European Journal of Psychological Assessment, Vol. 20, Issue 1, pp. 10–26

Does Cultural Background Influence the Intellectual Performance of Children from Immigrant Groups?

The RAKIT Intelligence Test for Immigrant Children

Jan te Nijenhuis[1], Elsbeth Tolboom[1], Wilma Resing[2], and Nico Bleichrodt[3]

[1]Work and Organizational Psychology, University of Amsterdam
[2]Developmental and Educational Psychology, University of Leiden
[3]Work and Organizational Psychology, Vrije Universiteit, Amsterdam, all The Netherlands

Keywords: Intelligence, testing, bias, immigrants, children

Summary: This paper addresses both the construct validity and the criterion-related validity of the "Revisie Amsterdamse Kinder Intelligentie Test" (RAKIT), which is a cognitive ability test developed for primary school children. The present study compared immigrant primary school children (N = 559) and Dutch children (N = 604). The mean scores of Surinamese/Netherlands Antillean, Moroccan, and Turkish children differed from each other and were lower than those of the Dutch children. Comparison of the test dimensions showed that group differences with respect to the construct validity were small. We found some item bias, but the combined effects on the sum score were not large. The estimate of general intelligence *(g)* as computed with the RAKIT showed strong predictive validity for most school subjects and standardized achievement tests. Although some criteria revealed significant prediction bias, the effects were very small. Most of the analyses we performed on differences in test scores and differences in criterion scores supported Spearman's hypothesis that g is the predominant factor determining the size of the differences between two groups. The conclusion that the RAKIT can be used for the assessment of groups from various backgrounds seems warranted.

Introduction

There has been an increasing interest in research questions about the influence of cultural background on the nature and development of cognitive abilities. It will be clear that research on these – assumed – cultural influences on cognitive performance, often operationalized by cognitive tests, should not only be restricted to cross-cultural comparisons between populations from different countries (for instance Western versus East-Asian countries) but should also look at cultural differences within one and the same country, especially after periods of substantial immigration.

Intelligence is a very important factor within any educational environment. It is supposed to tell something about one's cognitive abilities, such as problem solving (Resing & Hessels, 2001). Research on the predictive validity of intelligence test scores provides evidence that general intelligence, or g, can be seen as the predominant single predictor of future success: Scores on intelligence tests are fairly good predictors of educational and occupational achievement. The highest validity coefficients reported in schools vary around .50 (e.g., Jensen, 1998;

EJPA 20 (1), © 2004 Hogrefe & Huber Publishers
DOI: 10.1027//1015-5759.20.1.10

Neisser et al., 1996; Snow & Yalow, 1982). Based on various meta-analyses, Schmidt and Hunter (1998) estimated the mean validity coefficient between measures for cognitive ability and occupational success at $r = .51$. The highest validity coefficients for intelligence test scores and scholastic achievement (.50–.65) are found in primary schools. These values drop slightly in secondary and higher education, mainly because the groups of pupils become more and more homogeneous. Jensen (1980) stated that "Children with higher IQs generally acquire more scholastic knowledge more quickly and easily, get better marks, like school better, and stay in school longer" (p. 317).

It is therefore very disturbing that minority groups in the United States such as blacks and Mexican-Americans, in general score substantially lower on standardized intelligence tests than do whites. This may be due to test- or item-bias against ethnic groups, or it may reflect a lower mean level of cognitive abilities. Jensen (1980), Hunter, Schmidt, and Hunter (1979), and Schmidt, Pearlman, and Hunter (1980) concluded that most of the widely used standardized tests of mental ability – IQ, scholastic aptitude, and achievement tests –, are not biased against the native-born English-speaking minority groups on which the amount of evidence is sufficient for an objective determination of bias, if tests were, in fact, biased. More specifically, Jensen (1980) came to the conclusion that ". . . with a good choice of predictors . . . it is possible to predict elementary grades with considerable validity (about .70) for white, black, and Mexican-American children without having to take their ethnicity into account" (p. 474). These group differences in mean intelligence can have a great impact, starting as early as elementary school, and are a disadvantage for minority groups.

However, one cannot simply generalize these findings to the Dutch situation. The conclusions drawn by Jensen (1980), Hunter et al. (1979), and Schmidt et al. (1980) are based on immigrant groups who have grown up with the English language and who are familiar with the dominant culture. A large part of the Dutch ethnic groups, often immigrants from Turkey and Morocco, are unfamiliar with the Dutch language and have grown up in a traditional culture quite dissimilar from the Dutch culture. Children from these immigrant groups come to primary school at the age of 4 or 5, with considerably less preschool education than most of the Dutch children. Other immigrant groups come from the overseas colonies (Netherlands Antilles) and former colonies (Surinam). The majority have a good command of the Dutch language, but have grown up in their own specific cultural environment.

The number of children with an ethnic background in Dutch elementary schools has increased tremendously over the last twenty-five years. In 1996, the percentages for the three largest Dutch cities were: Amsterdam (51.8), Rotterdam (48.9), The Hague (38.1), and Utrecht (37.1). At present this percentage is still increasing (Verweij, Latuheru, Rodenburg, & Wijers, 1998). The percentage of immigrant group participation in Dutch high schools, in general, increased from 3.7 to 7.3 between 1992 and 1996 (Dutch Ministry of Education, Culture, and Science, 1997). Members of the immigrant groups in the Netherlands are less successful in schools compared to members of the Dutch group (e.g., Penninx, 1988; Roelandt & Veenman, 1988; de Jong, 1985, 1987; Lathuheru & Hessels, 1994; van Langen & Jungbluth, 1990). This situation signals the need for good and careful assessment of cognitive abilities in the schools.

As intelligence tests generally show high construct- and criterion-related validity, they are a legitimate instrument to assess the cognitive abilities of children both in the Netherlands and in other parts of the Western world. Within the educational sector, intelligence measures are considered the best single predictors of future success, although their predictive validity certainly is not perfect: Other factors, such as motivation and concentration, have their own specific, but small, predictive value. Despite the importance of the IQ tests in the schools, however, few studies have focused on the validity of these tests for children from ethnic groups. Therefore, an examination of this validity is strongly called for. It is of great importance that the available intelligence tests are equally valid for both minority groups and Dutch groups since they are at the very basis of important decisions for one's educational and professional career.

Te Nijenhuis and van der Flier (1997, 1999) found that cognitive test scores of adult immigrant job applicants are only slightly biased. Te Nijenhuis, Evers, and Mur (2000) concluded in their evaluation of the Dutch adapted version of the Differential Aptitude Test (Evers & Lucassen, 1992), an intelligence test for children aged 12 years and older, that the group differences in test scores found between the majority group and the immigrant group are for the greater part not caused by test bias. De Jong and van Batenburg (1984) showed that IQ test scores of immigrant children from primary schools are lower than those of their Dutch classmates, but they reported no evidence of bias. Although these are promising results from the perspective of professional test users, further research into the validity of IQ tests for Dutch immigrants is essential, because so much is at stake.

Research Question

The central question of this study is whether the standardized cognitive ability test RAKIT, the Revised Amsterdam Intelligence Test for Children (Bleichrodt, Drenth, Resing, & Zaal, 1984), has the same construct and predictive validity for immigrant children as for groups of Dutch elementary school children. Messick (1989) regarded validity as “an integrated judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p. 13). When researching the differences in test scores between immigrants and the Dutch, the question is whether the same conclusions can be drawn from the same test scores for members of both groups, or in other words, whether the validity of the test is the same for the two groups. The comparability of scores on intelligence tests depends on the answers to two different questions. The first question is whether immigrants and Dutch with the same test scores have the same level and pattern of cognitive functioning. The second question is whether members of both groups with the same test scores have the same chance of showing specific criterion behavior in the future. These two questions coincide with the traditional classification into construct validity and criterion-related validity. The question of construct validity consists of whether the same dimensions are measured, in the same measurement units, and at the same level. The measurement of the same dimensions in different groups will be confirmed if the test is connected with relevant constructs in a comparable fashion for the different groups. For this purpose, outcomes of analyses based on correlational techniques, like correlations and factor matrices, are often compared.

Similarity of measurement units in different groups presupposes that the same dimensions are being measured in the different groups. The question of comparability of measurement units can be answered in a positive way if a regression analysis of criterion variables on the test scores shows coinciding regression lines. To answer the question of construct validity on the item level, a large number of techniques for the analysis of biased items have been developed. Examples are techniques that are based on latent trait models and techniques where conditioning takes place on the total score. If a scale measures the same dimensions in an insufficient fashion, it is of little practical significance to carry out certain item bias analyses, since the starting assumption of these analyses is that the scale score is a good estimate for the position on the latent trait. If a test measures, to a high degree, the same dimensions at the level of the scale, but does not measure these dimensions in comparable measurement units, it can be meaningful to look at whether the same dimensions are being measured at the item level.

When answering the question of construct validity and criterion-related validity, use is partially being made of the same evidence, namely the relations between the test and the measurements of the criterion. The answer to the question of construct validity consists of a number of research analyses in which the relations with criteria that are representative of the nomological network are being examined. The answer to the question of criterion-related validity can be limited to the connection between a test and a limited number of criteria. It should be noted that in practice a statement about the equality of measurement units may be limited to one criterion, so that only nonvarying relations between the criterion and the scale score have to be demonstrated.

Method

Research Participants

Half of the data used in this study are taken from a group of Dutch children who constituted the norm sample of the RAKIT (Bleichrodt et al., 1984). This group is representative of the Dutch population. The other half are taken from the largest groups of immigrant children in the Netherlands: children with a Moroccan, Turkish, or Surinamese or Netherlands Antillean background. These children were selected from primary schools with a relatively high percentage of immigrant children according to the Dutch Central Bureau for Statistics. They are from both small and large cities, which are spread all over the country, in areas of varying degrees of urbanization. The ethnic sample is, due to its careful sampling and its large numbers, a good approximation of a representative sample of immigrant children in the Netherlands. Table 1 shows the distribution of the groups in terms of age and sample size.

The immigrant children had to meet the following criteria: Both parents were born in the country of their nationality; they are residents of the Netherlands and have attended Dutch schools for at least six months; they can understand the instructions and the meaning of the various subtests of the RAKIT; their age is respectively between 5.6 and 5.10, or 7.6 and 7.10, or 9.6 and 9.10 years. Apart from test scores on the RAKIT and scores on var-

Table 1. Composition of research participants.

			Group		
Age	Dutch	Turks	Moroccans	Sur./Neth.Ant	Total
5.8	204	61	62	60	387
7.8	196	71	60	61	388
9.8	204	62	60	62	388
Total	604	194	182	183	1163

Note. Sur./Neth.Ant. = Surinamese/Netherlands Antilleans
All data were analyzed, but only the results for the 7-year-olds are reported. The interested reader is referred to the first author, who will gladly supply all results upon written request.

Table 2. Composition research participants by length of residence in the Netherlands in percentages.

Number of years in the Netherlands	S/A	M	T
0–3	7	21	38
3–5	11	17	17
>5, not born in Neth.	11	10	8
>5, born in Neth.	31	31	28
Not known	39	21	8
Total	99	100	99

Note. S/A = Surinamese/Netherlands Antilleans; T = Turks; M = Moroccans; Neth. = The Netherlands.

ious criteria, which will be discussed later, data on sex and length of stay in the Netherlands were collected (see Table 2). All data were analyzed, but only the results for the seven-year-olds are reported. The first author will gladly supply all results upon written request.

Test

The RAKIT (Bleichrodt et al., 1984) is a carefully and well- constructed intelligence test for children aged four to eleven. It is based upon Thurstone's primary factor theory, and also contains tests for associative memory and complex visual-motor performance. The RAKIT has 12 highly differentiated subtests. The construction of the test is based on the assumption that a combination score of different aspects of the intelligence domain gives a good indication of the child's general cognitive ability level (Bleichrodt, Resing, Drenth, & Zaal 1987). The twelve subtests of the RAKIT are administered individually and do not involve writing or reading. The Dutch Committee on Test Evaluation (COTAN; Evers, Van Vliet-Mulder, & Ter Laak, 1992) gave the RAKIT the maximum judgment on all categories, including norms, reliability, and validity, whereby it becomes one of the top three cognitive ability tests in the Netherlands.

Carroll's (1993) hierarchical three-stratum model is a widely accepted intelligence model. It includes three levels of intelligence, the highest being general intelligence (*g)* (Stratum III). At the second level are the broad abilities Fluid Intelligence (G_f), Crystallized Intelligence ($G_{cr)}$, General Memory and Learning(G_m), Broad Visual Perception (G_v), Broad Auditory Perception, Broad Retrieval Ability, and Broad Cognitive Speediness (Stratum II). At the third and lowest level (Stratum I) are narrow abilities such as Sequential Reasoning, Spelling Ability, and Visualization. Although the RAKIT is not developed along the lines of this hierarchical factor model, one can describe the subtests of the RAKIT in terms of the three-stratum model of intelligence outlined above.

The Subtests of the RAKIT (Bleichrodt, Drenth, Zaal, & Resing, 1987)

1. *Closure:* The child is shown very incomplete pictures and has to figure out the complete picture. According to Carroll's taxonomy this subtest is a measure of Closure Speed at Stratum I, which makes this subtest a measure of Broad Visual Perception at Stratum II.
2. *Exclusion:* Out of four abstract figures, the child selects the one that is different from the other three. The child has to detect the necessary rule to solve the task. This subtest measures Induction at Stratum I, which makes it a measure of Fluid Intelligence at Stratum II.
3. *Memory Span:* The child has to memorize figures put on cards and the sequence in which they are presented. After five seconds the card is turned and the child reproduces the figures, in the right sequence, using blocks on which the figures are printed. The subtest contains a series with concrete figures and a series with abstract figures. Both series measure (Visual) Memory Span at Stratum I. Both series fall under General Memory and Learning at Stratum II.
4. *Verbal Meaning:* Words are presented to the child in an auditory fashion and from four figures the child chooses the one which resembles the word just heard. This subtest measures Lexical Knowledge at Stratum I and is a measure of Crystallized Intelligence at Stratum II.
5. *Mazes:* The child has to go through a maze with a stick as fast as possible. Because of the speed factor this subtest is a measure of Spatial Scanning at Stratum I, which falls under Broad Visual Perception at Stratum II.
6. *Analogies:* The child has to complete verbal analogies that are stated as follows: A: B is like C: . . . (there are four options to choose from). The constructors of this subtest tried to avoid measuring Lexical Knowledge, by including only those words that are very frequently used in ordinary life. All words in the analogy items are accompanied by illustrations, so as to reduce the verbal aspect of the task to a minimum. This subtest is a measure of Induction at Stratum I, which makes it a measure of Fluid Intelligence at Stratum II.
7. *Quantity:* In this multiple-choice test, the child has to make comparisons between pictures of objects differing in volume, length, weight, and surface. This subtest is a measure of Quantitative Reasoning at Stratum I, which measures Fluid Intelligence at Stratum II.
8. *Disks:* The child has to put disks with two, three, or

four holes on a board with pins as fast as possible until three layers of disks are on the board. This subtest is a measure of Spatial Relations at Stratum I, which measures Broad Visual Perception at Stratum II.

9. *Learning Names:* The child has to memorize the names of different butterflies and cats using pictures presented on cardboard. This subtest measures Associative Memory at Stratum I, which makes it a measure of General Memory and Learning at Stratum II.
10. *Hidden Figures:* The child has to discover which out of six figures is hidden in a complex drawing. This subtest is a measure of Flexibility of Closure at Stratum I, which makes it a measure of Broad Visual Perception at Stratum II.
11. *Idea Production:* The child has to name as many words, objects, or situations as possible that can be associated with a broad category within a certain time span, for example: "What can you eat?" This subtest is a measure of Ideational Fluency at Stratum I, which is a measure of Broad Retrieval Ability at Stratum II.
12. *Storytelling:* The child has to tell as much as possible about a picture on a board and what could happen to the persons or objects in the picture. The total score of the child is composed of both quantitative measures (number of words, number of relations, did or didn't develop a plot, etc.) and qualitative measures (did the child grasp the central meaning of the story). This subtest consists of different elements and measures at Stratum I: Naming Facility and Ideational Fluency, Sequential Reasoning, and, to some extent, Communication Ability. These Stratum I abilities are respectively measures of Broad Retrieval Ability, Fluid Intelligence, and Crystallized Intelligence at Stratum II.

All tests were administered according to the conditions prescribed in the manual, such as the use of standard termination rules (Bleichrodt, Drenth, et al., 1987).

Adaptations of the RAKIT are being used in countries such as Indonesia (Test Intelligensi Anak; Drenth, Bleichrodt, Setiono, & Poespadibrata, 1975), East-Africa (African Child Intelligence Test; Drenth, Van der Flier, Muinde, Otaala, Omari, & Opolot, 1980) and India (Indian Child Intelligence Test; Bleichrodt, Hoksbergen, Athavale, Kher, & Khire, 1991). Cross-cultural research has shown that the meaning of the tests is highly comparable for the various countries (Bleichrodt, Hoksbergen, Khire, & Dekker, 1998; Bleichrodt, Hoksbergen, & Khire, 1999).

The WISC-R is one of the most often used intelligence tests in the world and is considered a solid, classical test, that gives a good estimate of general intelligence. Just as the RAKIT, it consists of a large, highly differentiated collection of 12 subtests, broadly covering the intelligence domain. Bleichrodt, Resing, et al. (1987) conducted a study in which 469 Dutch children, aged 6.6–9.10, half of whom were boys, took both the RAKIT and the Dutch WISC-R; RAKIT IQ and WISC-R IQ showed a congruent validity of $r = .86$ (uncorrected for attenuation: $r = .79$). Bleichrodt, Resing, et al. did not report *g* scores, so they were computed by the present authors, resulting in correlations between WISC-R *g* and RAKIT *g* of $r = .79$ (uncorrected $r = .78$).

The question now is to find out what these correlations mean. Jensen (1980, Table 5) reports correlations between the WISC and other IQ tests: Median correlations range from .41–.80. It thus appears that the RAKIT correlates higher with the WISC than practically all of the standardized intelligence tests reported in the literature. This leads us to conclude that the RAKIT and the WISC-R measure the same underlying construct, that is general intelligence, to a very high extent.

Criteria

A teacher evaluated each child on five-point scales on Arithmetic, Dutch, Technical Reading, Comprehensive Reading, and Handicrafts (such as drawing, painting, and working with wood and clay). The reliabilities of the teacher's evaluations could not be calculated because each child was evaluated by only one teacher. These reliabilities were set to a value of .60; this value is based on de Groot (1978) reporting median values in the mid .40s. These empirical values are just slightly lower than those found in the meta-analysis of Rothstein (1990) on managers' evaluations of the quality of work of their employees, resulting in, on average, values of .52. However, to minimize the risk of overcorrecting for unreliability, a higher value of .60 is often chosen.

Statistical Analyses

Means and Reliability

To answer the question of group differences, the means and standard deviations of the total score on the RAKIT and of the different subtests were computed for the Dutch group and the various immigrant groups. The size of the mean differences in test scores between the Dutch group and the various immigrant groups were computed in terms of the Dutch group members' standard deviations. In order to get a measure of general intelligence, *g* scores for the participants were also computed. This was done by multiplying the *z* scores of a participant on the different subtests with the *g* loading of the subtest, and summing the products per participant on the different subtests. For all groups, the calculation of the *g* score was based on the *g* loadings of the subtests derived from the analyses on the Dutch groups, as these *g* loadings are free from bias. Correlations between the participants' total RAKIT score and the participants' *g* score were also calculated.

In order to calculate the *g* score of each participant, the *g* loading of a subtest must be computed. According to Jensen and Weng (1994) a good estimate of the subtest's *g* value can be made when a wide range of broad cognitive abilities (Stratum II) is measured, each by at least three first-order cognitive abilities (Stratum I). According to Ree and Earles (1991) the outcomes of several techniques to estimate *g* are highly comparable. In line with these findings, two different estimation techniques were used and their outcomes were compared. First, *g* scores were estimated taking into account the Schmid-Leiman (1957) decomposition. The *g* values of the subtests were computed by using path-tracing rules applied to a hierarchical model supplied by EQS (Bentler, 1996). The paths are traced from hierarchical *g* to the subtest, multiplying the path-coefficients of the paths taken along the way (Mulaik & Quartetti, 1997). Both Jensen and Weng (1994), and Thorndike (1985) state that the first unrotated principle-axis factor provides a good estimate of the subtest's *g* value. Therefore, the subtests' loadings on the first unrotated factor of a principle factor analysis were calculated as a second estimate of the subtests' *g* value. Following these two methods, congruence indices between the resulting *g* values were calculated and are reported in the Appendix.

In addition to the reliabilities of the subtest scores (Cronbach's α or split-half coefficients), the reliabilities of the RAKIT total score and the RAKIT *g* score were calculated. These reliabilities were estimated by using a method designed to estimate the reliability of a weighted sum (Nunnally, 1978, p. 250, formula 7–16). For the RAKIT total score the weights of the subtests were set at 1, for the RAKIT *g* score the weights of the subtests equaled the subtests' *g* value. These reliabilities were calculated separately for the complete Dutch group, all immigrant groups combined, and the individual immigrant groups.

Dimensional Comparability

In order to investigate whether the tests measured the same dimensions in the various groups, we tested for dimensional comparability. The fit of Carroll's hierarchical model of intelligence for our data was first tested on the majority group by means of structural-equation modeling using EQS (Bentler, 1996), which resulted in a model. Comparability of the meaning of the RAKIT for the Dutch group and the various immigrant groups was examined by means of a Multi-Group Analysis, in which each immigrant group was compared with the Dutch group. The basis for these comparisons was the established hierarchical model for the Dutch group described above. First, the equality of the covariance matrices was tested, secondly, the invariance of factor structure across Dutch and immigrant groups, and finally the invariance of factor structure across Dutch and immigrant group with the loadings of the subtests on the first order Broad Cognitive Abilities (Stratum II abilities) and the loadings of the Broad Cognitive Abilities on hierarchical *g* (Stratum III) constrained to be equal across groups.

Congruence indices (Tucker's ϕ) were calculated between the loadings of the subtests on the Broad Cognitive Abilities, and between the loadings of the Broad Cognitive Abilities on hierarchical *g* between the Dutch group and the various immigrant groups. Values above .85 are regarded as an indication of comparability of overall factor interpretation between the Dutch group and the immigrant group.

Furthermore, differences in loadings on a subtest of .10 between the majority and the immigrant group are interpreted as substantial. This interpretation is based on a study by Carretta and Ree (1995) on sex differences in cognitive abilities, in which most of the discrepancies in factor loadings were smaller then .05, the largest difference being .12.

Differential Item Functioning

When groups with the same cognitive abilities do not have the same probability of responding to an item of a test then the item functions differently or is biased. The Mantel-Haenszel method (MH; Holland & Thayer, 1988), one of the more widely used methods for detecting item-level bias in measurement (Millsap & Everson, 1993), was used to detect biased items. The scores of the participants on the subtests were divided into four to six successive score categories, depending on the number of items analyzed, for the Dutch and the various immigrant groups. An effect size was estimated for the difference between groups in the relationship of the proportion correct (*p* value) between score categories per item. This was tested on a 1% significance level. Only those items that were completed by 90% or more of all participants were subsequently tested for differential item functioning. Note that when a biased item is found with this technique, it implies that this is a statistically biased item, the item only deviates in a statistical sense. The next step after identifying the statistical biased item is trying to explain, on the basis of the statistical results and other information, what caused the item to deviate.

Differential Prediction

The predictive validity of a participant's *g* score was calculated for the teachers' evaluations for the Dutch group and the various immigrant groups. The regression equations for the Dutch group and the various immigrant groups were tested for prediction bias. Lautenschlager and Mendoza (1986) state that the best way to address differential prediction is to apply step-down hierarchical regression analyses to the data. The step-down hierarchical regression analysis starts with determining whether the criteria are significantly predicted by *g*. If they are, we tested whether the two variables *group* and *interaction of group with test* add significantly to the prediction of the criteria. And if these variables do significantly add to the prediction of the criteria, we tested whether the variable interaction of group with test adds significantly to the prediction (slope bias) and/or whether the variable group adds significantly to the prediction (intercept bias).

Spearman's Hypothesis Tested with RAKIT and Criterion Scores

Previous research has shown that the average test scores of immigrants are almost always lower than those of the Dutch group (te Nijenhuis & van der Flier, 1997; te Nijenhuis, Evers, & Mur, 2000). However, differences between groups vary tremendously, from practically none to as much as 2 *SD*s. In various studies this variation is explained by method bias (van de Vijver & Tanzer, 1997), focusing on factors such as language skills, scholastic knowledge, cultural influences, and acculturation. For instance, insufficient command of the language that is used in the test may have a general influence on the scores of tests with a substantial language component and may hardly affect scores on tests without a substantial language component. Furthermore, Crystallized tests may be more susceptible to scholastic knowledge than tests of Broad Visual Perception.

The question is how to check for method bias. Van de Vijver and Bleichrodt (2001) state that classical bias techniques, such as those for item bias, are not very suitable to detect factors that influence entire tests as opposed to single items. We tested for method bias using Spearman's hypothesis (Braden, 1989; Jensen, 1985, 1993, 1998; Jensen & Whang, 1993; Ja-Song & Lynn, 1992; Lynn & Holmshaw, 1990; Lynn & Shigehisa, 1991; Lynn, Chan, & Eysenck, 1991). Spearman's hypothesis states that the higher a test's *g* loading, the larger the mean score differences between two groups. This was tested by computing a correlation between the *g* value of the subtests and the difference between the means of the Dutch group and of each of the various immigrant groups. The higher the correlation, the more the difference between Dutch and immigrant groups can be attributed to group differences in *g* and less to method bias.

To estimate the importance of *g* in the differences between the Dutch group and immigrant groups, the mean differences were regressed on the *g* value of the subtests. To test Spearman's hypothesis, Jensen (1993) states the following methodological requirements:

1. The samples should not be selected on highly *g*-loaded criteria.
2. The variables should have a reliable variation in their *g* values.
3. The variables must measure the same latent traits in the different groups.
4. The variables must measure equal *g* values in the subgroups, which means that the congruence coefficient of the estimates should be above .95.
5. The *g* values should be computed separately; if the congruence coefficient indicates a high degree of similarity the *g* values can be averaged.
6. To prevent the correlation between the *g* value and the mean differences between the Dutch group and the different immigrant groups from being influenced by the different reliability of the variables, the variables should be corrected for attenuation.
7. The test of Spearman's hypothesis is the Pearson correlation between the *g* value of the subtests and the differences in means between the groups; these correlations should be statistically significant.

Consequently, if subtests that are dependent on Dutch language skills result in substantially larger group differences than is to be expected on the basis of their *g* loadings, the hypothesis that language proficiency is a factor of method bias would be confirmed.

In accordance with the test of Spearman's hypothesis with RAKIT scores, Spearman's hypothesis has also been computed with criterion scores (teachers' evaluations). The teachers' evaluations were not developed to measure *g*, so the correlation of the criteria with the *g* score measured by the RAKIT was used as an estimate of their *g* loadedness. Otherwise, the same methodology as Spearman's hypothesis tested with RAKIT scores applies.

Results

Means

As we are dealing with norm samples, the mean subtest scores of the Dutch group all equal the mean norm score of 15. The mean standard deviations are 5. The mean RAKIT IQ scores for this group are 100. Table 3 also

Table 3. Means and standard deviations for the Dutch group and immigrant group children, and deviation of the immigrant group children from the Dutch group in terms of the Dutch group's standard deviation (dev.) for children with a mean age of 7.8 years.

	Group										
	Dutch		Surinamese/Neth.Ant.			Turks			Moroccans		
Subtests	*M*	*SD*	*M*	*SD*	Dev.	*M*	*SD*	Dev.	*M*	*SD*	Dev.
Closure	14.85	5.06	13.77	6.28	.21	11.87	5.82	.59	8.60	5.51	1.24
Exclusion	14.96	5.07	12.36	5.26	.51	11.94	4.55	.60	10.52	4.71	.88
Memory Span	15.05	4.94	15.39	5.47	–.07	15.27	6.34	–.02	13.53	5.52	.31
Verbal Meaning	15.03	5.14	9.03	6.11	1.17	4.14	4.74	2.11	3.52	4.52	2.24
Mazes	15.02	5.03	12.74	5.69	.45	11.79	5.15	.64	11.37	4.56	.73
Analogies	15.03	4.94	14.57	4.28	.09	10.42	4.58	.93	10.95	4.92	.83
Quantity	15.21	5.10	11.69	5.61	.69	10.46	5.83	.93	8.38	4.96	1.34
Disks	15.01	5.05	12.41	4.61	.51	11.77	4.95	.64	9.68	4.41	1.06
Learning Names	15.05	5.05	13.13	4.28	.38	9.61	4.81	1.08	8.67	5.31	1.26
Hidden Figures	14.94	4.93	13.02	4.92	.39	11.32	5.36	.73	10.52	4.05	.90
Idea Production	15.06	5.18	11.84	4.97	.62	10.87	5.26	.81	11.27	5.64	.73
Storytelling	14.99	5.05	11.03	5.16	.78	10.14	4.97	.96	10.25	5.55	.94
RAKIT IQ	99.93	15.00	86.46	16.51	.89	76.97	14.13	1.53	72.10	14.09	1.85
RAKIT *g*	1.83	2.72	–.37	3.00	.81	–2.07	2.74	1.43	–3.14	2.70	1.83

shows the mean RAKIT *g* scores for these groups, the corresponding scores for the various immigrant groups, and, in addition, the deviation from the Dutch group in the Dutch group's *SD*s.

The scores reveal a clear pattern: The Dutch children outscore the immigrant children, and the Surinamese/Netherlands Antillean children outscore the Turkish and Moroccan children. The Surinamese/Netherlands Antillean children deviate, on average, up to one standard deviation on both RAKIT IQ and RAKIT *g*. The scores of Turkish children deviate more than one standard deviation on RAKIT IQ and RAKIT *g*. The scores of the Moroccan children are highly comparable to those of the Turkish children. For all immigrant groups, the highest deviations are on the subtest Verbal Meaning, which has a substantial verbal component, and smallest on the subtest Memory Span, on which the Surinamese children score even slightly better than the children from the majority group. The Turkish and Moroccan children show relatively large deviations on other language-related subtests (Idea Fluency, Learning Names, and Storytelling), whereas the Surinamese/Netherlands Antillean children show relatively smaller deviations on these tests. However, subtests having no, or only a small, language component (Disks, Hidden Figures, Exclusion, and Quantity) also reveal a substantial deviation from the majority group for all immigrant groups.

Reliability

Table 4 shows the reliabilities of the subtest, RAKIT IQ, and RAKIT *g* for the Dutch group, the complete immigrant group, and the individual immigrant groups. Except for the subtest Disks, which has a very low reliability for the Moroccan group ($\alpha = .44$), the reliabilities of the RAKIT subtests are satisfactory; besides, they are highly comparable across the majority group and the various immigrant groups. The reliabilities of RAKIT IQ and RAKIT *g* are very high and consistent across groups.

Dimensional Comparability

The fit of Carroll's hierarchical model of intelligence for our data was first tested on the Dutch group, using EQS (Bentler, 1996). Subtests' intercorrelations are not reported due to space limitations. The first author will gladly supply all correlation matrices upon written request. An initial model (not reported) with the four lower order factors Hybrid (G_h), Broad Visual Perception (G_v), General Memory and Learning (G_m), Broad Retrieval Ability (G_r), and a hierarchical factor *g*, showed an acceptable fit ($CFI > .90$), but examination of the outcomes of the Lagrange Multiplier test (LM-test) indicated a substantial increase in fit when three error correlations were freely estimated. These error correlations belonged to the subtests Disks, Mazes, and Idea Production, being the three subtests with a speed component. These outcomes supply both a statistical and a theoretical rationale for the inclusion of the three freely estimated error correlations into the final model. The Comparative Fit Index for the majority group is: $CFI = .976$, $\chi^2(48) = 56.72$, $p = .181$; this high *CFI* is indicative of a good fit.

Before testing the invariance of the factorial structure between the majority group and the various immigrant

Table 4. Reliability coefficients for the subtests of the RAKIT, RAKIT IQ, and RAKIT *g*.

			Group		
Subtests	Dutch	All Immigrants	Sur./Neth.Ant.	Turks	Moroccans
Closure	.81	.88	.89	.86	.86
Exclusion	.85	.86	.88	.84	.90
Memory Span*	.75	.88	.85	.91	.92
Verbal Meaning	.72	.89	.82	.90	.88
Mazes	.82	.73	.73	.71	.76
Analogies	.88	.83	.81	.81	.83
Quantity	.87	.93	.92	.93	.88
Disks*	.71	.69	.69	.64	.44
Learning Names	.83	.84	.75	.81	.92
Hidden Figures	.85	.83	.83	.85	.81
Idea Production	.84	.82	.83	.81	.88
Storytelling	.68	.82	.73	.73	.86
RAKIT IQ	.94	.97	.95	.95	.96
RAKIT *g*	.99	.99	.99	.99	.99

Note. Sur./Neth.Ant. = Surinamese/Netherlands Antilleans. * = for these subtests the split-half coefficient was used instead of Cronbach's α.

Table 5. Results of the EQS analyses comparing the Dutch group with each immigrant group.

Model	Group	χ^2	*df*	CFI
All subtests included				
1. Equality of Covariance Matrices	Sur./Neth.Ant.	109.48**	78	.942
	Turks	68.41*	78	1.000
	Moroccans	98.38**	78	.940
2. Equality of Factor Models	Sur./Antill.	132.01**	96	.933
	Turks	111.32*	96	.969
	Moroccans	130.37**	97	.939
3. Equality of Factor Loadings	Sur./Antill.	140.48**	108	.940
	Turks	115.76*	108	.984
	Moroccans	162.75***	109	.901
Verbal Meaning excluded				
1. Equality of Covariance Matrices	Sur./Neth.Ant.	80.05*	66	.970
	Turks	53.45*	66	1.000
	Moroccans	88.38**	66	.954
2. Equality of Factor Models	Sur./Neth.Ant.	98.47*	76	.972
	Turks	82.94*	76	.984
	Moroccans	96.44*	77	.960
3. Equality of Factor Loadings	Sur./Neth.Ant.	96.90*	87	.979
	Turks	87.72*	86	.998
	Moroccans	130.65**	88	.912

Note. CFI = Comparative Fit Index, Sur./Neth.Ant. = Surinamese/Netherlands Antilleans, *$p < .05$, **$p < .01$, ***$p < .001$.

groups, the covariance matrices were tested for equality. Table 5 show all *CFI*s. The *CFI*s for these analyses were all within acceptable range and varied from *CFI* = .940 to *CFI* = 1.00. Secondly, the data were tested for invariance of factorial structure across the Dutch group and the various immigrant groups. Again, all *CFI*s were within acceptable range and varied from *CFI* = .933 to *CFI* = .969 . Finally, the data were tested for invariance of factor loadings (invariance of factorial structure with the loadings of the subtests on the lower order factors and the loadings of the lower order factors on hierarchical *g* constrained to be equal across groups). The *CFI*s were within acceptable range and varied from *CFI* = .901 to *CFI* = .984.

Table 6. g Values of subtests, loadings of subtests on broad cognitive abilities, loadings of broad cognitive abilities on hierarchical *g*, and congruence coefficients for the Dutch group and the immigrant groups.

			Group					
	Dutch		Sur./Neth.Ant.		Turks		Moroccans	
Subtest, grouped by factor	*g* value	Factor load.	*g* value	Factor load.	*g* value	Factor load.	*g* value	Factor load.
Hybrid factor (G_h)								
Analogies	.528	.528	.853	.853	.626	.626	.481	.481
Verbal Meaning	.441	.441	.505	.505	.580	.580	.564	.564
Quantity	.647	.647	.705	.705	.660	.660	.715	.715
Visual factor (G_v)								
Disks	.493	.540	.494	.719	.367	.427	.500	.645
Exclusion	.580	.635	.452	.658	.506	.589	.298	.385
Mazes	.391	.428	.439	.639	.332	.387	.429	.553
Hidden Figures	.590	.646	.418	.609	.509	.592	.207	.267
Closure	.383	.419	.339	.494	.313	.364	.465	.600
Memory factor (G_m)								
Memory Span	.419	.533	.493	.598	.312	.315	.393	.393
Learning Names	.399	.507	.512	.621	.420	.425	.822	.822
Retrieval factor (G_r)								
Idea Production	.245	.604	.285	.461	.418	.748	.679	.898
Storytelling	.227	.561	.266	.430	.406	.726	.515	.681
Congruence coefficient			.971	.977	.973	.979	.890	.943
Factors								
Hybrid (G_h)	1.000		1.000		1.000		1.000	
Visual (G_v)	.913		.687		.859		.775	
Memory (G_m)	.787		.825		.989		1.000	
Retrieval (G_r)	.405		.619		.559		.756	
Congruence coefficient			.981		.991		.972	

Note. g computed by means of Schmid-Leiman decomposition. Sur./Neth.Ant. = Surinamese/Netherlands Antilleans.

Both theoretical considerations and the results of the LM-tests made us decide to rerun all analyses leaving out the subtest Verbal Meaning, with its susceptibility to language bias. This increased the comparability for all immigrant groups. In general, the model shows excellent fit for most groups. This is a strong indication that the same dimensions are being measured in both the Dutch group and the immigrant groups.

Table 6 shows the loadings of the subtests on the Broad Cognitive Abilities (first-order factors), and the loadings of the Broad Cognitive Abilities on *g*, provided by EQS. Congruence coefficients were computed based on the loadings of the subtests on the Broad Cognitive Abilities (the four lower-order factors), and based on the loadings of the Broad Cognitive Abilities on *g*, comparing the various immigrant groups with the majority group. In general, the congruence coefficients are high, indicating highly comparable overall interpretations of factors between the Dutch and the immigrant group. However, although the overall picture is one of generally highly comparable factors, there are substantial differences between factor loadings of individual subtests. About half of the differences in loading of subtests between the Dutch and the various immigrant groups is somewhat larger than .10, and there are some even larger differences. The differences between loadings of subtests are, generally speaking, largest on the subtests Idea Production and Storytelling. The Surinamese/Netherlands Antillean groups show less difference in loadings than do other immigrant groups. There are also substantial differences between the loadings of the lower-order factors on *g*. All immigrant groups show a difference in factor loading larger than .10 for the Retrieval factor on *g*. Furthermore, the Visual factor shows a difference in loading larger than .10 for Surinamese/Netherlands Antilleans and Moroccans and the Memory factor shows a difference in loading larger than .10 for Turks.

Differential Item Functioning

Table 7 displays the items (by means of the Mantel-Haenszel statistic) identified as being statistically biased against the various immigrant groups as well as the effect of the statistically biased items on the scores of the immigrants. Because the Mantel-Haenszel statistic requires di-

Table 7. Differential item functioning using the Mantel-Haenszel statistic.

	Group								
	Sur/Neth.Ant.			Turks			Moroccans		
Subtest	No.	No. Bias	Eff.	No.	No. Bias	Eff.	No.	No. Bias	Eff.
Analogies	10	2	.00	–	–	–	–	–	–
Verbal Meaning	36	10	.20	22	1	.00	22	1	.00
Learning Names	24	0	–	24	3	.17	24	4	.22
Closure	36	10	.11	36	15	.34	31	13	.25
Exclusion	29	1	.01	28	2	.04	26	6	.11
Quantity	33	10	.13	30	7	.04	28	8	.07

Note. MH statistic is computed on the basis of items answered by 90% of the participants in the group. Sur./Neth.Ant. = Surinamese/Netherlands Antilleans; No. = number of items analyzed; No. Bias = Number of biased items; Eff. = estimated score improvement (effect) if the biased items were replaced.

Table 8. Means and standard deviations for the Dutch and immigrant groups, and deviation from the Dutch group in terms of the Dutch group's standard deviation (dev.) for criteria.

	Group										
	Dutch		Sur./Neth.Ant.			Turks			Moroccans		
Criteria	Mean	*SD*	Mean	*SD*	Dev.	Mean	*SD*	Dev.	Mean	*SD*	Dev.
Arithmetic	5.74	1.99	5.06	2.10	.34	4.43	2.26	.66	4.56	2.42	.59
Dutch	5.73	2.09	5.37	1.97	.17	3.92	2.14	.87	4.74	2.35	.47
Technical Reading	5.80	2.23	5.77	1.99	.01	4.38	2.21	.64	5.38	2.48	.19
Compr. Reading	5.70	2.09	5.20	2.01	.24	3.44	1.95	1.08	4.34	2.42	.65
Handicrafts	5.45	1.75	5.44	1.95	.01	5.08	2.00	.21	4.96	1.87	.28
Total School Grade	5.67	1.69	5.58	1.87	.05	4.25	1.93	.84	4.83	2.18	.50
Est.Teach.IQ Child*	5.91	1.70	5.43	1.78	.28	4.55	1.86	.80	5.02	1.95	.52

Note. *Est. Teach.IQ Child = teacher estimation of child's IQ

chotomic scoring of the items, correct or not correct, analyses were not carried out for subtests that could not be scored this way. Furthermore, analyses could not be carried out for Hidden Figures because of data loss.

The number of biased items varied strongly among the subtests, among others, due to the termination rules. However, more important than the number of significantly biased items is their negative effect on the mean scores of immigrants. For example, for the Moroccans there is substantial bias in the subtest Learning Names. If the four statistically biased items are replaced with nonbiased items they could be expected to have, on average, a score of .22 *SD* higher. For Turks this effect is somewhat smaller: .17 *SD*. However, these effects could be regarded as underestimates, because not all items were analyzed with the MH statistic; potential bias in the remaining, nonanalyzed items could, therefore, have not been detected.

Post hoc inspection of the biased items was conducted and revealed words that could be interpreted as being difficult to understand for immigrants. It may be that the subtests call more strongly upon Dutch language skills than is desirable, given what the test is supposed to measure. On the other hand, some other items that would seem to be of comparable difficulty turned out to be not statistically biased. This is in line with the conclusion drawn in previous item-bias research that item bias is not always predictable.

Differential Prediction

Table 8 shows the means and standard deviations for various criteria; for the immigrant groups, the deviation from the Dutch group in terms of the Dutch group's standard deviation is shown. For all immigrant groups, deviations from the Dutch group are relatively large for criteria with a language component (for instance Vocabulary and Language Usage), however, criteria that are less language-related show relatively large deviations as well. Surinamese/Netherlands Antilleans show the smallest deviation from the majority group, and Turks and Moroccans show the largest deviations from the Dutch group.

Correlations between criteria and RAKIT IQ, and between the criteria and RAKIT *g* for the majority groups and various immigrant groups were computed (and are available from the first author). For all majority groups and for most immigrant groups, RAKIT *g* generally showed consistently higher correlations with the criteria

Table 9a. First step in hierarchical regression procedure: are the criterion measures (teacher evaluations) significantly predicted by *g*?

	Groups								
	Dutch with Sur./Neth.Ant.			Dutch with Turks			Dutch with Moroccans		
Criterion Measure	*R*	R^2	*F*	*R*	R^2	*F*	*R*	R^2	*F*
Arithmetic	.50	.25	71.37*	.54	.29	96.37*	.50	.25	75.26*
Dutch	.41	.17	43.74*	.53	.28	90.40*	.42	.18	48.02*
Technical Reading	.26	.07	16.40*	.39	.15	42.12*	.27	.08	18.07*
Comprehensive Reading	.44	.19	52.08*	.56	.31	102.62*	.45	.20	54.36*
Handicrafts	.26	.07	15.52*	.33	.12	28.05*	.29	.08	20.25*
Total School grade	.46	.22	60.53*	.59	.34	120.31*	.50	.25	73.04*
Est.Teach.IQ Child	.49	.24	64.77*	.55	.31	101.04*	.49	.24	69.76*

Note. Sur./Neth.Ant. = Surinamese/Netherlands Antilleans; Est.Teach.IQ Child = Teacher estimation of child's IQ; **p* < .001.

Table 9b. Second step in the hierarchical regression procedure: increase in prediction if different regression equations for the Dutch and immigrant groups are assumed.

	Groups					
	Dutch with Sur./Neth.Ant.		Dutch with Turks		Dutch with Moroccans	
Criterion Measure	ΔR^2	*F*	ΔR^2	*F*	ΔR^2	*F*
Arithmetic	.00	.14	.00	.17	.01	1.50
Dutch	.01	.65	.01	1.10	.01	.92
Technical Reading	.01	1.01	.01	.74	.01	1.66
Comprehensive Reading	.00	.20	.03	4.84**	.00	.20
Handicrafts	.01	.95	.02	2.36	.01	.74
Total School Grade	.02	2.81	.00	.16	.02	3.32**
Est.Teach.IQ Child	.01	1.18	.01	.78	.02	2.20

Note. Sur./Neth.Ant. = Surinamese/Netherlands Antilleans; Est.Teach.IQ Child = Teacher estimation of child's IQ; ***p* < .05.

Table 9c. Third step in the hierarchical regression procedure: increase in prediction if different regression equations for the Dutch and immigrant groups are assumed.

		Slope Bias		Intercept Bias	
Groups Criterion Measure		ΔR^2	*F*	ΔR^2	*F*
Dutch+Turks	Comprehensive Reading	.00	.23	.03	9.50**
Dutch+Moroccans	Total School Grade	.00	.69	.02	6.35**

Note. ***p* < .05.

than RAKIT IQ, although the differences were quite small.

To find out whether the *g* scores predicted the criteria differently for the majority and various immigrant groups, the data were applied to a step-down hierarchical regression procedure (Lautenschlager & Mendoza, 1986). The first step in this procedure is to test whether the criteria are, by any means, significantly predicted by the *g* scores. The second step tests whether the prediction is biased, either in intercept or slope. If the prediction is biased, the third step in this procedure would be to test whether the bias is in the intercept, in the slope, or in both.

Table 9a shows the results of the first step in the step-down hierarchical regression procedure. All criteria were significantly predicted by the *g* scores. Table 9b shows the results of the second step. For the Surinamese children, none of the criteria was predicted better when one assumed different regression equations for the Dutch and the immigrant group. For the Turkish children, one criterion out of seven was predicted better when one assumed different regression equations for the Dutch and the immigrant group. For the Moroccan children the prediction also improved for one of the seven criteria when one assumed different regression equations for the majority group and the immigrant group. Table 9c shows the results of the third step. In both instances intercept bias was found: Comprehensive Reading for Turks, $F(1, 229) = 9.50$, $p = .002$; and Total School Grade for Moroccans, $F(1, 219) = 6.35$, $p = .012$. In only 10% of the instances was a very small prediction bias found.

Table 10. Factor loadings of criteria on the first principle axis factor, correlation of the criteria with *g* score, *r*, and congruence indices.

	Groups							
	Dutch		Sur./Neth.Ant.		Turks		Moroccans	
Criteria	*FL*	*r*	*FL*	*r*	*FL*	*r*	*FL*	*r*
Arithmetic	.65	.49	.71	.45	.78	.48	.76	.42
Dutch	.83	.42	.98	.38	.83	.44	.87	.33
Technical Reading	.80	.28	.87	.25	.86	.34	.84	.30
Compr. Reading	.82	.42	.81	.46	.84	.40	.86	.25
Handicrafts	.49	.29	.52	.22	.35	.43	.60	.24
Total School Grade	.99	.51	.97	.42	.99	.51	.98	.46
Est.Teach.IQ Child	.81	.52	.92	.34	.83	.36	.92	.33
Congruence coefficient			.999	.987	.990	.980	.996	.985

Note. FL = Factor Loading; Sur./Neth.Ant. = Surinamese/Netherlands Antilleans; Est. Teach.IQ Child = teacher estimation of child's IQ

Spearman's Hypothesis Tested With RAKIT Scores

The sample was representative both for the Dutch and the immigrant population and was not selected on criteria with high *g* loadings. The variables show reasonable variation in their *g* values, although the subtests with low *g* loadings are somewhat underrepresented. The third requirement, stating that the variables should measure the same latent traits in the different groups, was met. The fourth requirement states that the variables must measure equal *g* values in the subgroups, which means that the congruence indices should be above .95. This requirement is not met for the Moroccans, so Spearman's hypothesis was not tested for this group. The *g* values were averaged and the group differences on subtests were corrected for attenuation.

In order to test Spearman's hypothesis, the correlations between V_{es} and V_g were computed (r = Pearson correlation; r_s = Spearman's rank order correlation): Surinamese/Netherlands Antilleans $r = –.18$, $r_s = –.063$, $p = .846$, and ES = –.49 g+.79; Turks $r = .32$, $r_s = .26$, $p = .417$, and ES = 1.60 g+.13. When Spearman's hypothesis is confirmed the group differences can be largely attributed to *g*. However, there is no substantial (Turks) or even a negative correlation (Surinamese/Netherlands Antilleans).

Spearman's Hypothesis Tested With Criterion Scores

Jensen's (1993) first requirement was met in the same way as in the previous section. Table 10 shows the correlations of the criterion variables with *g,* the loadings of the criteria on the first principle-axis factor and congruence indices (Tucker's ϕ). The second criterion was reasonably met, only criteria with low *g* loadings were underrepresented. The third requirement, stating that the same latent traits should be measured, was also met; the congruence indices between the loadings on the first principle-axis factor of the majority group and the various immigrant groups were high. Jensen's fourth requirement states that the variables should measure the same *g* in the different groups, indicated by a congruence coefficient above .95. This criterion was also met. The *g* loadings of the majority group and each immigrant group were averaged.

For the immigrant groups the correlations between V_{es} and V_g were: Surinamese/Netherlands Antilleans $r = .74$, $r_s = .92$, $p = .000$, and ES = 1.43 g–1.53; Turks $r = .40$, $r_s = .36$, $p = .216$, and ES = 1.16 g+.18; and Moroccans $r = .64$, $r_s = .50$, $p = .127$, and ES = 1.26 g–.05. In most instances, the group differences in criterion scores were strongly accounted for by the group differences in *g*. Criteria that lie above the regression line were relatively more difficult for immigrant group members than might be expected from their *g* loadedness, criteria under the regression line were relatively easier. The criteria that lie above the regression line nearly all have a substantial language component (Language Usage, Vocabulary, Dutch, Technical Reading, and Comprehensive Reading).

Discussion

The results for 7-year-old children provide important indications that the standardized ability test RAKIT is highly, though not perfectly, valid for the assessment of immigrant children. Results for 5-year-old and 9-year-old children are not reported, but yield highly comparable outcomes.

The subtests that call upon Dutch language skills showed the largest differences between the means of the Dutch group and the immigrant groups. However, and less so, tests that do not call for knowledge of the Dutch language also showed a large difference between the scores of the Dutch group and the immigrant groups. The assumption that the differences on the subtests can only be attributed to differences in proficiency in Dutch is not, therefore, supported by these outcomes. All reliabilities are satisfactory and highly comparable among all groups.

In general, the RAKIT subtest scores measure the same cognitive abilities as strongly in the various immigrant groups as in the Dutch group. Only the subtest Verbal Meaning does not appear to be comparable across groups, probably because it has a strong language component. However, notwithstanding the overall comparability, there are several substantial discrepancies from the Dutch group in factor loadings on the broad cognitive abilities of the individual subtests. For the Surinamese children these discrepancies are on average .10, for the Turkish and Moroccan children they are on average .15, clearly higher than the differences found by Carretta and Ree (1995) in their study on sex differences.

The analyses concerning the question of whether the same test scores predict the same criterion behavior showed the *g* scores are good predictors of the teachers' evaluations. Only in a small minority of cases did the prediction for the criteria improve significantly if different regression equations were taken for the Dutch as compared to the immigrant group. However, the effects were very small. For criteria with a substantial language component (Vocabulary, Language Usage, and Comprehensive Reading), the common regression lines overestimated the teachers' evaluation in some cases. The other criteria found to show intercept bias were underestimated by the common regression line. Teachers sometimes judge immigrant students on other grounds than just their command of the subject; Van de Vijver and Willemse (1991), for example, found that immigrant group grades reflected improvement rather than their command of the subject.

In general, item bias has the smallest effect on the mean scores of Surinamese/Netherlands Antilleans; this is not surprising since these children have grown up in a culture where Dutch is the official language of education. Post hoc inspection of the content of the statistically biased items revealed that in most instances these items contained idiom that is relatively difficult for immigrant children. Item bias in the subtest Closure also showed substantial effects on the mean scores of immigrant children. Closure might rely partly on Lexical Knowledge because children not only have to figure out what the incomplete picture is supposed to represent, but they also have to label it correctly. However, some of the biased items may be considered more difficult for immigrant children while others may not. Therefore, it is at this point impossible to draw unequivocal conclusions. The occurrence of statistical bias among the last in the series of analyzed items, which occurred quite often, might be an artifact of the method used, in which case the statistical bias could therefore be interpreted as position bias. There are a few cases in which both tests with and without a language component showed substantial bias, which we find hard to interpret. Practitioners usually work with the sum score of a whole battery, so that small to medium effects of bias in a number of subtests result in an even smaller bias effect on the sum score.

Spearman's hypothesis offers a simple, straightforward explanation for the great variance in mean differences in test scores and criterion scores between the immigrant children and the Dutch children. A moderate to strong relation was found between differences in score means of the groups and *g* loadings of the scores in four of five instances, making *g* highly accountable for differences between the majority group and the immigrant groups. So, *g* is the dominant factor in accounting for differences between majority and immigrant groups. However, the relations were not as strong as demonstrated elsewhere in the literature. The data clearly indicated method bias in relation to differential command of the Dutch language: Scores on criteria with a substantial language component are higher above the regression line than one would expect from their *g* loadings. The influence of language skills on tests with a substantial language component was also shown in the item bias analyses and has now been clearly documented. How plausible is the existence of other forms of method bias? Although research on scholastic skills, cultural influences, and acculturation has, to the best of our knowledge, not been published, it seems plausible these factors play some role. Further research is needed to address the issue of other forms of method bias.

On a cautionary note, we would like to state explicitly that the tests of Spearman's hypothesis as used in this paper only allow for conclusions about group differences in mean phenotypic intelligence, and not about group differences in mean genotypic intelligence. Convincing studies about group differences in mean genotypic intelligence would require quasi-experimental designs using, for instance, monozygotic and dizygotic twins reared in different surroundings, i.e., apart, in both Dutch and immigrant families; the samples in our study clearly do not meet these requirements.

An interesting finding in our study is that whereas all previous studies found a confirmation of the weak form of Spearman's hypothesis, meaning a positive correlation between mean group differences and *g* loading, we found one instance of a negative correlation, meaning a disconfirmation of Spearman's hypothesis; however, all other correlations were positive.

Limitations of this Study

With regard to the use of various statistical techniques, the sample sizes for the Dutch group are adequate, but the sample sizes for the three immigrant groups at each of the three age groups are somewhat small. This affects, for instance, the power of the various significance tests, and the outcomes of the structural equations modeling analyses. However, all of the findings are highly comparable over the various immigrant groups and age groups – there is only little variation in effect sizes –, and the general findings of our study – only little bias – fit in quite well with the rest of the empirical studies, so they strengthen the conclusions from this study. It appears that the finding of limited bias is quite generalizable.

When considering criterion validity, ratings by teachers are not broad enough. There is, however, an extensive nomological network of studies, showing that IQ scores of primary school children predict various long-term criteria (Jensen, 1980, 1998).

Conclusions

This paper addresses the suitability of the RAKIT IQ tests for immigrants. The analyses show that the dimensions of the RAKIT are highly comparable between the Dutch group and the immigrant groups, differential prediction has no strong effects, and item bias seldom has strong effects. This leads to the conclusion that, in general, the RAKIT is a legitimate instrument for the assessment of minorities. However, a test user should be careful interpreting the scores on the subtests of the RAKIT with a language component, especially for the Turkish and the Moroccan children. To the best of our knowledge, no peer-reviewed studies on test bias against immigrant children in West-European countries have been published outside the Netherlands. As many West-European countries resemble each other in that most of their immigrants come from third-world countries, including former colonies, the Dutch findings may probably be generalized to other West-European countries and probably even to specific groups of immigrants in the United States, such as Hispanics or Mexican-Americans.

Acknowledgment

We thank Peter Dekker (Vrije Universiteit, Amsterdam) for his help in organizing the data.

References

Bentler, P.M. (1996). *EQS, a structural equation program version 5.4*[Conputer software]. Encino, CA: Multivariate Software, Inc.

Bleichrodt, N., Drenth, P.J.D., Zaal, J.N., & Resing, W.C.M. (1984). *Revisie Amsterdamse Kinder Intelligentie Test* [Revision Amsterdam Child Intelligence Test]. Lisse, The Netherlands: Swets.

Bleichrodt, N., Resing, W.C.M., Drenth, P.J.D., & Zaal, J.N. (1987). *Intelligentiemeting bij kinderen* [The measurement of children's intelligence]. Lisse, The Netherlands: Swets.

Bleichrodt, N., Drenth, P.J.D., Zaal, J.N., & Resing, W.C.M. (1987). *Revisie Amsterdamse Kinder Intelligentie Test. Handleiding* [Revision Amsterdam Child Intelligence Test. Manual]. Lisse, The Netherlands: Swets.

Bleichrodt, N., Hoksbergen, R.A.C., Athavale, U., Kher, R., & Khire, U. (1991). *Indian Child Intelligence Test*. Pune, India: Jnana Prabodhini.

Bleichrodt, N., Hoksbergen, R.A.C., Khire, U., & Dekker, P.H. (1998). Vergelijkbaarheid van een intelligentietest in India en in Nederland [Comparability of an intelligence test in India and in the Netherlands]. *Kind en Adolescent, 19,* 396–412.

Bleichrodt, N., Hoksbergen, R.A.C., & Khire, U. (1999). Cross-cultural testing of intelligence. *Cross-Cultural Research, 33,* 3–25.

Braden, J.P. (1989). Fact or artifact? An empirical test of Spearman's hypothesis. *Intelligence, 13,* 149–155.

Carretta, T.R., & Ree, M.J. (1995). Near identity of cognitive structure in sex and ethnic groups. *Personality and Individual Differences, 19,* 149–155.

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

de Groot, A.D. (1978). *Vijven en zessen* [Fives and sixes/Humming and hawing]. Groningen, The Netherlands: Wolters-Noordhoff.

de Jong, M.J. (1985). Het onderwijsniveau van allochtone leerlingen in Rotterdam [The educational level of immigrant children in Rotterdam]. In M.J. de Jong (Ed.), *Allochtone kinderen op Nederlandse scholen: Prestaties, problemen en houdingen* [Immigrant children in Dutch schools: Achievements, problems, and attitudes] (pp. 3–12). Lisse, The Netherlands: Swets.

de Jong, M.J. (1987). *Herkomst, kennis en kansen. Allochtone en autochtone leerlingen tijdens de overgang van het basis naar het voortgezet onderwijs* [Origin, knowledge, and opportunities. Immigrant and majority group pupils during the transition from primary to secondary education]. Lisse, The Netherlands: Swets.

de Jong, M.J., & Van Batenburg, Th. A. (1984). Etnische herkomst, intelligentie en schoolkeuze advies [Ethnic origin, intelligence, and school choice advice]. *Pedagogische Studiën, 61,* 362–371.

Drenth, P.J.D., Bleichrodt, N., Setiono, T., & Poespadibrata, S. (1975). *Test Intelligensi Anak* [Child Intelligence Test]. Amsterdam, The Netherlands: Vrije Universiteit.

Drenth, P.J.D., Van der Flier, H., Muinde, N.P., Otaala, B., Omari, I.M., & Opolot, J.A. (1980). *Jatibio Akili Mtot Afrika* [African Child Intelligence Test]. Amsterdam, The Netherlands: Vrije Universiteit.

Evers, A., & Lucassen, W. (1991). *Handleiding DAT '83. Differ-*

entiële aanleg testserie [Manual DAT'83. Differential aptitude test series]. Lisse, The Netherlands: Swets.

Evers, A., Vliet-Mulder, J.C., & Ter Laak, J.(1992). *Documentatie van tests en test research in Nederland* [Documentation of tests and test research in the Netherlands]. Assen, The Netherlands: Van Gorcum.

Gottfredson, L.S., (1997). Why *g* matters: The complexity of everyday life. *Intelligence, 24,* 79–132.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Hunter, J.E., Schmidt, F.L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721–735.

Ja-Song, M., & Lynn, R. (1992). Reaction times and intelligence in Korean children. *Journal of Psychology, 126,* 421–428.

Jensen, A.R. (1980). *Bias in mental testing.* London: Methuen.

Jensen, A.R. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *The Behavioral and Brain Sciences, 8,* 193–219.

Jensen, A.R. (1993). Spearman's hypothesis tested with chronometric information-processing tasks. *Intelligence, 17,* 47–77.

Jensen, A.R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Jensen, A.R., & Weng, L.-J. (1994). What is a good *g*? *Intelligence, 18,* 231–258.

Jensen, A.R., & Whang, P.A. (1993). Reaction times and intelligence. A comparison of Chinese-American and Anglo-American children. *Journal of Biosocial Science, 25,* 397–410.

Latuheru, E.J., & Hessels, M.G.P. (1994). Schoolprestaties van allochtone leerlingen: De invloed van etnische herkomst [School achievements of immigrant pupils: The influence of ethnic origin]. *Tijdschrift voor Onderwijsreseach, 3,* 227–239.

Lautenschlager, G.J., & Mendoza, J.L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 133–139.

Lynn, R., Chan, J.W.C., & Eysenck, H.J. (1991). Reaction times and intelligence in Chinese and British children. *Perceptual and Motor Skills, 72,* 443–452.

Lynn, R., & Holmshaw, M. (1990). Black-white differences in reaction times and intelligence. *Social Behavior and Personality, 18,* 299–308.

Lynn, R., & Shigehisa, T. (1991). Reaction times and intelligence: A comparison of Japanese and British children. *Journal of Biosocial Science, 23,* 409–416.

Messick, S. (1989). Validity. In R.L. Linn (Ed.). *Educational measurement* (3rd ed.) (pp. 13–103). Washington, DC: American Counsel on Education, MacMillan.

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Ministerie van Onderwijs, Cultuur & Wetenschappen [Dutch Ministry of Education, Culture, & Science] (1997). *Voortgezet onderwijs in cijfers 1997* [Secondary education in figures 1997]. Available at www.minocw.nl.

Mulaik, S.A., & Quartetti, D.A. (1997). First order or higher order general factor? *Structural Equation Modeling, 4,* 193–211.

Neisser, U., Boodoo, G., Bouchard, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., & Urbina, S. (1996). Intelligence, knowns, and unknowns. *American Psychologist, 51,* 77–101.

Nunnally, J.C. (1978). *Psychometric theory.* New York: McGraw Hill.

Penninx, R. (1988). *Minderheidsvorming en emancipatie: Balans van kennisverwerving ten aanzien van immigranten en woonwagenbewoners* [The formation of minorities and emancipation: An account of knowledge acquisition with regard to immigrants and trailer park residents]. Alphen aan den Rijn, The Netherlands: Samson.

Ree, M.J., & Earles, J.A. (1991a). The stability of *g* across different methods of estimation. *Intelligence, 15,* 271–278.

Ree, M.J., & Earles, J.A. (1991b). Predicting training success: Not much more than *g*. *Personnel Psychology, 44,* 321–332

Ree, M.J., Earles, J.A., & Teachout, M.S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79,* 518–524.

Resing, W.C.M., Bleichrodt N., & Drenth, P.J.D. (1986). Het gebruik van de RAKIT bij allochtoon etnische groepen [Use of the RAKIT for the assessment of immigrants]. *Nederlands Tijdschrift Voor de Psychologie, 41,* 179–188.

Resing, W.C.M., & Hessels, M.P.G. (2001). Het meten van de cognitieve mogelijkheden en het schoolgedrag van allochtone kinderen [The measurement of cognitive opportunities and the school behavior of immigrant children]. In N. Bleichrodt & F. van de Vijver (Eds.), *Diagnostiek bij allochtonen* [Diagnosing immigrants] (pp. 89–118). Lisse, The Netherlands: Swets.

Roelandt, Th., & Veenman, J. (1988). *Minderheden in Nederland. Positie in het onderwijs* [Minorities in the Netherlands. Their educational situation]. Rotterdam, The Netherlands: Erasmus Universiteit/ISEO.

Rothstein, H.R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunities to observe. *Journal of Applied Psychology, 75,* 322–327.

Schmid, J., & Leiman, J.M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61.

Schmidt, F.L., Pearlman, K., & Hunter, J.E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33,* 705–724.

Schmidt, F.L., & Hunter, J.E. (1989). Interrater reliabilities cannot be computed when only one stimulus is rated. *Journal of Applied Psychologie, 74,* 368–370.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Snow, R.E., & Yalow, E. (1982). Education and intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 493–585). Cambridge: Cambridge University Press.

te Nijenhuis, J. (1997). *Comparability of test scores for immigrants and majority group members in the Netherlands.* Unpublished doctoral dissertation, Vrije Universiteit, Amsterdam, The Netherlands.

te Nijenhuis, J., & Van der Flier, H. (1997). Comparability of GATB scores for immigrant and majority group members: Some Dutch findings. *Journal of Applied Psychology, 82,* 675–687.

te Nijenhuis, J., & Van der Flier, H. (1999). Bias research in the Netherlands: Review and implications. *European Journal of Psychological Assessment, 15,* 165–175.

te Nijenhuis, J., Evers, A., & Mur, J. (2000). The validity of the Differential Aptitude Test for the assessment of immigrant children. *Educational Psychology, 20,* 99–115.

Thorndike, R.L. (1985). The central role of general ability in prediction. *Multivariate Behavioral Research, 20,* 241–254.

van Langen, A., & Jungbluth, P. (1990). *Onderwijskansen van migranten; de rol van sociaal-economische en culturele factoren* [Educational opportunities of immigrants: The role of socioeconomic and cultural factors]. Lisse, The Netherlands: Swets.

Verweij, A.O., Latuheru, E.J., Rodenburg, A.M., & Wijers, Y.M.R. (1998). *Jaarboek 1997 Grote-Stedenbeleid. Deel 2: Statistische gegevens* [Yearbook 1997 large-cities policy. Part 2: Statistical data]. Rotterdam, The Netherlands: ISEO.

van de Vijver, F.J.R., & Bleichrodt, N. (2001). Conclusies [Conclusions]. In N. Bleichrodt & F.J.R. van de Vijver (Eds.), *Diagnostiek bij allochtonen: Mogelijkheden en beperkingen van psychologische tests* [Diagnosing immigrants: Possibilities and limitations of psychological tests] (pp. 237–243). Lisse, The Netherlands: Swets.

van de Vijver, F.J.R., & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263–280.

van de Vijver, F.J.R., & Willemse, G.R.W.M. (1991). Are reaction time tasks better suited for ethnic minorities than paper-and-pencil tests? In N. Bleichrodt & P.J.D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology* (pp. 450–464). Lisse, The Netherlands: Swets.

Jan te Nijenhuis
Social and Organizational Psychology
Leiden University
P.O.Box 9555
NL-2300 RB Leiden
The Netherlands
Tel. +31 71 527-3705
Fax +31 71 527-3619
E-mail JanteNijenhuis@planet.nl

Appendix

Table A1. The g values of the RAKIT subtests, estimated both by EQS and by taking the loadings of the subtests on the first unrotated principle factor of a principle axis factor analysis (PAF), and congruence indices between the two methods of estimation.

RAKIT Subtests	*g* EQS	*g* PAF
Closure	.38	.42
Exclusion	.58	.58
Memory span	.42	.42
Verbal meaning	.44	.44
Mazes	.39	.49
Analogies	.53	.49
Quantity	.65	.62
Disks	.49	.57
Learning names	.40	.42
Hidden figures	.59	.59
Idea production	.25	.35
Storytelling	.23	.25
Congruence coefficient	.99	.99

Note. g EQS = *g* value supplied by EQS; *g* PAF = *g* value estimated by taking the loading of the subtest on the first unrotated principle axis factor.