Educational Research and Evaluation  
An International Journal on Theory and Practice  

Publication details, including instructions for authors and subscription information:  
http://www.informaworld.com/smpp/title~content=t714592776

On the meaning of cross-cultural differences in simple cognitive measures  
Fons J. R. van de Vijver  
Tilburg University, Tilburg, The Netherlands  

Online Publication Date: 01 June 2008  


To link to this article: DOI: 10.1080/13803610802048833  
URL: http://dx.doi.org/10.1080/13803610802048833

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf  
This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.  

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
On the meaning of cross-cultural differences in simple cognitive measures

Fons J.R. van de Vijver*

Tilburg University, Tilburg, The Netherlands

(Received 7 January 2007; final version received 27 January 2008)

A set of 5 reaction time tests of increasing cognitive complexity were administered to 35 secondary school pupils in Zimbabwe and The Netherlands at 4 consecutive school days in order to explore the existence and nature of cross-cultural differences on reaction time tests measuring basic cognitive operations. No cross-cultural differences were observed on the simple tests, while the Dutch pupils were faster than Zimbabwean pupils on the more complex tests. Zimbabwean pupils showed larger performance increments at retesting. Two Wechsler Intelligence Scale for Children (WISC) subtests (Vocabulary and Digit Span Forward and Backward), a tapping test, and parental socioeconomic status were used to predict the cross-cultural performance differences. Socioeconomic status and digit span could explain all these differences. It is argued that the cross-cultural performance differences on elementary cognitive tests cannot be interpreted at face value because of their susceptibility to non-target factors, such as test understanding and previous test exposure.

Keywords: cross-cultural differences; elementary cognitive tests; cognitive complexity; Netherlands; Zimbabwe

Introduction

The literature on cross-cultural studies of cognitive tests shows a persistent stalemate. On the one hand, there is a remarkable consistency in empirical findings in that the structure of the cognitive system is found to be stable across cultures (e.g., Irvine, 1979; Jensen, 1998; Van de Vijver, 1997). Carroll's (1993) model of cognitive abilities may well claim universality, despite the culturally restricted, Western research on which it is based. The consistency also applies to cross-cultural differences in score levels. Groups of East-Asian and European descent tend to score higher than groups of African descent (cf. Berry, Poortinga, Segall, & Dasen, 2002; Jensen, 1980, 1998; Van de Vijver, 1997, 2002). On the other hand, interpreting these differences has turned out to be controversial, as exemplified in the old IQ debate recently invigorated by the publication of Herrnstein and Murray's (1994) book (cf. Neisser et al., 1996). Cross-cultural performance differences on intelligence tests may be due to construct-relevant factors, such as reasoning and memory, and to factors other than the target construct, such as motivation, test-wiseness (Sarnacki, 1979), and previous test exposure. Van de Vijver and Leung (1997) have coined the term

*Email: Fons.vandevijver@uvt.nl

ISSN 1380-3611 print/ISSN 1744-4187 online
© 2008 Taylor & Francis
DOI: 10.1080/13803610802048833
http://www.informaworld.com
“method bias” to refer to all method-related factors that give rise to cross-cultural score differences that are not essential to the target construct.

The current study addressed the nature of cross-cultural differences in cognitive tasks. As explained below, the current study attempted to minimize interpretation problems by (a) choosing particular cognitive tasks, namely elementary cognitive tests (ECTs); (b) including measures in the study that are expected to statistically explain observed cross-cultural differences in cognitive test performance; (c) administering the same tests in a longitudinal design and making a cross-cultural comparison of the score gains. Each of these factors which are aimed at maximizing the interpretability of observed performance differences across cultures is described below.

**Elementary cognitive tests**

The current study addressed elementary cognitive tests (ECTs) (Jensen, 1998; Posner & McLeod, 1982; Vernon, 1987), based on the reasoning that cross-cultural differences on these tasks are easier to interpret than score differences on complex cognitive tests such as intelligence tests, which can be influenced by a myriad of factors. However, ECTs should not be viewed as culture-free tests of cognitive functioning. Frijda and Jahoda (1966) argued that culture-free and culture-fair intelligence tests do not exist, as all tests are based on performance that always uses a cultural medium of expression. Still, some tests of cognitive functioning are more influenced by cultural factors than others. Jensen proposed the term “culture-reduced tests” to refer to tests that are relatively little influenced by (unintended) cultural factors. ECTs could be seen as examples of such culture-reduced tests.

Four studies in our research group compared performance differences of Dutch mainstream and immigrant pupils in both ECTs and more complex cognitive tasks. Van de Vijver and Willemse (1991) administered five relatively simple cognitive reaction time measures to 59 mainstream and 47 immigrant children of Grade 6 in The Netherlands. The latter group was mainly of Turkish and Moroccan descent. The test battery (which is described in more detail in the Method section, as it is also used in the present study) uses figures as stimuli. The tests vary in cognitive complexity and range from a simple reaction time task to a task in which geometric patterns have to be mentally added to each other. In addition, the pupils’ scores on the CITO test (a nationwide administered test of school achievement) were available. The test has three subtests: Reading (reading comprehension), Arithmetic (mental arithmetic), and Information (school knowledge, mainly derived from the Science and Social Studies curricula). The cross-cultural differences were small and nonsignificant for the ECTs (average effect size, Cohen’s $d = 0.07$), while the mainstream pupils showed significantly higher scores on Reading ($d = 0.41$) and Information ($d = 0.62$); the difference for Arithmetic was in the same direction ($d = 0.23$) but failed to reach significance. The ECTs showed weak negative correlations with the scholastic achievement scores; in the group of mainstreamers, the correlation went from 0.00 for the simple reaction test to –0.22 for the most complex test (averaged across the three scholastic measures). The correlations were stronger (particularly for the more complex tests) in the immigrant group and ranged from –0.09 to –0.54 (note that negative correlations are expected, because pupils with a better school performance will be faster and will have shorter reaction times). A similar pattern was found for average grade mark. In the mainstream group, the latter variable was unrelated to ECT performance, while in the immigrant group all correlations were negative and three out of the five were significant (average of –0.30). The differences in performance on the ECTs were small and not clearly patterned.
Van de Vijver, Willemse, and Van de Rijt (1993) administered four of the five ECTs of the previous experiment, alongside the Wechsler Intelligence Scale for Children (WISC), to four groups of primary school pupils in The Netherlands from 8 to 11 years of age: 49 mainstream, 34 first-generation, 24 second-generation pupils, and 21 pupils from intercultural marriages (with one Dutch parent). Using tables developed for Dutch mainstream children, the IQs of these four groups were found to be 100, 79, 90, and 97 (effect sizes were –1.60, –0.79, and –0.21, respectively). In both generations, the verbal IQ showed larger differences than the performance IQ. The ECT differences were, again, relatively small and not clearly patterned.

Wagenmakers (1994) administered three tests of the same ECT battery (leaving out the most simple and difficult task) to 73 mainstream and 60 immigrant pupils (mainly of Turkish, Moroccan, and Surinamese descent) of the third grade in The Netherlands. In addition, Raven’s Standard Progressive Matrices were administered. No ECT showed a significant cross-cultural difference (average $d = 0.01$), while the mainstream children showed a significantly higher score on the Raven ($d = 0.63$). The correlations of the ECTs with the Raven were negative and significant; for the two most difficult reaction time tests highly significant correlations of about –0.55 were found in both groups. Correlations between ECTs and average report marks were negative and slightly stronger in the immigrant group than in the mainstream group.

Finally, Helms-Lorenz, Van de Vijver, and Poortinga (2003) administered a simple and a complex ECT to 747 Dutch majority group members and 474 second-generation immigrant pupils in The Netherlands (mainly of Moroccan, Turkish, Surinamese, and Antillean background), aged between 6 and 12 years. The instrument of this study differed somewhat from the previous studies in that the present study employed a mouse as response device while a board with response buttons was used in the above studies. In addition, either of two omnibus intelligence tests was administered to the children. Mainstream pupils showed a higher performance than immigrant pupils; the differences were larger for the intelligence test (average $d$ across subtests $= 0.54$) than for ECTs (average $d = 0.38$). The correlations between the ECTs and IQ (averaged across the two intelligence tests) were about –0.15 for the simple reaction time measure and about –0.25 for the complex measure in both cultural groups.

The four studies showed convergent findings. First, the relationship between speed and intelligence was largely identical for mainstream and immigrant children; in a study among primary school pupils in Guatemala and Austria, Neubauer and Benischke (2002) also found that this relationship was stable across cultures. Second, cross-cultural differences were smaller on ECTs than on intelligence tests and scholastic achievement tests. Third, both in groups of Dutch mainstreamers and immigrants, the correlation between ECTs and both scholastic measures and intelligence test scores tended to become stronger for ECTs of higher cognitive complexity. The upper limit of 0.30 of the (absolute) correlation between ECTs and intelligence or school performance, reported before (e.g., Vernon, 1987), was replicated here.

Our findings are essentially in line with Jensen (e.g., 1985, 1993, 1998), who conducted many studies of ECTs involving European and African Americans. He consistently found performance differences to increase with task complexity. No consistent cross-cultural performance differences were found on the simple reaction time tests; however, more complex tests tended to show a higher performance of European Americans. Jensen has proposed a theoretical framework to explain these findings. His so-called “Spearman Hypothesis” postulates that cross-cultural performance differences on cognitive tests are related to cognitive complexity; tasks with more cognitive complexity (operationalized as
tests with a higher loading on g) showed larger cross-cultural differences, which supports this theory: “Since Spearman’s hypothesis has been consistently borne out in many independent sets of appropriate data, and no contrary data have been found, it may legitimately claim the status of empirical fact” (Jensen, 1993, p. 48).

Jensen’s claim has not remained undisputed. The study by Helms-Lorenz et al. (2003) may seem to support Spearman’s Hypothesis, as larger cross-cultural differences were reported for the various subtests of the intelligence tests than for the cognitively less complex ECTs. However, these authors also got ratings (by independent experts) of the cultural entrenchment of the tests of the battery. They found that these ratings were better predictors of cross-cultural performance differences than were measures of cognitive complexity. Spearman’s Hypothesis has also been challenged on methodological grounds. Thus, Dolan, Roorda, and Wicherts (2004) reanalyzed data from two previously published studies that reported support for Spearman’s Hypothesis (a South-African study by Lynn & Owen, 1994, and a Dutch study by Te Nijenhuis & Van der Flier, 2003), using multigroup confirmatory factor analysis. Their analyses focused on testing factorial invariance, which amounts to a test of the equality of factor loadings, correlations, and error variances across cultural groups. Factorial invariance, a prerequisite for a test of Spearman’s Hypothesis, was not “remotely tenable” (p. 170). In conclusion, the evidence regarding Spearman’s Hypothesis is inconclusive. Supportive evidence has been challenged on both substantive grounds (i.e., neglect of cultural bias in instruments) and methodological grounds (i.e., neglect of psychometric bias in instruments). The previous discussion indicates that there are at least three explanations of the patterning of cross-cultural score differences at simple and cognitive tests, which are partly competing and partly complementary: the increase of cognitive complexity (Spearman’s Hypothesis), the increase of cultural complexity, and methodological artifacts (e.g., test-wiseness and previous test exposure).

Statistical explanations of cross-cultural differences

A second way to maximize the interpretability of observed cross-cultural performance differences on cognitive tests against rival explanations involves the measurement of variables in the study that could have a bearing on the cross-cultural differences. The idea to include explanatory variables in cross-cultural designs has been called “unpackaging”, because the researcher tries to “unpackage” observed cross-cultural score differences in a target variable by measuring presumably relevant explanatory variables (Whiting, 1976; see also Bond & Van de Vijver, in press; Poortinga & Van de Vijver, 1987). Statistical techniques such as analysis of covariance and regression analysis are used to evaluate to what extent cross-cultural differences in the explanatory variables can statistically account for the observed cross-cultural score differences in the target variable. Cross-cultural differences on simple tests (if present) may be more straightforward to interpret, particularly when alongside the target tests, measures of presumably relevant sources of cross-cultural differences (e.g., a test of memory capacity) are administered.

Longitudinal designs

The third way to bolster the interpretability of cross-cultural score differences implemented in the current study involved the use of a longitudinal design. The patterning of cross-cultural performance differences at retesting could shed light on the nature of these differences. Invariance of these differences across cultures would underscore the
similarity in psychological meaning of cognitive test scores, whereas cross-cultural differences in gain patterns point to differences in meaning. Jensen (1998) reports support for the invariance assumption; he did not find any change in cross-cultural score differences between African Americans and European Americans in choice reaction time tasks with retesting. One of the earliest systematic cross-cultural studies of elementary cognitive processing was carried out by Poortinga (1971). He administered several series of ECTs, consisting of simple, two-, four-, and eight-choice reaction time measures involving both auditory and visual stimuli to 40 white and 40 black students in South Africa. No cross-cultural differences were found for the first two tasks, while the white group showed shorter reaction times on both types of stimuli in the last two tasks; the effect size was very large (1.64 SD for the visual stimuli and 1.32 SD for the auditory stimuli). An analysis of the change of the reaction times across the series of the last two tasks showed that the black students had longer reaction times in the beginning, but their reaction times decreased more rapidly than those of the white students. In an analysis of variance, the interaction between series and culture was significant, which was interpreted by the author as evidence that the tasks were not equivalent. It is interesting to note that Jensen (1998, p. 392) quotes Poortinga’s (1971) study as evidence in favor of the Spearman Hypothesis (i.e., larger cross-cultural differences in the more complex ECTs), whereas Poortinga uses the same data to demonstrate that more complex ECTs tend to be more biased than simpler measures. The differences in results between the studies of Jensen and Poortinga are striking. Although the studies did not use exactly the same kind of research methods and stimuli, both worked in the so-called Hick paradigm (that compares reaction times to simple and choice-reaction stimuli; Hick, 1952), which has shown robust results across various different stimulus procedures. As a consequence, differences in instrumentation are unlikely to explain the differences in results. It could be speculated that the contrasting results found in the studies by Jensen and Poortinga are due to the cultural and socioeconomic distance between the groups, which is larger in South Africa than in the USA.

Two related kinds of explanations can be envisaged for the decreasing cross-cultural performance differences on complex tasks after retesting. The first draws on the literature about skill learning and practice effects in retesting (e.g., Ackerman, Kyllo, & Roberts, 1999; Logan, 1988). According to Ackerman (1986), cognitive tasks are either consistent (data limited) or inconsistent (resource limited). Performance on consistent tasks is limited by the available knowledge, such as the performance on a test of vocabulary knowledge. Repeated test administration of such tests can be expected to quickly show an asymptotic result, because performance is limited by knowledge available in long-term memory (which can be seen as fixed across the repeated administrations), whereas performance on inconsistent tasks is more dependent on available resources (e.g., effort). Simple ECTs can be expected to be more data limited, while complex ECTs can be expected to be more resource limited. The classical studies by Fleishman and Hempel (1954) indicated that in the initial stages of learning of a psychomotor task, intellectual components show stronger relationships with intelligence than motor performance, and this pattern reverses with prolonged training. In the initial stages, tasks could be assumed to be more complex for groups with little previous experience on similar tasks, such as groups with lower levels of education. In this line of reasoning, score gains depend on initial level and are unrelated to cultural factors; individuals from different cultural groups with the same score at the first test administration (or cultural groups with the same initial scores) are expected to show the same gain pattern.

The second explanation of decreasing cross-cultural performance differences on complex tasks after retesting involves method bias. Van de Vijver and Leung (1997) have
coined the terms “instrument bias” and “administration bias” (which are examples of method bias) to refer to all those aspects of an instrument or its administration that lead to cross-cultural score differences. There is cross-cultural evidence supporting this explanation. For example, Van de Vijver, Daal, and Van Zonneveld (1986) asked primary school pupils in Surinam, The Netherlands, and Zambia to solve inductive reasoning items. Randomly selected groups in each country were trained to solve items quite similar to those in the pretest and posttest. Zambian pupils showed the lowest scores at the pretest and the largest increments at the posttest. Interestingly, these increments were found both in the trained (experimental) and untrained (control) group. The authors argued that this increase was caused by improved test-taking skills, learned at the pretest. Nkaya, Huteau, and Bonnet (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. Similarly, Ombrédane, Robaye, and Plumail (1956) have shown that the predictive validity of the Raven test increased after repeated administration in a group of illiterate, Congolese mine workers. Working with primary school children in Tanzania, Sternberg et al. (2002; see also Grigorenko & Sternberg, 1998) used a dynamic testing procedure to assess cognitive skills. Compared to pretest scores on the dynamic tests, posttest scores generally were better predictors of reference ability and achievement measures, indicating that the retesting changed the psychological nature of the task and that the performance at the retest was less influenced by fewer additional skills. Retest effects due to method factors are not restricted to non-Western participants and also prevail among Westerners (e.g., Blieszner, Willis, & Baltes, 1981; Wing, 1980). It is clear that repeated test administrations can yield valuable insights about the nature of cross-cultural score differences at the first occasion.

The present study

The present study examined the nature of cross-cultural differences in performance on ECTs in Zimbabwe and The Netherlands. These countries were chosen because of their differences in culture, affluence, and education. School life expectancy is 9 years in Zimbabwe and 17 years in The Netherlands (United Nations Statistics Division, 2007). Educational expenditure for secondary education was US$ 6,403 in The Netherlands and US$ 1,368 in Zimbabwe in 2001 (both amounts are converted from local currencies using Power Purchasing Parity; Organisation for Economic Co-operation and Development, 2004). A test battery consisting of ECTs of varying complexity was administered to secondary school pupils in a repeated-measures design. In line with previous studies (e.g., Helms-Lorenz et al., 2003; Jensen, 1998; Poortinga, 1971), cross-cultural differences in reaction times are expected to increase with ECT complexity (with the Dutch pupils showing a higher performance) (Hypothesis 1). In line with Poortinga’s (1971) study, it is expected that the repeated test administration reduces the size of the cross-cultural differences on the ECTs (Hypothesis 2). Furthermore, additional instruments to measure cognitive skills that could be relevant for validating observed cross-cultural differences on ECTs (crystallized intelligence, as measured by Vocabulary in the present study, and memory), noncognitive skills (manual dexterity), and socioeconomic status are administered. These measures are used to provide a partial or complete explanation of the observed cross-cultural performance differences. It is expected that these measures can significantly decrease the observed performance differences (Hypothesis 3).
Method

Participants

All pupils (N = 35 in each culture) attended the first grade of secondary school. The Zimbabwean (black) pupils attended a secondary school about 30 km from the capital Harare. The school is state owned; all teachers were qualified. The classrooms had concrete floors; no classrooms had sufficient furniture for all pupils. In The Netherlands, pupils were recruited from a secondary school in a village close to Tilburg and a school in Wageningen. Different types of secondary schools were sampled in The Netherlands so as to get a fair picture of the intellectual heterogeneity of the population. The average age of the Dutch pupils was 12.57 (SD = 0.66), while the average age of the Zimbabwean pupils was 14.80 (SD = 0.80); the difference between the samples is highly significant, t(68) = –12.78, p < .01, d = –3.06. The age difference of about 2 years is due to a combination of two factors: the age of enrolment in primary school differs 1 year (6 years in The Netherlands and 7 years in Zimbabwe), and the Dutch primary school curriculum has six grades and the Zimbabwean seven grades.

Instruments

Noncognitive measures

A biographical questionnaire was administered. Information about the following indicators of socioeconomic status was collected: parental education and profession, number of books in the household (5-point scale: 0–10 books; 11–20 books; 21–50 books; 51–100 books; more than 100 books), and availability of radio and television in the household (both scored dichotomously, sum score of both was used). These indicators were all positively correlated (a first analysis revealed that the first factor explained 46.8% of the variance). The average z score on these variables was used as SES score. The sample differences in SES were highly significant (Zimbabwe: M = –0.56, SD = 0.33, range = –1.07, 0.29; Netherlands: M = 0.56, SD = .39, range: –0.16, 1.36), t(68) = –12.68, d = –3.08).

Cognitive measures

Five ECTs were administered. The first task, Task A, was a simple reaction time measure. The pupils were seated behind a computer monitor and response panel; two buttons visible were visible on this panel (the solid squares on the response panel of Figure 1). The screen was white at the beginning of the trial. After an auditory warning stimulus, the black contour lines of a square became visible. After this warning stimulus, the pupil was asked to press the home button of the response panel. After a fixed period of 2 s and a random period between 0 and 2 s, the square became black. The pupil was instructed to lift his or her index finger from the home button and to press the button at the top of the response panel as soon as the square became black. Reaction time (measured in milliseconds) was defined as the time elapsed between the stimulus onset and the release of the home button of the response panel. Movement time was the time needed to release the home button and press the response button. The intertrial interval was 5 s. The test consisted of two series of 10 trials (same duration of foreperiod and intertrial intervals were used for all tasks).

Task B was a choice reaction time measure that used five stimuli. Six buttons were visible on the response panel (see Figure 1). After an auditory warning signal, the contours of five squares became visible on the monitor. One of the squares became black after the
foreperiod. The squares which became black varied across trials but were identical for all pupils; for example, the lower left square always became black in the first trial. The pupil had to press the corresponding button of the response panel (i.e., the lower left button of the response panel). A similar spatial pattern (of two squares at the left side, one in the upper middle, and two at the right side) was used on the monitor and response device so that pupils would find it easy to determine which button had to be pressed. In order to ensure that pupils identified the target figure before lifting the home button, the squares became gray, and only the borders remained visible, once the home button was lifted. The test consisted of two series of 10 trials. In Task C, the ninth figure of Figure 2 was always used. The other four squares were randomly chosen from the remaining set of 15 figures of Figure 2. The pupil was instructed to press the button corresponding to the position of the ninth figure. In Task D, the stimulus set consisted of two pairs of identical figures and an “odd one out”. The pairs of identical figures and the target figure were chosen randomly; for example, two squares contained the 6th figure of Figure 2, two squares the 9th figure, while the 12th figure was the “odd one out”, which had to be identified. The pupil had to press the button of this latter figure.

Task E used the concept of “complementary squares”; these are squares that form one black square (without overlap) when combined. The eight rows of Figure 2 make up eight complementary sets of two figures. Two such sets, together with a randomly chosen other figure, were displayed in the five squares (e.g., the Figures 5 and 6 as the first pair, 11 and 12 as the second pair, and 16 as the “odd one out”). The pupil had to identify the “odd one out” and press the corresponding button. The last three tasks consisted of two series of 12 items.

Incorrect responses were treated as missing values in the data. In an analysis that is not further documented here, the proportion of errors was found to be small and similar for Zimbabwean and Dutch pupils.
The Dutch version of the Vocabulary subtest of the third version of the Wechsler Intelligence Scale for Children was administered (Vander Steene et al., 1991). In Zimbabwe, the English-language version was applied. The Digit Span Forward and Backward from the same test were also administered. This test is identical for the two languages, whereas the Vocabulary subtests use entirely different stimuli in the two languages, precluding a direct comparison.

Finally, a tapping task was used, in which the pupil had to press the button of a hand counter as often as possible in 30 s. After a brief break, the task was repeated.

**Procedure**

All instruments were administered by the (Dutch) author. The ECTs were administered on 5 subsequent school days. The 1st day was used to get the pupils acquainted with the tests and the testing procedure; data presented here are based on the last 4 days. The order of the ECTs was fixed on the 1st day (from Task A to Task E in the order in which they were described above); on the last day, the reverse order was used. The order was randomized on the other days. The testing language in Zimbabwe was English (the language of instruction), which was the second language of the pupils; their first language was Shona.
Results
The psychometric properties of the instruments are presented first, followed by a cross-cultural comparison of the gain patterns of the reaction time on the ECTs. The final analysis addressed the question to what extent tapping, memory, and Vocabulary scores could explain the cross-cultural differences observed.

Reliabilities
The internal consistencies of the five ECTs, given in Table 1, were high across all tasks and days of administration (with an overall median of 0.93) and were similar across the two countries. The correlations of the Forward and Backward Digit Span scores were 0.62 in Zimbabwe and 0.71 in The Netherlands (both $p < .01$). The tapping scores of the first series and second series correlated 0.85 in The Netherlands and 0.88 in The Netherlands (both $p < .01$). These correlations pointed to a fair reliability. For Vocabulary, no estimate of the internal consistency was available.

Gain patterns
The ECTs were administered on 4 consecutive days, and on each day, two series were administered, making a total of eight series. The cultural group means of the series are presented in Figure 3. Analysis of variance (repeated measures) was used to examine the size of the cross-cultural differences and of the decrements of the RTs; per task an analysis was carried out. The effect sizes (proportion of variance accounted for) have been presented in Table 2. The size of the main effect of culture increased with the complexity of the test (ranging from a nonsignificant value of 0.03 for Task A to a highly significant value of 0.45 for Task E). These findings supported the first hypothesis. The mean reaction times of the Dutch pupils were consistently shorter than those of the Zimbabwean pupils. The effect of series was significant for all tasks; reaction times decreased with practice for all tasks. The size of the effect increased with task complexity, going from 0.04 for Task A to 0.33 for Task E; more complex tasks showed more performance increments. The interaction of culture and series also increased with complexity; yet, its size was never larger than 0.06 (a medium effect size according to Cohen).

Table 1. Internal consistencies of reaction time tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Culture</th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task A</td>
<td>Zimbabwean</td>
<td>0.96</td>
<td>0.93</td>
<td>0.91</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>0.96</td>
<td>0.94</td>
<td>0.94</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>Task B</td>
<td>Zimbabwean</td>
<td>0.96</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>0.96</td>
<td>0.95</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>Task C</td>
<td>Zimbabwean</td>
<td>0.93</td>
<td>0.93</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>0.91</td>
<td>0.90</td>
<td>0.88</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Task D</td>
<td>Zimbabwean</td>
<td>0.92</td>
<td>0.91</td>
<td>0.92</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>0.93</td>
<td>0.89</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Task E</td>
<td>Zimbabwean</td>
<td>0.91</td>
<td>0.91</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>0.86</td>
<td>0.80</td>
<td>0.87</td>
<td>0.90</td>
<td>0.87</td>
</tr>
<tr>
<td>Median</td>
<td>Zimbabwean</td>
<td>0.93</td>
<td>0.93</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>0.93</td>
<td>0.90</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Figure 3. Mean reaction times per task and series (in ms).
Table 2. Effect sizes (proportion of variance accounted for) of culture, series, and their interaction.

<table>
<thead>
<tr>
<th>Task</th>
<th>Source</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Culture</td>
<td>Series*</td>
<td>Culture × Series*</td>
</tr>
<tr>
<td>Task A</td>
<td>0.03</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Task B</td>
<td>0.03</td>
<td>0.06**</td>
<td>0.03</td>
</tr>
<tr>
<td>Task C</td>
<td>0.25**</td>
<td>0.29**</td>
<td>0.06**</td>
</tr>
<tr>
<td>Task D</td>
<td>0.44**</td>
<td>0.26**</td>
<td>0.06**</td>
</tr>
<tr>
<td>Task E</td>
<td>0.45**</td>
<td>0.33**</td>
<td>0.05**</td>
</tr>
</tbody>
</table>

*Significance is based on the Greenhouse-Geisser corrected number of degrees of freedom. *p < .05. **p < .01.

Several mathematical models can be used to model RT decrements in a repeated administration paradigm, such as a power function (Anderson, 2001; Kirsner & Speelman, 1996), differential equations (Smith, 2000), and logarithmic functions (Kail, 1988). The current study employed a power function, which postulates that an observed reaction time, $RT$, can be written as $RT = at^b$, in which $a$ represents the reaction time at time 1, which is minimum or maximum value of the curve (as reaction times tend to decrease with practice, $a$ will usually represent the highest value), $t$ is the measurement occasion (here ranging from one to eight), and $b$ indicates the degree of increase (in case $b$ is positive) or decrease (if $b$ is negative). A common representation of the power law is found by taking the natural logarithm of the model equation, which gives: $\ln(RT) = \ln(a) + b \ln(t)$. This model states that the logarithm of the reaction time is a linear function of the logarithm of the measurement occasion.

Data were analyzed both at individual and country level. The former amounts to fitting the power law to the data of each participant separately, while in the latter a single function is fitted through the country means of the reaction times. A graphical rendering of the country-level data is presented in Figure 4. As can be seen there, the more complex tasks showed gain patterns that were more in line with the power law than did the less complex tasks. This pattern can be confirmed in Table 3, in which the parameters of the power function are given for both samples. Both the country- and individual-level regression parameters are given; the former express the relationship between country means and measurement occasion, while the latter represent the average of the 35 regression functions per country, fitted on the individual-level data. The intercepts and regression coefficients were similar at both levels, the squared multiple correlations were higher at country level (as could be expected on the basis of aggregation effects). The latter were higher for the more complex tasks.

The cross-cultural differences in regression function showed an interesting pattern. The last column of Table 3 presents effect size differences in the individual-level parameters of the regression function (Cohen’s $d$). The size of the intercept differences increased with task complexity. The intercept of the Dutch sample was smaller than the intercept of the Zimbabwean sample, and this effect became more pronounced with task complexity. This pattern of increasing cross-cultural differences with cognitive complexity resembles Spearman’s hypothesis. However, the regression coefficients revealed a different picture. With the exception of Task A, all reaction time tasks were negative, indicating faster learning by the Zimbabwean pupils. The average effect size is $-0.46$, which is almost a medium effect size. So, although the difference was significant only for Task C ($d = -0.56, t(68) = -2.23, p < .05$), all choice reaction time tasks showed a pattern of larger intercepts and faster learning for Zimbabwean pupils than for Dutch pupils.
Figure 4. Observed (solid curve) and expected (dotted curve) country differences in reaction time per task and series (in ms).

Note: A power law is fitted to the differences, using series as predictor. The equation in the figure presents the parameters of the regression function ($RT$ represents the country mean of the reaction time, $t$ refers to the series). All regression parameters are significant ($p < .01$), with the exception of the parameters of Task A.

For each task, the size of the cross-cultural differences in mean RT is plotted against the series in Figure 4. The differences in mean RTs increased with task complexity and decreased with session (the latter confirms the second hypothesis). Remarkably, the
observed differences fitted the power law well (see Figure 4). With the exception of data of Task A which did not follow a power function, all tasks showed high and significant proportions of explained variance, ranging from 0.60 for Task C to 0.95 for Task D (all ps < .01). It can be concluded that the size of the cross-cultural differences in the choice reaction time tasks were well patterned across the eight series, following a power function.

**Explanation of cross-cultural performance differences**

A final analysis determined to what extent individual differences in vocabulary, digit span, and tapping could explain the cross-cultural differences in the ECTs. An overview of the cross-cultural differences in the predictors is given in Table 4. No test of the cross-cultural differences in vocabulary is presented here, as scores were standardized per country. As can be seen in Table 4, tapping did not show a significant difference (d = 0.25), while the differences for digit span and socioeconomic status were significant and large (d = −1.42 and −3.08, respectively). The Dutch pupils showed, on average, a longer digit span and a higher socioeconomic status.

In the regression analyses used to explore the nature of the cross-cultural differences in the reaction time measures, the independent variables were vocabulary, digit span, tapping, and socioeconomic status. The dependent variables were parameters of the regression function (regression coefficient, intercept, and proportion of explained variance) fitted to the eight series. The results are presented in Table 5. Effect sizes of the cross-cultural differences are also given there. The first number in the table (0.28) indicates that the regression coefficients of the individual-level regressions showed an average effect size.
of 0.28 (a positive sign indicates here that the Zimbabwean mean is higher), which was not significant. The number next to it (–0.02) gives the (nonsignificant) effect size which remained after the four independent variables were regressed on the regression coefficient of Task A. Various conclusions can be drawn from Table 5. First, regression coefficients and (squared) multiple regression coefficients of the power law were poorly predicted, while intercept differences were more strongly related to the predictors. Second, differences in intercept were larger than differences in regression coefficients and multiple correlations. Third, no effect sizes remained significant after correcting for the predictors (which supports the third hypothesis). Even the relatively large effect sizes of the intercepts disappeared altogether after correcting for the predictors. Socioeconomic status and digit span were effective predictors of intercept differences. For the most complex tests (Task D and E), both the digit span and socioeconomic status were significant predictors. The nature of these findings is explored in the next section.

Table 4. Means and standard deviations of the predictors in Zimbabwe and The Netherlands and test of significance of their differences.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Zimbabwe</th>
<th>The Netherlands</th>
<th>t (df = 68)</th>
<th>Cohen's d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Digit Span</td>
<td>4.13</td>
<td>0.52</td>
<td>5.11</td>
<td>0.83</td>
</tr>
<tr>
<td>Tapping</td>
<td>157.44</td>
<td>13.82</td>
<td>152.97</td>
<td>21.45</td>
</tr>
<tr>
<td>Socioeconomic status</td>
<td>−0.56</td>
<td>0.33</td>
<td>0.56</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Table 5. Effect sizes (Cohen’s d) of cross-cultural differences in parameters of power regression function before and after regressing background variables on these parameters and results of the regression analyses.

<table>
<thead>
<tr>
<th>Regression parameters</th>
<th>Cohen’s d</th>
<th>Regression parameters</th>
<th>Effect size before</th>
<th>Effect size after</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vocabulary</td>
<td>Digit Span</td>
</tr>
<tr>
<td>Regression coefficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task A</td>
<td>0.28</td>
<td>−0.02</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td>Task B</td>
<td>−0.43</td>
<td>−0.12</td>
<td>−0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>Task C</td>
<td>−0.56*</td>
<td>0.02</td>
<td>−0.21</td>
<td>−0.08</td>
</tr>
<tr>
<td>Task D</td>
<td>−0.47</td>
<td>−0.01</td>
<td>−0.24*</td>
<td>0.40**</td>
</tr>
<tr>
<td>Task E</td>
<td>−0.39</td>
<td>0.03</td>
<td>−0.12</td>
<td>0.26</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task A</td>
<td>0.14</td>
<td>0.09</td>
<td>−0.22</td>
<td>−0.19</td>
</tr>
<tr>
<td>Task B</td>
<td>0.53*</td>
<td>0.16</td>
<td>−0.19</td>
<td>−0.23</td>
</tr>
<tr>
<td>Task C</td>
<td>1.21**</td>
<td>0.11</td>
<td>0.01</td>
<td>−0.15</td>
</tr>
<tr>
<td>Task D</td>
<td>1.21**</td>
<td>0.11</td>
<td>0.08</td>
<td>−0.38**</td>
</tr>
<tr>
<td>Task E</td>
<td>1.58**</td>
<td>0.24</td>
<td>−0.11</td>
<td>−0.27*</td>
</tr>
<tr>
<td>Explained variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task A</td>
<td>−0.11</td>
<td>0.02</td>
<td>−0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>Task B</td>
<td>0.34</td>
<td>0.00</td>
<td>0.01</td>
<td>−0.12</td>
</tr>
<tr>
<td>Task C</td>
<td>0.43</td>
<td>−0.01</td>
<td>0.18</td>
<td>0.00</td>
</tr>
<tr>
<td>Task D</td>
<td>0.13</td>
<td>−0.03</td>
<td>0.14</td>
<td>−0.36*</td>
</tr>
<tr>
<td>Task E</td>
<td>0.52*</td>
<td>0.02</td>
<td>0.08</td>
<td>−0.27</td>
</tr>
</tbody>
</table>

*p < 0.05. **p < 0.01.
Discussion and conclusions

In order to explore the poorly understood nature of cross-cultural differences on tests of basic cognitive functions, five reaction time tests of increasing complexity were administered to 35 secondary school pupils in Zimbabwe and The Netherlands on 5 consecutive school days. The simple tests did not show cross-cultural differences, whereas Dutch pupils outperformed Zimbabwean pupils on the more complex tests (effect sizes of the latter tests were large). Zimbabwean pupils showed larger performance increments at retesting. Two WISC subtests (Vocabulary and Digit Span Forward and Backward), a tapping test, and parental socioeconomic status were used to explain the cross-cultural differences in ECT performance. It was found that these differences could be explained by digit span and socioeconomic status.

The finding that socioeconomic status and digit span are good predictors of cross-cultural performance differences on simple reaction time measures requires further scrutiny. Let us first consider cross-cultural differences in digit span. The size of the cross-cultural differences in digit span might have been inflated by the use of English as testing language in Zimbabwe; after all, it is uncommon to find large cross-cultural differences in short-term memory (Wagner, 1981). The reported cross-cultural differences in short-term memory appear to follow Baddeley’s (1997) phonological loop model. In this model, short-term memory span is operationally defined as the number of items (usually digits) an individual can read aloud in 1.5 s. Speakers of languages with longer words for digits (such as Welsh) tend to have a shorter span than speakers of languages with shorter words for digits (cf. Shebani, Van de Vijver, & Poortinga, 2005, in press). However, words in English and Dutch, which are both Germanic languages, are of about the same length, which makes significant differences in span of native speakers of these languages unlikely. However, the Zimbabwean participants were not native speakers and could be assumed to have more problems with the quick internal processing of English-language numerals.

Socioeconomic status is a proxy variable that should be “unpackaged”. The substantive interpretation of socioeconomic status is compromised by the huge differences in scores across the two countries in its constituent variables. Socioeconomic status showed relatively little within-country and much cross-country variation. As a consequence, any variable with a similar pattern of small within- and large cross-country variation can statistically explain observed cross-cultural differences in reaction times. A good example of such a variable is testing language. Dutch pupils were assessed in their first language, whereas Zimbabwean pupils were assessed in English, which usually is their second or third language. Also, although extensive training preceded the actual testing session, it is impossible to rule out the possibility that the performance of the Zimbabwean pupils would have been better if they were assessed by a Zimbabwean tester or had a longer training session. Finally, it was impossible to match the samples on both educational and chronological age. It could well be that age difference also influenced the test results. It can be concluded that the success of socioeconomic status in the explanation of cross-cultural performance differences is an important first step in their interpretation; however, more research is needed to obtain a better understanding of the psychological meaning of socioeconomic status.

The importance of non-target performance moderators is confirmed in the analysis of gain patterns. Compared to Dutch pupils, Zimbabwean pupils showed both longer reaction times in the beginning (in line with Spearman’s Hypothesis). However, the Zimbabwean pupils also showed more gain at retesting (which is hard to reconcile with Spearman’s hypothesis). If the Zimbabwean pupils would have a lower mental speed, there
would be no reason to expect that they are faster learners. The most likely and parsimonious explanation of our findings is that the cross-cultural differences observed are influenced by various unintended factors (such as test-wiseness and previous exposure to similar tasks). The question as to whether Dutch or Zimbabwean pupils show higher levels of skill in elementary cognitive skills apparently requires a complex answer. At surface level, there are no differences for simple tests, and as tests become more complex, the Dutch increasingly outperform the Zimbabwean pupils (thereby confirming Spearman’s Hypothesis). However, if elementary cognitive skills are defined in terms of learning, the Zimbabwean pupils show a superior performance to the Dutch pupils, thereby disconfirming Spearman’s Hypothesis. Clearly, the question of which pupils show a higher skill level defies a simple answer. ECTs measure more than elementary cognitive skills; the influence of these supplementary skills increases with task complexity. The Zimbabwean pupils seem to pick up some of the latter skills during the first test administrations. As a consequence, the ECTs are increasingly comparable across test occasions. The size of the cross-cultural differences decreases with test administration. So, the more comparable tests are, the smaller the differences they show.

A straightforward mathematical extrapolation of the curves of the complex tasks of Figure 4 suggests that the curves will eventually reach a value of zero; yet, no curve of Figure 4 came close to its asymptotic value of zero. The question can be asked after how many trials the performance differences would have become negligible. In theory, the power law allows for its prediction. For example, for Task E, the difference is predicted to become zero after about 1,000 sessions. However, the extrapolation has to make the dubious assumption of the long-range applicability of the power law; yet, it is clear that the superiority of the performance of the Dutch pupils can be expected to remain for at least some time with continued administration.

For the ECTs as studied here, the support of Spearman’s Hypothesis is more seeming than real, as it is partly based on cross-cultural differences that are essentially unrelated to the intellectual processes (general intelligence) which are supposed to underlie the differences. Our study replicated and extended findings by Poortinga (1971) on adults. He reported larger cross-cultural differences and more gain on more complex tasks. Jensen (1998) interpreted Poortinga’s study as supporting Spearman’s Hypothesis, whereas Poortinga favored an interpretation in terms of biasing factors. The current study would be more supportive of the latter position, because the cross-cultural differences could be “explained away” by method-related (biasing) factors and the evidence that an increase of test validity is accompanied by a decrease in cross-cultural performance differences. The bias interpretation is also in line with the study by Helms-Lorenz et al. (2003), in which knowledge with the Dutch language and culture was found to “explain away” performance differences on ECTs of mainstream and immigrant Dutch children.

The current study has implications for cross-cultural cognitive assessment. It was found that even for tests which, in comparison with complex cognitive instruments, are supposed to show less susceptibility to the influence of ambient cultural factors, cross-cultural performance differences cannot be interpreted in a straightforward manner. Assessment-related factors such as previous exposure are assumed to influence the cross-cultural differences observed. The influence of these factors will certainly not be less for culturally more entrenched tests. Caution is required in the interpretation of these differences if no estimate is available of the size of impact of the performance-moderating factors. Repeated test administrations as applied here, the usage of measures of test exposure (Sarnacki, 1979), the estimation of the cultural entrenchment of tests (Helms-Lorenz et al., 2003), and dynamic testing procedures (e.g., Sternberg et al., 2002) are
examples of procedures which can be used to increase the validity of instruments and, hence, the validity of the cross-cultural inferences. All these procedures point to the need to be cautious in drawing inferences about single applications of cognitive tests in culturally heterogeneous contexts. It is only through a further exploration of the nature of the cross-cultural differences that we can deepen our insight in culture–cognition relationships. Cross-cultural assessment procedures employed to gain insight into cross-cultural performance differences introduce new sources of cross-cultural differences. This is in itself paradoxical.

Note on contributor
Fons J.R. van de Vijver is professor of Cross-Cultural Psychology at Tilburg University, Tilburg, The Netherlands, and North-West University, Potchefstroom, South Africa. He has over 250 publications on cross-cultural topics such as research methods, acculturation, and intelligence.

References


