Bias

Bias, or lack of equivalence, refers to the sources that can distort valid comparisons (e.g., Van de Vijver & Leung, 2021). Major sources of bias are construct bias, construct underrepresentation and irrelevance, method bias, and item bias.

Construct Bias

Construct bias means that a theoretical framework, or a part of it, is tied to the context of a specific population, and poorly crosses population boundaries. Theories are typically developed within a specific context and may confound universal and population-specific aspects. For example, the operational rules for multiplication with and multiplication without an abacus differ, even though they are based on the same arithmetical principles. The development of a theory for numerical ability may be influenced by whether or not an abacus is used in the local context.

Construct Underrepresentation and Construct Irrelevance

Construct underrepresentation and irrelevance can occur when there is less than full overlap across populations of the domains accounted for by the construct. A simple example is the inclusion of items requiring root extraction in a test of numerical ability when in some populations root extraction forms part of the school curriculum for a certain age group and in other populations it is not included in the curriculum. Having root extraction items in a test will lead to construct irrelevance in the latter populations, while omitting such items will create construct underrepresentation in populations where root extraction has been taught.

Construct irrelevance can be identified by applying psychometric analyses for item bias on the data obtained with the test itself. Construct underrepresentation requires additional information on the domain in populations where the instrument is going to be applied. Such an analysis is conducted with exploratory (often qualitative) methods.

Method Bias

Method bias refers to differential distortion between populations by biasing factors affecting most or all items in a test, for example due to differential stimulus familiarity or differences in guessing strategy with multiple-choice items. For tracing method bias in test scores, multimethod approaches involving additional instruments are needed.

Item Bias

Item bias entails a distortion in (one or more) separate items, for example due to poor translation. Item bias can be identified by psychometric analyses on the data obtained with the same instrument across populations.

These four sources of incomparability can be linked to three of the four conceptual positions. Since no comparisons are made in the full relativism position, none of these four types of bias applies there. A threat to this position is the lack of validity evidence for population-specific constructs. Just like not all concepts proposed in Western psychology have received empirical support, the fact that a population-specific concept is identified in exploratory (qualitative) research does not guarantee without further analysis of construct validity that it is a valid psychological construct for the population concerned. For construct universalism, a major threat is construct bias, while for domain universalism the major threat consists of construct underrepresentation and irrelevance. Full universalism can be distorted both by method bias affecting all items in an instrument and by item bias making direct comparisons of scores or score differences between populations flawed.

Applying the Framework to Intelligence

One of the more recent traditions identified in the first section focuses on what is common in intellectual functioning across human populations, and the other tradition looks for context-specific forms of intelligence. The framework presented in the second section is meant to provide a set of tools to clarify the strengths and the weaknesses of both traditions, and how these affect the transfer of intelligence tests across populations and test score interpretations.

Context-Focused and Assessment-Focused Approaches to Intelligence

The strength of relativist approaches to intelligence is the study of the construct and the domain as they emerge through performance in a specific context. The focus is on what appears to be relevant for a specific context. However, most studies just *assume* both construct validity within the specific context and lack of comparability across contexts (see examples mentioned in the first section of this chapter). Convergent evidence may be quoted extensively, but critical examination is lacking. Identifying how a domain of psychological functioning is conceptualized within a particular population does not make in itself a valid psychological construct in that population (as the history of psychology has amply demonstrated). Assessment of intelligence does require not only the operationalization of specific forms of intelligence but also empirical evidence that a context-specific test behaves as the context-specific theory would suggest. Moreover, the fact that a context-specific instrument provides a valid assessment of intelligence in a certain population does not in itself justify the claim that the construct measured is a context-specific construct. To the best of our knowledge, there is no research ruling out that a universal intelligence theory can account for what is measured with context-specific instruments. Rather, there is contrary evidence. One of the strongest claims for the indigenous nature of intelligence is the focus on intra- and interpersonal functioning in non-Western populations that has been argued to be incompatible with the Western intelligence construct that is technologically focused. However, with the emerging construct of emotional intelligence in Western psychology this claim no longer holds. Maximum performance tests have been developed in Western populations assessing the abilities to perceive, understand, and regulate emotions, and substantial correlations with traditional cognitive tests of intelligence have been found (MacCann et al., 2014).

A strong point of the tradition focused on comparison of intellectual functioning with intelligence tests is that it does not take equivalence of these instruments for granted. Scores on existing instruments are analyzed for lack of equivalence, or “bias”. Such psychometric analyses investigate the comparability of the internal structure and, in addition, allow the identification of items that function differently across populations (item bias). Most difficult to identify are biasing factors that affect most or all items in an instrument and that are referred to as “method bias” (see above). An example of a possible distorting factor with the RPM (and similar tests) is an interaction between direction of reading and writing in a language and the left-right orientation embedded in the construction of the items.

A weak point of the assessment-focused approach is that it concentrates on the test itself (and possibly method factors affecting all items in the test). However, such procedures do not address identification of construct underrepresentation. Moreover, as a rule studies in this tradition simply assume construct validity across populations without providing empirical evidence, beyond what is generated by the psychometric analysis of bias. The nomological network of a test or test battery is seldom studied. Processes and mechanisms underlying responses are even less studied. Moreover, population-comparative research has focused on nonverbal tests at the expense of verbal tests. While verbal tests tend to be more context sensitive than performance tests, simply omitting verbal tests leads to gross underrepresentation of the intelligence domain. Verbal abilities form an essential part of the intelligence construct that cannot be captured with nonverbal tests.

In summary, the two research traditions tend to either simply reject (relativist approach) or simply assume (instrument approach) construct equivalence. The position of construct universalism with its requirement of construct equivalence has not been a major target of investigation. Constructing a test that can only assess intelligence within a specific population does not in itself justify the claim that the intelligence construct is context-bound.

Trade-offs Between Positions

Within the framework outlined in the previous section, the choice is not between either accepting full comparability of intelligence scores or assuming context specificity of the intelligence construct. There are intermediate positions that require increasingly restrictive forms of equivalence.

Within the relativist position, environmental factors can be investigated only within a specific context. If the intelligence construct does not cross borders, then this also applies to the factors that affect intelligence. It requires at least a position of construct universalism to investigate meaningfully whether the same environmental variable (e.g., the amount of time a child interacts with adults) has a similar impact across contexts. A position of domain universalism makes research easier, as it allows research with the same assessment instruments across contexts, but (as in the position of construct universalism), it does not allow to compare the size of impact of an environmental factor. With both positions, the scale units on which the construct is measured can differ between groups. Only within the position of full universalism, studies of the size of environmental factors across contexts can be undertaken. Depending on the type of comparison one wants to make, different psychometric conditions associated with different levels of equivalence apply (see Table 7.1).

We give an example of a comparative study where only score differences needed to be analyzed in order to answer the question posed. Brouwers et al. (2009) analyzed studies of the Flynn effect during the last century based on the RPM tests across the world. They found that this effect was steeper in non-Western compared to Western populations and attributed this to a much larger change in environmental factors stimulating intellectual performance in the non-Western populations, notably implementation of universal education in countries with previously high levels of illiteracy. As the Flynn effect offers one of the strongest forms of evidence for environmental impact on intelligence, the claim that it emerges more strongly in non-Western than in Western societies is highly informative. However, this claim can only be justified if score differences have the same meaning across contexts and populations (requiring the form of equivalence called "metric equivalence" earlier on (see Table 7.1). For making direct comparisons of test score levels, full score equivalence is needed. For instance, the studies by the OECD on school performance, the PISA studies (https://www.oecd.org/pisa/) link country differences in cognitive achievement tests to differences in the educational system of countries and formulate advice on how to improve the educational system. Such interpretations and such advice can only be justified if a given test score allows the same interpretation, independent of the population in which that score was obtained.

This analysis shows that a relativist position is not the only way to do justice to the effects of the context in which a population lives. A relativist position allows to make some claims about environmental effects (those that are unique to a specific context), but precludes other claims (identifying effects that are similar across contexts). There is thus a trade-off between the position one takes and the type of claims one can make about contextual factors. In an increasingly globalizing world, in which resources and especially adverse life conditions are unevenly distributed, being able to study effects of context factors on (forms of) intelligence is highly relevant. Factors such as poverty, lead pollution, COVID infections, and so on all have demonstrably negative effects on intellectual performance, and some populations (within countries and across countries) are more affected by these factors than other populations. Being able to identify these effects, the underlying processes, and, especially, how they can be mitigated is highly relevant. One has to avoid errors on both sides: one can unjustifiably compare populations and misrepresent how context-specific factors mold the construct and expression of intelligence, or one can unjustifiably reject comparisons and miss out on identifying context factors that operate on expression of intelligence across populations.

Transfer of Instruments

The four positions identified earlier on are hierarchically ordered in terms of increasing restrictions on equivalence. Thus, the positions represent four possible states of affairs, in which some claims can be meaningfully made and other claims are precluded. Which position best represents the state of affairs in a particular instance requires empirical scrutiny. A central theme in this endeavor is the transfer of instruments. Three types of instrument transfer have been distinguished in the literature: assembly, revision, and adoption (e.g., Van de Vijver & Poortinga, 2005, 2020). With *adoption,* the same instrument is applied across populations after careful translation of the instructions, the items, and other test materials. *Revision* implies that the original instrument is taken as a basis and elements of the instrument (such as some of the items or the response scale) are changed to make them more appropriate to the local context. When only the design and the broad themes and goals of the original instrument are kept, but new content and possibly other administration methods are developed, one speaks of *assembly*. The zero option is to develop a completely new instrument for a specific population. In this way, an instrument can be optimally adapted to the behavior repertoire of that population. However, this is a time- and cost-intensive procedure. Probably more important, there is no accumulation of knowledge on the target construct and/or domain across populations.

The type of transfer that has been chosen needs to be justified by the level of equivalence that can be demonstrated for the adapted version of the instrument in the target population. When a completely new instrument is developed to assess a context-specific form of intelligence, predictions characteristic for the context-specific form of intelligence need to be examined and justified empirically within the target population. Construct equivalence needs to be demonstrated across the populations concerned when test adaptation amount to assembly. In the case of revision, equivalence of the domain as well as structural equivalence across populations has to be shown. In the case of adoption and quantitative comparison of scores across populations, additionally strict requirements for score equivalence have to be met.

Ideally, an instrument, new or adapted, is developed by a group of experts representing each of the populations where it will be used. This strategy is followed in large-scale country projects, comparing school performance data, such as Program for International Student Assessment (PISA) of the OECD and Trends in International Mathematics and Science Study (TIMMS) of the IEA. By involving experts from all populations right from the start, context specificity and bias can be avoided in the theoretical conceptualization, the item content of instruments, as well as the way in which the items are displayed. Unfortunately, this ideal situation is only possible for large-scale well-funded projects. For most research, this ideal situation is not within reach. At the same time, it is clear that only a careful translation and investigating the internal structure and item bias is insufficient. The relevance and representativeness, as well as the adequacy of the method through which the content is offered within the new context, needs to be investigated. The ITC Guidelines for Translating and Adapting Tests (International Test Commission, 2017) contain many practical recommendations.

Whether and to which extent instruments are transferable and which level of equivalence can be reached will depend on the (lack of) overlap in behavior repertoire between the populations involved, as well as on the aspects of the intelligence domain that are studied. The larger the differences in behavior repertoire and the more a test assesses acquired knowledge (think of the vocabulary subtest of the Wechsler scales) rather than underlying information processes (think of the digit span test that assesses the size of the phonological loop in the Wechsler scales), the less likely adoption can be justified and the less likely that strict demands on equivalence will be met. Even for populations with large overlap in behavior repertoire and for tests that assess underlying information processes, test adoption needs to be supported empirically and strict conditions of equivalence need to be demonstrated.

Conclusions

There lies a whole world between the assertion that tests prove innate differences in intelligence between populations and the claim that intelligence can only be defined and studied within the specific context in which it becomes manifest. These extremes are often the only positions that are articulated. However, there is a danger that they function as straw men in debates; finding evidence that rejects one extreme position does not necessarily form evidence supporting the other extreme position. Convincing evidence for population differences in distributions of test scores does not imply that test scores are strictly comparable. Equally, the demonstration of context specificity in performance on cognitive tasks does not necessarily imply that the intelligence construct is context-specific.

In this chapter, we have argued that fighting the misuse of intelligence tests, typical for the early history of psychology, does not require the concept of intelligence to be abandoned. Intelligence should be defined contextually where needed, but as a common human capacity where possible. Transfer and adaptation of tests to other populations as where they originated is connected to the transfer of insights about validity. Only for (the aspects of) intelligence that can be defined and assessed in the same way across contexts and populations can factors be identified that hamper intellectual development across these populations and that contribute to marginalization and exclusion. Comparability of psychological data can neither be accepted nor be rejected out of hand, but is a matter of empirical scrutiny.

References

Berry, J. W., Poortinga, Y. H., Breugelmans, S. M., Chasiotis, A., & Sam, D. L. (2011). *Cross-cultural psychology: Research and applications* (3rd ed.). Cambridge University Press.

Brouwers, S. A., Van de Vijver, F. J. R., & Van Hemert, D. A. (2009). Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences, 19*, 330–338.

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

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1–22.

Cronbach, L. J., & Drenth, P. J. D. (Eds.). (1972). *Mental tests and cultural adaptation*. Mouton.

Dasen, P. R., Dembele, B., Ettien, K., Kabran, K., Kamagaté, D., Koffi, D. A., & N'Guessan, A. (1985). N'gloulé, l'intelligence chez les Baoulé [N'goulé, intelligence with the Baoule]. *Archives de Psychologie, 53*, 293–324.

Fitts, P. M., & Posner, M. I. (1968). *Human performance*. Brooks/Cole.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191.

Fontaine, J. R. J. (2005). Equivalence. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, pp. 803–813). Academic Press.

Fontaine, J. R. J. (2008). Traditional and multilevel approaches in cross-cultural research: An integration of methodological frameworks. In F. J. R. Van de Vijver, D. A. Van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 65–92). Lawrence Erlbaum Associates.

Fontaine, J. R. J. (2011). A fourfold conceptual framework for cultural and cross-cultural psychology: Relativism, construct universalism, repertoire universalism and absolutism. In F. J. R. van de Vijver, A. Chasiotis, & S. M. Breugelmans (Eds.), *Fundamental questions in cross-cultural psychology* (pp. 165–189). Cambridge University Press.

Fontaine, J. R. J., & Breugelmans, S. M. (2021). Emotion between universalism and relativism: Finding a standard for comparison in cross-cultural emotion research. In M. Bender & B. G. Adams (Eds.), *Methods and assessment in culture and psychology* (pp. 144–169). Cambridge University Press.

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

Helms-Lorenz, M., van de Vijver, F. J. R., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: g or c? *Intelligence, 31*, 9–29.

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). [www.InTestCom.org].

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 90*, 185–244.

Kan, K.-J., Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. J. (2013). On the nature and nurture of intelligence and specific cognitive abilities: The more heritable, the more culture dependent. *Psychological Science, 24*, 2420–2428.

Klineberg, O. (1935). *Race differences*. Harper & Row.

Lee, J. J., Wedow, R., Okbay, A., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics, 50*, 1112–1121.

Lynn, R. (2006). *Race differences in intelligence, an evolutionary analysis*. Qashington Summit Publishers.

MacCann, C., Joseph, D. L., Newman, D. A., & Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: Evidence from hierarchical and bifactor models. *Emotion, 14*, 358–374.

Mann, C. W. (1940). Mental measurements in primitive communities. *Psychological Bulletin, 37*, 366–395.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Mundy-Castle, A. C. (1974). Social and technological intelligence in Western and non-Western cultures. *Universitas, 4*, 46–52.

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science, 10*, 282–306.

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

Poortinga, Y. H. (2021). *Concept and method in cross-cultural and cultural psychology*. Cambridge University Press.

Poortinga, Y. H., & Van de Vijver, F. J. R. (1987). Explaining cross-cultural differences: Bias analysis and beyond. *Journal of Cross-Cultural Psychology, 18*, 259–282.

Porteus, S. D. (1937). *Primitive intelligence and environment*. Macmillan.

Raven, J., Raven, J. C., & Court, J. H. (2004). *Manual for Raven's progressive matrices and vocabulary scales*. Harcourt Assessment.

Reuning, H., & Wortley, W. (1973). Psychological studies of the Bushmen. *Psychologia Africana, Monograph Supplement, No. 7.*

Scribner, S. (1979). Modes of thinking and ways of speaking: Culture and logic reconsidered. In R. O. Freedle (Ed.), *New directions in discourse processing* (pp. 223–243). Ablex.

Sonke, C. J., Poortinga, Y. H., & De Kuijer, J. H. J. (1999). Cross-cultural differences on cognitive task performance: The influence of stimulus familiarity. In W. J. Lonner, D. L. Dinnel, D. K. Forgays, & S. A. Hayes (Eds.), *Merging past, present, and future in cross-cultural psychology* (pp. 146–158). Swets and Zeitlinger.

Sonke, C., Van Boxtel, G., Griesel, R., & Poortinga, Y. H. (2008). Brain wave concomitants of cross-cultural differences in scores on simple cognitive tasks. *Journal of Cross-Cultural Psychology, 39*, 37–54.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41*, 37–55.

Van de Vijver, F. J. R., & Leung, K. (2021). *Methods and data analysis for cross-cultural research* (2nd ed.). 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., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Erlbaum.

Van De Vijver, F. J. R., & Poortinga, Y. H. (2020). Dealing with methodological pitfalls in cross-cultural studies of stress. In T. Ringeisen, P. Genkova, & F. T. L. Leong (Eds.), *Handbuch Stress und Kultur: Interkulturelle und kulturvergleichende Perspektiven* (pp. 1–19). [Chapter 2-1]. Springer.

Van de Vijver, F. J. R., Van Hemert, D. A., & Poortinga, Y. H. (2008). Conceptual issues in multilevel models. In F. J. R. Van de Vijver, D. A. Van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 3–26). Erlbaum.

Van de Vijver, F. J. R., & Willemsen, M. E. (1993). Abstract thinking. In J. Altarriba (Ed.), *Culture and cognition* (pp. 317–342). North Holland.