ELSEVIER

Intelligence 31 (2003) 107–121

INTELLIGENCE

Do age-group differences on mental tests imitate racial differences?

Arthur R. Jensen*

School of Education, University of California, Berkeley, CA 94720-1670, USA

Received 23 December 2000; received in revised form 27 March 2001; accepted 10 July 2001

Abstract

Previous studies have shown that the pattern of mean differences on various mental tests (and other psychometric features) between black (B) and white (W) children all of the same age is imitated by comparing racially homogeneous groups composed of all W or all B children that differ in chronological age (CA). When the younger/older CA ratio is between 0.80 and 0.90, raw score differences approximate the B–W mental age (MA) ratio of same-age B and W children on IQ tests. The typical W–B IQ difference is 10–20 points. The imitation of actual same-age W–B mean differences by different age groups of W (or B) has been observed in a number of psychometric characteristics besides total raw scores. Does the same "imitation" phenomenon occur for Spearman's hypothesis? This hypothesis states that the standardized mean W–B differences on various tests are directly related to the magnitudes of the tests' *g* loadings? In a battery of 17 diverse tests, the actual W–B comparisons strongly bear out Spearman's hypothesis. The comparisons of different age groups (of the same race) resemble the Spearman effect, but also clearly differ, as the least *g*-loaded tests show larger age-group differences than the corresponding differences between the age-matched W and B groups. The findings support Spearman's hypothesis that the W–B difference in test performance is predominantly a *g* difference rather than a unitary developmental difference affecting all factors in test performance.

Keywords: IQ; *g* factor; Race; Age; Spearman's hypothesis

* 30 Canyon View Drive, Orinda, CA 94563-1504, USA.
E-mail address: nesnejanda@aol.com (A.R. Jensen).

0160-2896/03/$ – see front matter © 2003 Elsevier Science Inc. All rights reserved.
PII: S0160-2896(01)00090-3

1. Introduction

Does the commonly observed white–black (W–B) population difference on various cognitive tests represent mainly a unitary or general difference in developmental growth rates in test performance? Or does it reflect more specifically a developmental growth-rate difference mainly in the *g* factor common to all tests of cognitive ability?

One way to investigate this question is by means of Spearman's hypothesis. According to this hypothesis, the variable standardized mean W–B differences on various tests are directly related to the tests' *g* loadings. This hypothesis has been substantiated by a significant positive correlation between the vector of mean W–B differences on various tests and the corresponding vector of those tests' *g* loadings—a phenomenon termed the *Spearman effect*. (Spearman's hypothesis, the methodology for testing it, and the results of 18 independent studies that show the reality of the Spearman effect at a high level of statistical significance are presented in Jensen, 1998a, pp. 369–402).

Hypothesis: The W–B difference in test performance is a developmental difference between the groups' mental growth curves from early childhood to maturity. Hence, it should be possible to imitate the Spearman effect by comparing groups of older and younger children of the same race. The *chronological ages* (CAs) of the younger/older groups must be in approximately the same ratio as the B–W ratio of mental ages (MAs) as estimated by standard IQ tests. The older/younger groups of the same race are termed *pseudo-race* groups.

The rationale for pseudo-race comparisons was first spelled out as a means for distinguishing between racial/cultural bias in mental tests and developmental differences in mental growth (Jensen, 1980, pp. 450–452). Item intercorrelations and other internal features of test performance that differ systematically between different racial/cultural groups are often imitated by racially/culturally homogeneous groups that differ in CA and therefore in absolute ability level. The nature of such differences therefore appears mainly developmental rather than cultural.

The question is: Do the pseudo-race groups imitate the Spearman effect observed in actual W–B comparisons? If so, the vector of pseudo-race group differences on various tests should be correlated with the tests' *g* loadings to about the same degree as actual W–B differences for groups of the same age.

The possibility that the Spearman effect might be imitated by pseudo-race groups was suggested by previous findings that W–B differences in certain other features of test performance closely imitate differences between older and younger age groups of the same race. The modal choice of error distractors on the multiple-choice Raven matrices test differs between B and W children of the same age, but the B choice of error distractors closely matches the choice made by W children approximately two years younger (e.g., sixth grade B's and fourth grade W's) (Jensen, 1980, p. 584). On a continuum of MA, some error distractors are intellectually more sophisticated than others.

Similarly, B preschoolers show a higher percentage of less 'mature' preferential responses in a test that calls for matching cards on the basis of color, form, number, and size. Also, on Gesell's developmental tests of figure copying, on free drawings, and on developmental Piagetian tests of conservation of number and volume and the horizontality of water level in a tilted bottle, B

children (under 8 years of age) perform the same as W children who are about one year younger (review and references in Jensen, 1980, pp. 662–677). Because these *g*-loaded tests were examined in isolation from other tests, they do not throw much light on the question of interest in this study. However, some other findings in the literature are pertinent.

1.1. Preliminary observations

A child's "MA" derived from a typical psychometric battery, such as the Stanford–Binet, does not represent a unitary or factorially homogeneous dimension of growth in mental capability (Siegler & Richards, 1982). The *g* factor does not account for all of the common factor variance in a diverse battery of tests at any given CA. Early on, Spearman (1927, p. 370) discovered that the non-*g* or specific variance (*s*) in tests derives in part from the influence of sensory and motor demands, and tests that depend more on these abilities than on relation eduction have growth trajectories from childhood to maturity that are different from highly *g*-loaded tests involving relation eduction. He also noted that the same thing is true for tests that only involve simple retentiveness, or reproduction, which depend hardly at all on relation eduction or inference for either acquisition of information or its long-term retention and retrieval (Spearman, 1927, Chapter XVI). Memorizing paired-associates nonsense syllables and defining the meaning of words in a vocabulary test differ markedly in their proportions of *g* and *s* variance. Vocabulary is normally acquired by inference or eduction of word meanings from the contexts in which they are encountered, but memorizing paired-associate nonsense syllables requires mainly "retentivity" (Spearman's term), which reflects little if any relation eduction. A broad generalization, though still hypothetical, is that *the similarity of various tests' mental growth curves is a function of the tests' g-loading*.

1.1.1. The Case et al. study

The *g* variance in a test battery can be divided into the part associated with mental growth during the childhood years and a residual part independent of these age effects. This was done in a study by Case, Demetriou, Platsidou, and Kazi (2001), although the point I am making with their data (shown in their Table 3) was not the purpose of their study. They gave a battery of 23 diverse cognitive tests to children ranging from 7 to 10 years of age. A hierarchical factor analysis with Schmid–Leiman orthogonalization resulted in five first-order factors (with eigenvalues > 1), labeled *spatial*, *numerical*, *social/verbal*, *logic/analysis*, and *social causality* (i.e., an aspect of social intelligence, inferring possible causes of various personal or social interactions) with factor intercorrelations yielding *g* as a second-order factor. The factor analyses were based both on the raw correlation matrix for all ages combined and on the correlation matrix after CA was partialed out.

The averages of all the tests' *g* loadings *before* and *after* age was partialed out were .55 and .35, respectively. This is a much larger difference than was found for all the residualized first-order factor loadings, which averaged .32 *before* and .37 *after* age was partialed out. The proportion of the total variance in the whole test battery attributable to *g* (with age-in) is .30; with age-out it is *reduced* to .12. The positive difference (+.18) may be considered a *developmental* component of the *g* variance. Removing this developmental component by partialing out age

from the correlation matrix has the opposite effect on the proportion of common factor variance due to non-*g*, which is *increased* slightly from .10 (for age-in) to .14 (for age-out). The correlation between the vectors of *g* loadings for age-in and age-out is only .61, while the correlation between the age-in and age-out vectors for non-*g* loadings is .83, further indicating that *g* is more reflective of age or developmental effects than is non-*g*. The most *g*-loaded tests also show the largest decrements with age partialed out ($g \times g$-decrement rank-order $r = +.50$, $P < .01$). The essential point here is that various tests are differentially reflective of the developmental component of the *g* factor. As children grow up they acquire more specific bits of information and particular skills, but also their learning ability, or the speed and efficiency of their information-processing capacity, increases and is reflected by the *g* factor.

1.1.2. The Garrett et al. study

In their classic paper on the "age differentiation" hypothesis, Garrett, Bryan, and Pearl (1935) provide means, S.D.s, and correlation matrices for 10 diverse mental tests given to groups of school children of ages 9 and 12 years, separately for boys and girls, with nearly equal *n*'s in each subgroup, for a total $N = 421$. As the data for boys and girls are highly similar on all these variables, the sexes were combined to increase the reliability of the analysis. (Garrett's 15-year age group is not used because comparing it with the age 12 group is not statistically independent of the comparison made between ages 9 and 12, as the two comparisons would have one age group in common.) These pseudo-race groups, then, are used here to determine whether the Spearman effect shows up between groups differing in CA by correlating the vector of mean standardized differences between the two age groups on each of the 11 tests with the corresponding vector of those tests' *g* loadings. (The *g* vectors were averaged across the sexes and age groups, as their congruence coefficients are all above .95, indicating that they are highly similar.) The vectors of standardized differences and *g* loadings showed a Pearson *r* of +.48 and a Spearman rank correlation of +.51; with $n = 10$ values in each vector, both correlations are significant ($P < .05$). This correlation, however, is somewhat smaller than the correlation of about .60 typically found in the Spearman effect based on the test score differences between same-age W's and B's. This lower correlation in the study by Garrett et al. is largely the result of restricted variation among the tests' *g* loadings. Because Garrett et al.'s age groups differed by 3 years (ages 9 and 12), the mean standardized difference between the two age groups on the 10 subtests was 1.33*z*; this is a much larger difference than the subtest differences between W and B groups, which typically average 0.6*z* to 0.7*z* (Jensen, 1998, p. 382).

1.1.3. The British Ability Scales (BAS)

The Technical Manual (Elliott, 1983) for the BAS provides a wealth of raw score data for different age groups representative of the British population, of which only the normative data on 8- and 10-year-old boys are examined here. These age groups were selected because 8/10 is fairly close to the ratio of average mental-age levels of B and W 10-year-olds, respectively, with the W IQ = 100, the B IQ = 80, a difference equivalent to 1.3σ. The mean of the standardized differences between the groups, ages 10 and 8 years, on the 18 subtests, is 0.59, which is close to the average of the subtest differences in W–B (same-age) comparisons.

The BAS is one of the most exceptionally well constructed of modern psychometric tests, with 18 subtests of highly diverse abilities. The *g* factor of this battery is represented by the first principal component of the 18 subtests obtained for these age groups. The vector of subtests' *g* loadings and the vector of standardized mean differences between 10- and 8-year-olds' test scores are significantly correlated, with Pearson $r = .49$ and Spearman's rank correlation = .57, values similar to those found in the Garrett data, and somewhat below the correlation typically found for the Spearman effect in W–B comparisons. However, it is impossible to determine the cause of this lower correlation without a comparison group of B's on the BAS.

1.1.4. The Kaufman Assessment Battery for Children (K-ABC)

The *K-ABC Interpretive Manual* (Kaufman & Kaufman, 1983) provides raw score data that can be treated in the same way as was done with the BAS. Two sets of comparisons were made with the K-ABC national standardization sample, based on test score differences between ages 8 and 10 years and between ages 9 and 11 years. The first principal factor of the 13 K-ABC subtests is used as the measure of *g*. The vector of the 13 subtests' *g* loadings and the standardized mean difference between the groups of age 10 and age 8 are correlated by Pearson $r = .75$ and Spearman rank correlation = .79; with $n = 13$, both correlations are significant ($P < .01$). The corresponding correlations based on the groups of age 11 and age 9 are Pearson $r = .76$, Spearman $r_s = .73$. This is the only test for which the Spearman effect for pseudo-race groups can be directly compared with same-age W and B groups, as seen in the following study.

1.1.5. The Naglieri and Jensen study

Naglieri and Jensen (1987) compared groups of 86 W and 86 B children matched pairwise for age (mean = 10.75 years), school, sex, and socioeconomic status (SES). The W–B IQ difference was 0.77σ, equivalent to 12 IQ points. The *g* factor loadings of the 13 K-ABC subtests were here represented by the first principal factor of the national standardization data. The vector of subtests' *g* loadings and the vector of standardized mean differences between the W and B groups are correlated by Pearson $r = .74$, Spearman $r_s = .77$. These correlations are nearly the same as the corresponding correlations obtained on the K-ABC in the pseudo-race comparisons (ages 9–11 and 8–10). Close matching of the racial samples for educational background and SES evidently increased, or at least did not lessen, the Spearman effect, that is, the predictability of the W–B differences in test scores by the sizes of their *g* loadings.

2. Method

2.1. Subjects

Subjects were all pupils enrolled in regular classes in Grades 3 through 8 in a California school district. A pupil's precise age and race (non-Hispanic W or African-

Table 1
Factor loading of race (W–B) on the *g* factor (PF1) of 11 to 17 tests[a] with *g* estimated separately in the correlation matrix of each group (W or B) in Grades 3 through 8

Race/grade	*g* loading		Mean age (years)		Sample size	
	White	Black	White	Black	White	Black
W3 B3	.48	.56	8.27	7.93	157	122
W4 B4	.54	.61	9.35	9.24	113	129
W5 B5	.54	.68	10.24	10.60	144	132
W6 B6	.59	.55	11.33	11.25	131	124
W7 B7	.65	.79	12.45	11.91	156	167
W8 B8	.50	.52	13.27	13.28	176	181
Mean	.55	.62		Total	877	855

[a] 11 tests in Grade 3; 17 tests in Grades 4–6; 15 tests in Grades 7–8.

American) were obtained from the school records. The mean and standard deviation (S.D.) of the Lorge–Thorndike IQ of the pupils in this study are: W: $\bar{X}=106.5$, S.D. = 13.5; B: $\bar{X}=91.1$, S.D. = 11.8; the mean standardized W–B difference is 1.2σ. The overall national norms of the Lorge–Thorndike IQ are scaled to $\mu=100$, $\sigma=16$ (Lorge & Thorndike, 1957). The data were originally obtained for a study of the validity of various cognitive tests for predicting scholastic achievement in different racial/ethnic groups (reported in Jensen, 1974).

To ensure uniformity of procedures, all tests were administered to intact classes by specially trained testers, while the classroom teachers acted as monitors. Hispanic and Asian pupils were not included in the present data analyses. Sample sizes for B's and W's in each grade, totaling 1732 pupils, are given in Table 1.

2.2. Tests

A wide variety of both standardized cognitive ability tests and scholastic achievement tests were administered according to the instructions and time limits given in the test manuals. Also included were some specially devised tests of short-term memory, ability to listen and pay attention to oral instructions, and tests of speed and persistence intended to assess pupils' motivation to work attentively at the assigned task.

2.2.1. Lorge–Thorndike Verbal IQ and Nonverbal IQ (Levels 3 and 4)

Lorge–Thorndike Verbal IQ and Nonverbal IQ (Levels 3 and 4) are described in the test manual (Lorge & Thorndike, 1957) as measures of verbal and nonverbal reasoning. The verbal battery comprises subtests of sentence completion, verbal analysis, arithmetic reasoning, synonyms, and number series. The nonverbal battery comprises spatial relations, figure classification, figure synthesis, and pictorial and figural analogies. The machine scoring of the tests provided only total scores for Verbal and Nonverbal IQ. As the total verbal and nonverbal scores are factorially complex, they would be located in terms of Carroll's (1993)

three-stratum hierarchical factor model in the second stratum, and both are highly loaded on the third stratum, i.e., the *g* factor. The verbal and nonverbal IQ are correlated at +.70 in Grade 4 and slightly more at higher grade levels.

2.2.2. Raven's Progressive Matrices

Raven's Progressive Matrices is a multiple-choice measure of nonverbal inductive and deductive reasoning based on visual figures of graded complexity. It is a second-stratum factor (fluid intelligence, *Gf*) in Carroll's model. The Colored Progressive Matrices (36 items) was used in Grades 3–6, the Standard Progressive Matrices (60 items) in Grades 7–8.

2.2.3. Figure Copying

This is a nonspeeded test of developmental age, composed of 10 geometric figures originally introduced by Binet and used by Piaget and Gesell to measure developmental (mental) age (Ilg & Ames, 1964; Jensen, 1980, pp. 165, 662–665). The test involves analytic spatial ability and is moderately *g*-loaded. The subject views a figure on the top half of each page of the test booklet and simply has to copy it as accurately as possible. Because the figure to be copied is constantly in the subject's view, accurate performance is not dependent on memory of the figure. Each of the 10 figures, ordered on a Guttman-like scale of difficulty, is presented singly on a separate page. Each copied figure is scored on a three-point scale (0–2) for its degree of conformity to the essential features of the target figure. This test was used only in Grades 3–6, because at higher grades there is a ceiling effect and restricted variance.

2.2.4. Listening–Attention

Listening–Attention is a test presented by tape recorder that assesses the child's ability to listen and pay attention to simple verbal instructions for about 5 min. Accuracy of performance depends mainly on the subject's ability to listen to directions and pay close attention continuously for several minutes. Although performance requires knowing the numbers from 0 to 9, this test makes virtually no other demand on knowledge, memory, or reasoning. As purely a measure of listening and sustained attention, its very simple cognitive demands make it difficult to classify in Carroll's factor hierarchy, but it seems best to belong under two of the first-stratum factors in the Carroll model: Factor AC (p. 547, attention, carefulness, counting, number checking) and Factor LS (p. 147, listening). Answer sheets consist of columns of 10 pairs of random but nonmatching single digits presented in five columns on two pages; each column is headed by a capital letter in alphabetical order. A tape-recorded male voice names the letter of the column to be attended to, then names, at a 2-s rate, one digit in each pair. Subjects are instructed to cross-out the named digit. The total score is the number of correct cross-outs.

2.2.5. Memory for Numbers

This is a test of digit span or short-term auditory memory, consisting of three subtests, each of which consists of six series of digits going from four digits in a series up to nine digits. It is a first-stratum factor (MS, p. 256) in Carroll's model. It is presented on a tape recording on

which the digits are spoken clearly by a male voice at the rate of precisely one digit per second. At the conclusion of each series, signaled by a 'bong' sound, the subject writes on a specially printed form as many digits as he or she can recall in the correct order. Each subtest is preceded by three practice trials. The three subtests are: (1) *Immediate Recall*, in which the 'bong' sounds 1 s after the last digit is heard; (2) *Delayed Recall*, in which the response is not written until signaled by the 'bong' 10 s after the last digit was spoken; and (3) *Repeated Series*, in which the same digit series is presented three times in succession, each repetition separated by a tone of 1-s duration, then is immediately recalled after the 'bong' at the end of the third presentation. The score for each subtest is the number of digits recalled in the correct serial position.

2.2.6. Making X's

Making X's assesses test-taking motivation, indicating the subject's willingness to comply with instructions in a group testing situation and to concentrate for a brief period of time. In Carroll's model (under the name "writing X's") it is a first-stratum factor (p. 537, psychomotor ability). The subject is simply asked to make X's in a series of squares for a period of 90 s. The test is given twice, under different instructions: (1) *nonmotivating* instructions in which nothing is said about speed, and (2) motivating instructions, in which subjects are told to work as fast as they can and try to exceed the number of X's they made on the first trial. The subject's score is the number of X's written in 90 s.

2.2.7. Stanford Achievement Tests (Kelley, Madden, Gardner, & Rudman, 1964)

Primary, intermediate, and advanced forms are nationally standardized tests of scholastic achievement, with separately scored subtests for *Vocabulary*, *Paragraph Comprehension*, *Spelling*, *Word Study Skills*, *Language Usage*, *Arithmetic Computation*, *Concepts*, and *Applications*.

3. Results

3.1. Spearman's hypothesis of the W–B difference

All analyses are based on the raw scores for each test. Standardized mean group differences were obtained by dividing the raw score mean group difference by the geometric mean of the raw score within-group S.D.s. First, it is essential to determine whether the present battery of tests shows the effect predicted by Spearman's hypothesis, viz., that the mean W–B difference in cognitive abilities is predominantly a difference in *g* and that therefore the variable magnitudes of the W–B differences on various cognitive tests are directly related to the tests' *g* loadings, which are here represented by the first principal factor of the test battery.

A test of Spearman's hypothesis requires that the *g* factors (in this case the first principal factor, PF1) extracted separately in the W and B samples represent the same factor. The

congruence coefficient between B's and W's in Grades 3 to 8 ranges between .94 and .98, averaging .97, indicating that the *g* factor is highly similar for both racial groups.

The dichotomous variable of race, with B and W quantified as 0 and 1, was entered into the correlation matrix for the whole test battery, separately for the correlation matrix of each racial group. The correlation of 'race' with each of the test variables, therefore, is a point-biserial correlation (r_{pbs}). A principal factor analysis was performed on the correlation matrices separately for W's and B's within each grade level from Grades 3 to 8. The point of interest is the loading of the race variable (W–B) on the *g* factor. This *g* factor loading provides a purer indicator of the racial difference in *g* than would a comparison of the groups' *g* factor scores, because these are always slightly contaminated by the non-*g* residue from lower-order factors and uniqueness.

Table 1 shows the results and also the mean ages and sample sizes of the W and B groups. The *g* loading of 'race' (equivalent to the point-biserial correlation of the W–B racial dichotomy with the *g* factor of the whole test battery within each racial group) averaged over Grades 3 to 8 is .55 for W's and .62 for B's. 'Race' had negligible loadings on all the other factors with eigenvalues >1. The magnitudes of the *g* loadings of race at all grade levels indicate standardized mean group differences larger than 1S.D. In terms of *g* factor scores (standardized to $\mu = 0$, $\sigma = 1$), the mean W–B difference is 1.25σ. (Within the range of $r_{pbs} < \pm .60$, the r_{pbs} is virtually a linear transformation of the standardized difference between groups. The transformation equation is given in Jensen, 1980, p. 122, Fig. 4N.3. Given equal S.D.s within each group, for example, a 1S.D. mean difference between the groups would amount to a point-biserial *r* of .45.)

The second test of Spearman's hypothesis is the correlation between the column vector of standardized mean W–B differences on the various tests and the corresponding column vector of those tests' *g* loadings. These correlations, at each grade level, are shown in Table 2 as the Pearson correlation coefficient (*r*), the Spearman rank-order correlation (r_s), and the Pearson *r* with the column vector of the tests' KR-20 reliability coefficients partialed out (r_c), which, if not markedly different from the other (zero order) correlations, indicates that the correlations are not mainly an artifact of variation in the tests' reliability coefficients. Spearman's hypothesis was borne out in every grade.

Table 2
Correlation between tests' *g* loadings and the standardized mean W–B difference in Grades 3 to 8, showing the Pearson correlation (*r*), Spearman's rank-order correlation (r_s), and Pearson *r* with test reliability partialed out (r_c)

Grade	*r*	r_s	r_c
3	.54	.84	.65
4	.77	.66	.72
5	.75	.71	.76
6	.59	.72	.70
7	.79	.80	.74
8	.77	.77	.79
Mean	.70	.75	.73

3.2. A simulated Spearman effect in pseudo-race groups

Pseudo-race comparisons are based on groups of children all of the same race but which differ in CA. The CA ratio of the younger group's CA to the older group's CA approximates values between 0.80 and 0.90. Multiplied by 100, this range of values is typical of the mean B IQ in nationally normed data on most tests of general cognitive ability. Up to about 15 years of CA, on average, children's MA (MA) increases linearly as a function of CA, so that in this age range IQ can be conceived as the average ratio of MA/CA × 100. By selecting groups within this range that differ by two grade levels (or by approximately 2 years in age), the ratio of younger/older age falls in the range of 0.80 to 0.90, which is also the ratio of B–W MA.

Grade levels separated by 2 years were compared only if the same tests were used at both grade levels, so the comparisons are based on the same tests that were used in the previous analyses of the actual Spearman's hypothesis based on B and W groups. This condition results in comparisons only of Grades 4 and 6 and Grades 5 and 7. The same analyses used in the actual racial comparisons were performed on the different age groups of the same race.

The *g* factor is highly similar across the compared age groups, as attested by congruence coefficients ranging from .94 to .99, averaging .96.

Table 3 (analogous to Table 1) shows the *g* loadings of the compared age groups when older/younger age is included as a dichotomous variable in the factor analysis of the correlation matrix of all the tests. Also shown is the exact CA ratio (younger/older) of the compared groups. These *g* loadings indicate that age differences in cognitive capabilities mainly reflect growth in *g* itself more than in any of the residual group factors. The *g* loadings in Table 3 are somewhat smaller than the analogous loadings in Table 1; that is, the same-age W–B differences in *g* are slightly larger than the differences between groups of same-race children that are separated by about 2 years in CA.

Table 4 (analogous to Table 2) shows the correlations between tests' *g* loadings and the size of the standardized mean differences between the age groups. Nearly all these correlations are considerably smaller than those for the actual race comparisons shown in Table 2. Examination of the tests that least conform to the hypothesized Spearman effect shows that they are the least *g*-loaded and evidently have quite different growth trajectories

Table 3
Factor loadings of age (higher/lower grade) on the *g* factor (PF1) in pseudo-race groups with *g* estimated separately in the correlation matrix of each grade and the age ratio (younger/older) for each comparison

g loading			
Race/grade	Higher	Lower	Age ratio
W6 W4	.66	.64	.83
B6 B4	.59	.69	.82
W7 W5	.45	.50	.82
B7 B5	.35	.39	.89
Mean	.51	.55	.84

The *n* for each race/grade group is shown in Table 1.

Table 4
Pseudo-race groups: correlation between tests' *g* loadings and the standardized mean difference between higher and lower grades (6–4 and 7–5) within W and B groups, showing the Pearson correlation (*r*), Spearman's rank-order correlation (r_s), and Pearson *r* with test reliability partialed out (r_c)

Race/grade	*r*	r_s	r_c
W6 W4	.45	.43	.37
B6 B4	.73	.75	.68
W7 W5	.42	.19	.36
B7 B5	.19	.19	.27
Mean	.45	.39	.42

than the more *g*-loaded tests. W–B differences on these low-*g* tests, especially the memory span tests, are much smaller than the age-group differences.

Table 5 helps elucidate the discrepancies between the racial and pseudo-racial comparisons. It shows the *g* loadings averaged over racial and age groups and the standardized mean differences (effect size) averaged over all comparisons of the same type: (A) same-age W–B, (B) same-race older/younger age, and (C) older B–younger W. The (C) comparison is

Table 5
Mean *g* loadings for all grades and mean effect sizes for three types of contrast: (A) W–B same age; (B) same-race older–younger age; (C) older B–younger W[a]

	g	Effect size		
		A	B	C
Verbal IQ	.796	1.175	1.26	0.160
Nonverbal IQ	.728	1.220	1.513	0.301
Raven Matrices	.509	0.965	0.516	−0.091
Figure Copying	.405	0.947	1.100	–[b]
Listening–Attention	.048	0.225	0.031	0.010
Digit Span (Immediate)	.402	0.209	0.573	0.431
Digit Span (Repeated)	.409	0.051	0.704	0.732
Digit Span (Delayed Recall)	.424	0.167	0.681	0.543
Making X's 1	.245	0.194	0.222	−0.116
Making X's 2	.316	0.178	0.439	0.021
Vocabulary	.785	0.957	–[b]	–[b]
Paragraph Comprehension	.774	1.103	1.065	−0.588
Spelling	.746	0.702	0.48	−0.257
Word Study	.758	0.910	0.517	–[b]
Language	.805	0.974	0.660	−0.313
Arithmetic Computation	.571	0.733	0.672	−0.310
Arithmetic Concepts	.634	0.934	0.714	−0.139
Arithmetic Applications	.666	0.910	0.538	−0.495
Mean	.557	0.697	0.689	0.089
S.D.	.223	0.404	0.371	0.563

[a] Contrasted groups differ by two grade levels.
[b] This test not included at one of the contrasted grade levels.

Table 6
Correlations between the columns *g*, A, B, and C in Table 5: Pearson *r* (above diagonal), Spearman r_s (below diagonal), and partial r_c (in parentheses) controlling for test reliability

Column	*g*	A	B	C
g	–	.80 (.82)	.59 (.55)	− .38 (− .44)
A	.71	–	.65 (.72)	− .50 (− .52)
B	.46	.54	–	.12 (.04)
C	− .43	− .40	.16	–

intended to indicate whether B children of a given age perform on the tests equally with W children who are about 2 years younger.

The correlations between the column vectors in Table 5 are shown in Table 6, which gives an overall picture that is clearly consistent with the results of the previous separate analyses based on different race/grade groups, shown in Tables 2–4. The largest correlation ($g \times A$) is for the actual race comparison (i.e., Spearman's hypothesis). The correlation between the vector of *g* loadings and the vector of pseudo-race (i.e., older/younger) differences ($g \times B$) is considerably lower. Yet the correlations between vectors of effect sizes for both the race and the pseudo-race comparisons ($A \times B$) are fairly substantial, showing that a 2-year age difference between same-race groups mimics the race difference between same-age groups to a considerable extent. The correlations involving C (older B's–younger W's) are mainly negative, indicating that W's still score somewhat higher than B's who are 2 years older and do so largely on the more highly *g*-loaded tests. Older B's exceed younger W's most on the digit span memory tests, with an average effect size of + 0.57.

4. Discussion

This study makes several points of theoretical and practical interest.

(1) Increases in test scores on various mental tests across age during childhood quite strongly reflect growth in the *g* component of various abilities, suggesting that increments in test performance with increasing age result not just from increases in specific types of acquired knowledge and skills, but in a general factor, *g*, that accounts for the intercorrelations of a wide variety of knowledge and skills and is probably the chief agent accounting for individual differences in their rate of acquisition. The *g* component not only differs across various psychometric tests but also varies, although to a lesser degree, on the same tests at different points on the mental growth trajectory.

(2) The data provide another demonstration of Spearman's hypothesis. The findings in each of the Grades 3 through 8 replicate those of 19 other studies based on a wide variety of tests, with B and W samples ranging in age from preschool to middle age (reviewed in Jensen, 1998a, Chap. 11; Nyborg & Jensen, 2000). No battery of diverse tests administered to representative samples of the W and B populations has yet been found that does not confirm Spearman's hypothesis. Combining all the previous independent studies of the Spearman effect, the regression of the mean W–B standardized differences on the *g* loadings of the tests

predicts a W–B difference of 1.31σ for a hypothetical pure test of *g*, i.e., a *g* loading of unity (Jensen, 1998a, p. 377). For the present study, the regression of the W–B differences on the tests' *g* loadings (i.e., the regression of column A on column *g* in Table 5) predicts a mean W–B difference of 1.34σ on a hypothetical pure test of *g*. The corresponding pseudo-race comparison (regression of column B on column *g* in Table 5) predicts a mean difference between the older and younger groups of only 1.14σ on a hypothetical pure test of *g*.

(3) An attempt was made to determine whether the Spearman effect is mimicked by pseudo-race groups composed of same-race children who were separated by two grades in elementary school, thereby differing in age, on average, by about 2 years. Although the attempted imitation gave some results resembling those obtained with the actual racial comparison of same-age W and B groups, in certain details it falls short of an accurate imitation, indicating that the B–W differences in performance on a variety of cognitive tests cannot be viewed as a uniform difference in mental growth rates for all cognitive abilities. The racial difference in mental growth trajectories is greater to the degree that the tests are more *g*-loaded. The comparison of same-race groups differing 2 years in age showed much larger differences on those tests with smaller *g* loading, particularly tests of short-term memory, than were seen in the W–B comparisons. This is a developmental corollary of Spearman's hypothesis.

Another hypothesis, as yet not tested, is whether a comparison of two groups of same-age and same-race individuals that have IQ distributions typical of the W and B populations will yield a near-perfect imitation of the Spearman effect. Ideally, to control cultural, socio-economic, and other between-family sources of variance, such a study would be done by allocating each member of fraternal twin pairs to the upper or the lower IQ group in such a way as to obtain a 1S.D. difference between the two groups.

If the results closely resemble those found in the W–B comparisons, it would prove that Spearman's hypothesis does not apply uniquely to a particular racial comparison but predicts the same phenomenon for any two groups, regardless of race, that differ in IQ. The Spearman effect has been found in comparisons between the majority population of the Netherlands and certain first-generation immigrant groups of Surinamese, Netherlands Antilleans, North Africans, and Turks (te Nijenhuis & van der Flier, 1997, 1999). Regarding explanations (e.g., linguistic/cultural test bias) for the mean test score differences between the majority and minority groups, the authors concluded, "Of the different explanations for the differences in means between the groups, Spearman's hypothesis received the strongest support. For all four of the immigrant groups, *g* is the predominant factor accounting for differences between the majority group and the immigrant groups" (p. 686). (Further examples of Spearman's hypothesis with other tests and immigrant groups in the Netherlands are in te Nijenhuis, 1997; te Nijenhuis, Evers, & Mur, 2000).

(4) The fact that the test battery used in this study contains several subtests of the Stanford Achievement Test that clearly measure scholastic attainment does not imply that the general factor extracted from the total battery is not a valid representation of Spearman's *g*. In a school population exposed to the same educational program for the same period of time, individual differences in scholastic knowledge, cognitive skills, and reasoning based thereon should, in theory and in fact, be highly *g*-loaded. Both the learning and use of symbol systems

in the acquisition of knowledge and their applications in reading comprehension, quantitative reasoning, and problem solving are well established hallmarks of psychometric *g* (Jensen, 1989, 1993). Nonverbal tests without any scholastic content also show substantial loadings on the first principal factor (PF1) used to represent *g* in the present battery.

Because *g* is the chief source of both individual and group differences in scholastic achievement, the main practical implication of the present finding is this: In elementary school the overall mean W–B difference in the *g* component of scholastic performance is equivalent to a difference of about 2 years of CA within either racial group. This raises the old-fashioned but still meaningful concept of *readiness* for various types of scholastic learning. A pupil's level of *g* at a given age predicts the pupil's rate of learning and ability to grasp increasingly more complex concepts that determine a pupil's progress and self-perception of success in scholastic performance.

The wide range of differences in readiness during the most critical period of school learning, both between and within racial groups, underlines the desirability of varying the introduction and pacing of instruction in basic school subjects so as fully to allow for individual differences in absolute *g* level regardless of pupils' CA. As I have explained elsewhere (Jensen, 1998b), this could probably be achieved most effectively by eschewing homogeneous ability grouping of entire classes at the elementary level. A complex combination of computerized instructional programs that automatically keep a running record of each pupil's progress could be instituted, along with the classroom teacher's interventions with individuals and with small ad hoc ability groups as may be called for by their particular instructional needs.

References

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

Case, R., Demetriou, A., Platsidou, M., & Kazi, S. (2001). Integrating concepts and tests of intelligence from the differential and developmental traditions. *Intelligence*, *29*, 307–336.

Elliott, C. D. (1983). *British Ability Scales, manual 2, technical handbook*. Windsor, Berks, UK: National Foundation for Educational Research, Nelson Publishing.

Garrett, H. E., Bryan, A. I., & Perl, R. (1935). The age factor in mental organization. *Archives of Psychology*, (176), 1–31.

Ilg, F. L., & Ames, L. B. (1964). *School readiness*. New York: Harper & Row.

Jensen, A. R. (1974). Equality for minorities. In H. J. Walberg (Ed.), *Evaluating educational performance* (pp. 175–222). Berkeley, CA: McCutchen.

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

Jensen, A. R. (1989). The relationship between learning and intelligence. *Learning and Individual Differences*, *1*, 37–62.

Jensen, A. R. (1993). Psychometric *g* and achievement. In B. R. Gifford (Ed.), *Policy perspectives on educational testing* (pp. 117–227). Norwell, MA: Kluwer Academic Publishing.

Jensen, A. R. (1998a). *The g factor*. Westport, CT: Praeger.

Jensen, A. R. (1998b). The *g* factor in the design of education. In R. J. Sternberg, & W. M. Williams (Eds.), *Intelligence, instruction, and assessment: theory into practice* (pp. 111–131). Mahwah, NJ: Erlbaum.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children: interpretive manual*. Circle Pines, MN: American Guidance Service.

Kelley, T. L., Madden, M., Gardner, E. F., & Rudman, H. C. (1964). *Stanford Achievement Test*. New York: Harcourt, Brace and World.

Lorge, I., & Thorndike, R. L. (1957). *The Lorge–Thorndike Intelligence Tests, Levels 2 to 4*. Boston: Houghton Mifflin.

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

Nyborg, H., & Jensen, A. R. (2000). Black–white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Personality and Individual Differences*, *28*, 593–599.

Siegler, R. S., & Richards, D. D. (1982). The development of intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 897–971). Cambridge, UK: Cambridge University Press.

Spearman, E. (1927). *The abilities of man: their nature and measurement*. New York: Macmillan.

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

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

te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants 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.