The Journal of Special Education
http://sed.sagepub.com/

The Black-White Difference On the K-ABC: Implications for Future Tests

Arthur R. Jensen

J Spec Educ 1984 18: 377
DOI: 10.1177/002246698401800316

The online version of this article can be found at:
http://sed.sagepub.com/content/18/3/377

Published by:
Hammill Institute on Disabilities

HAMMILL INSTITUTE ON DISABILITIES

and

SAGE

http://www.sagepublications.com

Additional services and information for *The Journal of Special Education* can be found at:

Email Alerts: http://sed.sagepub.com/cgi/alerts

Subscriptions: http://sed.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 1, 1984

What is This?

THE BLACK-WHITE DIFFERENCE ON THE K-ABC: IMPLICATIONS FOR FUTURE TESTS

Arthur R. Jensen, Ph.D.
School of Education, University of California, Berkeley

Claims that the Kaufman Assessment Battery for Children (K-ABC) is less culturally biased than other standard tests of intelligence and therefore shows a much smaller average difference between black and white children are critically examined in terms of the psychometric properties of the K-ABC. It is concluded that the apparently reduced difference between black and white samples, as compared with the one standard deviation difference typically found on other IQ tests, is not the result of greater validity or of less biased measurement of children's intelligence by the K-ABC. The diminished black-white difference on the K-ABC seems to be largely the result of psychometric and statistical artifacts: lower *g* loadings of the mental processing scales and greater heterogeneity of the standardization sample, which causes mean group differences to be smaller when they are expressed in standard score units. The general factor measured by the K-ABC is essentially the same *g* as that of the Stanford-Binet and Wechsler scales. But the K-ABC yields a more diluted and less valid measure of *g* than do the other tests. The K-ABC factors of successive and simultaneous mental processing, independent of the *g* factor, constitute only a small fraction of the total variance in K-ABC scores, and the predictive validity of these small factors per se is probably nil. The present criticisms of the K-ABC suggest new means for improving the design of future tests for measuring general intelligence.

It is a rare event when the publication of a new psychological test becomes a news item in the popular media. The first flurry of media news about the Kaufman Assessment Battery for Children (K-ABC), as I recall, highlighted two main claims about this new test: (1) that the K-ABC represents a radically new departure from the traditional IQ tests, and (2) that it manifests an average black-white difference only about half as large as that generally found with other tests. Having consistently promoted the idea that IQ tests are biased against minorities, the popular media were now quick to conclude, on the basis of purportedly smaller black-white differences, that the new K-ABC must surely be less culturally biased than such prevailing standardized tests as the Stanford-Binet and the Wechsler Intelligence Scale for Children–Revised (WISC-R). The popular consensus appeared to be that the K-ABC provided conceptually different and better measurement of mental ability than any previous IQ test—better not only for America's minority children, but for *all* American children.

These claims, if true, would be important news indeed. Like many other psychometricians and differential psychologists, undoubtedly, I, too, was eager to examine the empirical basis for these claims. The present paper comprises the results of my examination. I am not attempting here a comprehensive critique of the K-ABC. Rather, all present observations and analyses will focus directly upon the

0022-4669/84/1803-0016$03.00/0

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

effort to understand why the K-ABC might show a smaller average black-white difference than is generally found with other cognitive tests. This effort is necessarily limited by the information available to me in the *K-ABC Interpretive Manual* (Kaufman & Kaufman, 1983). Consequently, definitive answers to certain crucial questions will have to await new evidence, as I shall indicate.

PSYCHOMETRIC PROPERTIES OF THE K-ABC

A careful study of the *K-ABC Interpretive Manual* itself (henceforth abbreviated *IM*) provides sufficient grounds for disclaiming that the K-ABC is a radical departure from the Binet test and its well known descendants, or that the K-ABC is in any way more advanced or more sophisticated than the older tests, in terms of psychometric theory and technology. This is rather a disappointment—that a brand new test, on which so much effort and resources have obviously been lavished, should be fundamentally so close in nature to the old Simon-Binet test of 1905! Calculations of split-half reliabilities, scaled scores, percentile ranks, confidence bands, varimax factors, and the like—all of which were unknown to Binet, of course—are useful advances, to be sure; but these are essentially superficial psychometric trimmings, and, as such, do not mitigate the basic criticisms of the Binet model. Conceptually and psychometrically, this bright new K-ABC invites the same fundamental objections as the original Binet. But the K-ABC is not alone in this, of course. The same thing can be said of all the present-day Binet descendants, for example, the Stanford-Binet and the various Wechsler tests. These tests and a great many others in the same tradition, including now the K-ABC, are all basically kith and kin, not identical superficially, of course, but closely inbred siblings nevertheless. It is probably too much to hope that psychology will be spared any new additions to this family in the future.

All tests constructed in the tradition established at the beginning of this century by Binet are subject to a major theoretical liability, namely the arbitrary selection and weighting of items or subtests making up the composite score. The selection and weighting procedures used in such tests are arbitrary in the sense that they are not objectively determined in accord with any theoretical construct of what the test purports to measure, but are arrived at by subjective, armchair "psychologizing." As an alternative method, items and weights are sometimes selected and derived by the purely empirical means of maximizing the test's correlation with some arbitrary criterion, such as school grades or job performance ratings. In this case, no question of theory is involved; the multiple regression equation provides the optimal weights. This raw empiricism of multiple regression methodology works satisfactorily and is defensible on practical grounds provided the only issue of concern is the prediction of a clearly specified, limited, and objectively measurable criterion. Such an approach is hardly applicable to the measurement of a hypothetical construct such as intelligence, however.

The very real theoretical deficiencies of the Binet model of test construction, inherited by the Wechsler tests, and now by the K-ABC, were cogently explicated at least as early as 1927, by Charles Spearman, in Chapter V of *The Abilities of Man*. This brilliant chapter, published 57 years ago, contains a most pertinent

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

critique of the psychometric model of the K-ABC. Spearman's arguments have never been contradicted by later psychometricians. In its claim as a measure of intelligence (*IM*, p. 1), the K-ABC must bear the full brunt of Spearman's criticism. One possible escape would be to disavow the K-ABC's claim to measure intelligence, and defend only its measurement of sequential and simultaneous processing capacities, but this would itself invite other lines of criticism. The only other "out," conceptually, is the easy, but wholly unsatisfactory, tack of saying merely that intelligence is whatever any "intelligence test" measures. But unless one can go beyond such tautology, there can be no theoretical basis for deciding whether any one test is a better or worse measure of intelligence than any other. Our choice of tests, in this case, must be based purely upon their superficial features, the attractiveness of their packaging, or the advertising hype in our professional journals.

The basic problem pointed out by Spearman, briefly, is this: If we arbitrarily, or subjectively, make up a collection of diverse test items (or subtests of homogeneous item types), with a subject's responses to each of these units scorable as right or wrong (e.g., 1 or 0), the scores on the various units being added or averaged to obtain a total or composite score, who is to say (1) whether the items or subtests which make up the whole test are a good (or a bad) collection of vehicles for measuring intelligence? and (2) whether the various parts of the tests have (or have not) been given the right weights in arriving at the total score? Let us not be distracted by the claim that the construction of the K-ABC was guided by a psychological or neurological theory of sequential and simultaneous mental processing, or by the demonstration that factor analysis distinguishes between the sequential and simultaneous tests. These features of the K-ABC do not overcome the two most basic points of Spearman's complaint. For we still have no way of knowing whether (1) the particular collection of tests of sequential and simultaneous processing, or the factors of sequential and simultaneous processing themselves, are good vehicles for the measurement of intelligence; or (2) whether the sum of three tests of sequential processing and five tests of simultaneous processing should arbitrarily have weights in the ratio of 3 to 5 in determining the total mental processing score. It is not any particular set of weights, equal or otherwise, that is objectionable, but the fact that there is no apparent theoretical basis for these weights. Does sequential processing, for example, constitute three-fifths of mental ability, or three-fourths, or one-half? Back to Spearman's Chapter V! Is there any basis for believing that this K-ABC Mental Processing Composite (MPC), which is scaled like IQ, with mean = 100, *SD* = 15, is in any way preferable to the Stanford-Binet IQ or to the Wechsler Full Scale IQ? In fact, it appears that the chief liabilities of the Binet model have fallen even more heavily upon the K-ABC than upon either the Stanford-Binet or the Wechsler scales, mainly because of the arbitrary weights given to the different parts of the K-ABC in the Mental Processing Composite. In this respect, Binet and Wechsler were simply luckier, not better in principle, however. The Stanford-Binet yields only a single, highly *g*-loaded score, and the verbal and performance IQs of the Wechsler tests are each based on five or six subtests and have large and highly comparable correlations with *g*.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

Because group differences, as between blacks and whites, vary on different subtests, the size of the group difference in total score will depend upon just how the subtests are weighted in the composite. Unless there is some theoretically defensible rationale to justify the weights given to the various subtests, therefore, it is possible to manufacture tests "to order," tests that can show almost any size of mean black-white difference, at least within broad limits. The WISC-R can serve as an example. In the national standardization sample, the mean black-white difference on Full Scale IQ is 15.83 IQ points, or 1.14 σ, that is, slightly more than one standard deviation. (Note that here and throughout this article, σ signifies the mean group difference divided by the square root of the N-weighted mean within-groups variance.) When factor analyzed, the WISC-R yields three significant factors: a general factor, g, a verbal factor, V, a performance (or spatial) factor, P, and a short-term memory factor, M (Jensen & Reynolds, 1982). As the WISC-R is normally scored, the g factor accounts for about 30% of the total variance of Full Scale IQ, and each of the three other factors accounts for about 5% of the variance. Yet if we calculate factor scores for every subject in the standardization sample and obtain the average of the four factor scores for each subject, giving each of the four factor scores equal weight, to obtain a composite score, the mean black-white difference on the composite becomes only 0.305 σ (or 4.57 IQ points), a shrinkage of 73%. Similarly, without resorting to factor analysis, one can make up a special subscale of the WISC-R consisting of Digit Span, Coding, and Tapping (a subtest included in the standardization, but eliminated in the published version of the WISC-R), which has a mean black-white difference of 0.37 σ. Averaging this difference with the mean of the black-white differences on the remaining ten other subtests of the WISC-R yields a composite black-white difference of 0.59 σ, or 8.80 IQ points, a shrinkage of 48%. An effect similar to this kind of arbitrary weighting is seen in the black-white difference on the K-ABC Mental Processing Composite. In the school-age standardization sample, for example, the *three* sequential processing tests (Hand Movements, Number Recall, and Word Order) show an average black-white difference which is only about one-half as large as the difference on the *five* simultaneous processing tests. But the sequential score and the simultaneous score are given weights in the ratio of 3 to 5 in the Mental Processing Composite, which is interpreted as the "measurement of total intelligence in the assessment battery" (*IM*, p. 31).

It must also be kept in mind that the total variance on a test, and the size of the black-white difference (or any other kind of group difference) is related not only to the differences on each of the subtests separately, but to all the *covariances* among the subtests as well. In all good tests, it is the sum of all the covariances that contributes the largest part of the total variance; this same source also contributes the largest part of the mean black-white difference. It is also the sum of the covariances among a number of diverse tests which constitutes most of Spearman's general factor, or g. However, when small numbers of rather homogeneous subtests (e.g., the three sequential processing tests) are grouped together and scored as a single unit, such units reflect relatively little of the covariance that potentially exists in the test as a whole. The isolation of small groups of tests scored as separate composites in order to obtain a number of seemingly distinct diagnostic scales, therefore, re-

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

stricts the amount of *g* that is measured by any one of the part scores. This, in effect, underweights *g*, relative to other factors. If *g* is an important factor in the black-white difference (as I will later show that it is), any restrictions on the full emergence of *g* in the derived scores of a test will naturally tend to diminish the size of the difference. It is important to note that, as various tests are added together, the size of the black-white difference, in σ units, increases, up to a limit, and the rate of increase in the black-white difference is greater than the rate of increase in the total within-groups standard deviation of test scores. For example, the mean black-white difference on all the WISC-R subtests, each considered separately, is 0.70 σ, whereas on all the subtests combined, the black-white difference is 1.14 σ, an increase of 63%. Similarly, the mean of the black-white difference on the eight K-ABC mental processing subtests, when calculated separately, is 0.41 σ, whereas the mean black-white difference of the eight subtests combined is 0.65 σ, a 59% increase. Analysis of the WISC-R and many other tests shows that this consistent phenomenon is quite general and not restricted just to the Wechsler tests. It can be accounted for by four facts: (1) as the scores on various tests are successively added together, the total variance of the combined score increases with each successive addition; (2) much more of the increase in total variance is attributable to all the covariances among the tests than to the combined variances of the separate tests. Note that the total variance of combined tests *a* and *b* is $\sigma^2_{a+b} = \sigma^2_a + \sigma^2_b + 2\varrho_{ab}\sigma_a\sigma_b$, where $\varrho_{ab}\sigma_a\sigma_b$ is the covariance of tests *a* and *b*. Also, recall that for any given number, *n,* of separate test variances that go to make up part of the total variance, there are $n^2 - n$ covariance terms, and hence the sum of the covariances increases at a faster rate than the sum of the variances; (3) because the *g* factor arises largely from the covariances among diverse tests, the factor composition steadily changes, with the *g* factor accounting for a larger and larger proportion of the total variance at each successive addition of a test to the composite, and (4) because the black-white difference is mainly a difference in *g,* this difference becomes more markedly manifest as the composite score becomes increasingly *g* loaded with successive additions of scores on different tests. This last proposition is a corollary of Spearman's hypothesis concerning the nature of the black-white difference, namely, that it is mainly a difference in *g,* and, hence, that the varying magnitude of the black-white difference on various cognitive tests should be directly related to the *g* loadings of the tests (Spearman, 1927, p. 379). But before examining this hypothesis in greater detail relative to the K-ABC, a few words about Spearman's *g* are in order.

Spearman's g *and the theorem of the indifference of the indicator*

Despite their shortcomings, Binet's test and its modern descendants are surprisingly useful and valid measures of intelligence, not so much through knowing design on the part of their makers as, inadvertently, through the inevitable manifestation of a phenomenon described by Spearman as "the indifference of the indicator." Some 80 years ago, Spearman (1904) discovered that all cognitive tests, however diverse, provided they possess at least some minimum degree of complexity, are positively intercorrelated to varying degrees in the general population. This finding is consistent with Galton's (1883) and Spearman's (1904, 1927) theory that

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

there is a general ability which enters to some degree into the performance of every kind of mental task. This general ability factor is usually the largest source of individual difference variance among the significant factors of any diverse collection of cognitive tests; more than any other source of variance, this factor is reasonably identified as general mental ability, our best working definition of "intelligence."

Spearman's invention of factor analysis made it possible to quantify the degree to which any particular test measures the general factor that is common to the entire collection of various tests that are entered into a factor analysis. Provided that the collection of tests is fairly varied so that all the tests are not just different forms of the same type of test, the general factor, or *g*, in any particular collection is much the same as the *g* factor in any other collection. The different *g* factors found in various collections of diverse tests are more highly correlated with one another, moreover, than are the tests themselves; and the larger the number of tests in a given collection, and the more varied, the greater is the similarity of its *g* factor to the *g* derived from other collections of tests. The *g* factor extracted from even quite small collections, bearing hardly any superficial resemblance to one another, are remarkably similar. The *g* factors extracted separately from the six verbal subtests and from the six performance subtests of the Wechsler Adult Intelligence Scale, for example, are correlated about +0.8.

Because all cognitive tests measure *g* to some extent, and because any fair-sized collection of various tests measures *g* to a large extent, it is evident that the measurement of *g* is not tied to any particular test or to any particular collection of tests. One and the same *g* can be measured to slightly varying approximations by a virtually unlimited number of different test batteries. This is essentially the basis for Spearman's theorem of "the indifference of the indicator." Spearman (1927) himself is well worth quoting on this point:

> A corollary—more practical than theoretical—to be derived from the universality of *g* is what may be called the theorem of the indifference of the indicator. This means that, for the purpose of indicating the amount of *g* possessed by a person, any test will do just as well as any other, provided only that its correlation with *g* is equally high. With this proviso, the most ridiculous "stunts" will measure the selfsame *g* as the highest exploits of logic or flights of imagination. . . . And here, it should be noticed, we come at last upon the secret why all the current tests of "general intelligence" show high correlation with one another, as also with *g* itself. The reason lies, not in the theories inspiring these tests (which theories have been most confused), nor in any uniformity of constructions (for this has often been wildly heterogeneous), but wholly and solely in the above shown "indifference of the indicator." Indeed, were it worth while, tests could be constructed which had the most grotesque appearance, and yet after all would correlate quite well with all the others. (pp. 197-198)

In a sense, the K-ABC is both saved and defeated by *g*. Most of whatever practical validity the K-ABC will have, and some justification for its claim as a measure of intelligence, is due to its undoubtedly measuring *g* more than anything else. On the other hand, because the various scales of the K-ABC—Sequential Processing, Simultaneous Processing, Mental Processing Composite, Nonverbal, and Achievement—are all much more saturated with *g* than with any other source of usefully nonspecific factor variance, both the theoretical interpretation and practical utility of the scores derived from these various scales will be severely constrained. At the same time, the attempt by the K-ABC to measure other factors

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

besides *g* has most probably also diluted its *g* saturation to some extent, rendering its Mental Processing Composite a somewhat poorer measure of intelligence than either the Stanford-Binet IQ or the WISC-R Full Scale IQ. The K-ABC Mental Processing Composite is correlated about .61 with the Stanford-Binet IQ (*IM*, p. 116) and about .70 with the WISC-R Full Scale IQ. The corresponding correlations for the K-ABC Achievement Scale are .78 and .76 for normal school-age children. It thus would appear that the Achievement Scale is a better measure of intelligence than the Mental Processing Composite. But are these higher correlations of the Stanford-Binet and WISC-R with the Achievement Scale merely a result of the former tests' also being more scholastic-achievement oriented, or are they a result of all three tests' being better measures of *g* than is the Mental Processing Composite? The evidence quite clearly favors the latter interpretation.

Factor structure of the K-ABC

The *IM* (pp. 102–107) presents only varimax rotations of the two principal factors. By its very nature, varimax rotation completely obscures the *g* factor in the K-ABC battery, distributing and submerging the general factor (i.e., the first unrotated principal factor) among the rotated orthogonal (i.e., uncorrelated) factors. Alternatively, principal factor analysis without rotation of the factor axes permits us to apportion the total variance attributable to the general factor, or *g*, and to the only two other significant factors of the K-ABC battery, labeled sequential and simultaneous processing. I have done this analysis on the correlation matrices given on page 92 of the *IM*, with the following results.

When all the 11 subtests (7 mental processing and 4 achievement) for preschool children are factor analyzed, the first principal factor, *g*, accounts for 43% of the total variance; only 6% of the variance is divided between the sequential and simultaneous factors. The remaining 51% of the total variance is unique to each of the separate tests, and is theoretically uninterpretable in psychological terms. The four achievement tests show the larger *g* loadings, averaging .76, as compared with .57 for the seven mental processing tests. (Note that a *g* loading is defined as the correlation of a subtest with the *g* factor of the whole battery.) There is no independent achievement factor. Clearly, the achievement tests measure the same *g* factor as the mental processing tests, but the achievement tests do it better.

When all 13 subtests (8 mental processing and 5 achievement) for school-age children are factor analyzed together, *g* accounts for 44% of the total variance, leaving 5% divided between sequential and simultaneous processing, and 51% uniqueness, which is of questionable usefulness, diagnostically. Again, the five achievement tests show the highest *g* loadings, averaging .76, as compared with .59 for the eight mental processing tests.

But the K-ABC never combines all the mental processing tests and the achievement tests in a single score. So I have factor analyzed just the mental processing tests, with the results shown in Table 1. It is apparent that *g* accounts for some four to six times as much of the total variance of the Mental Processing Composite as do the sequential and simultaneous processing factors together. This means, of course, that the sequential and simultaneous processing tests measure *g* much more

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

TABLE 1
PERCENTAGE OF TOTAL VARIANCE IN THE K-ABC MENTAL PROCESSING COMPOSITE ATTRIBUTABLE TO COMMON FACTORS

	Percent Variance			
Factor	Preschool		School-Age	
g	35.6		38.4	
Sequential	2.9	5.7	3.4	8.5
Simultaneous	2.8		5.1	
Total	41.3		46.9	

than either measures any specific type of processing. The average of the *g* loadings of the sequential tests are .64 for preschool and .60 for school-age children; the corresponding average *g* loadings of the simultaneous tests are .56 and .62.

Actual scaled scores on the sequential and simultaneous tests of the K-ABC are not strongly associated with these specific different factors, moreover. In actual practice, regressed scores, with *g* regressed out, might be less misleading, as the large saturation of *g* in unregressed scaled scores prevents any clear interpretation of the various sequential and simultaneous tests. For example, Hand Movements and Word Order, which are both sequential tests, show lower correlations with each other than with Photo Series and Spatial Memory, which are simultaneous tests. Also, Matrix Analogies and Gestalt Closure, both simultaneous tests, show less correlation with each other than with Hand Movements and Word Order, both sequential tests. Thus, it is apparent that the sequential and simultaneous tests do not provide very clear measures of these two supposedly distinct dimensions of cognitive ability.

Much is made of the separation of the mental processing tests from the achievement tests, with the former treated as the only legitimate measure of intelligence. The Achievement Scale predicts scholastic success better than does the Mental Processing Composite, it is claimed, because the latter has eliminated factual and school-related tasks (*IM,* p. 13). A more defensible interpretation is simply that the Mental Processing Composite is less *g* loaded than the Achievement Scale. The *g* factor is the single best predictor of scholastic performance, which, of course, is itself highly *g* loaded. This is not because *g* depends upon the measurement of achievement per se, as can be clearly demonstrated by the fact that measures of individual differences in choice reaction time (which are *g* loaded but have absolutely no scholastic achievement content) also show substantial correlations with scholastic achievement (Carlson & Jensen, 1982). Scholastic achievement is strongly associated with *g* simply because such achievement generally allows more ample scope for the readily observable manifestation of individual differences in *g* than do most other activities in which children engage.

One of the many interesting phenomena discovered by Spearman in his investigations of *g* is the fact that highly *g*-loaded tests are generally more highly correlated with low-*g* tests than low-*g* tests are correlated with each other, even when the high- and low-*g* tests are made up of quite dissimilar contents and the low-*g* tests are of similar content. We can see this phenomenon in the WISC-R, for example.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

The two most highly *g*-loaded subtests are Vocabulary and Block Designs, which are correlated +0.43, despite the fact that, superficially, they are extremely dissimilar tests. The two least *g*-loaded subtests are Digit Span and Coding, which are correlated only .28, even though they appear somewhat similar, both involving short-term memory and concentrated attention. Yet Digit Span and Coding show slightly higher correlations with Vocabulary and Block Designs, ranging from .29 to .36.

This same phenomenon is found in the K-ABC. The mental processing subtest with the lowest *g* loading, Gestalt Closure, shows an average correlation with all the other mental processing tests of .35 for preschool and .30 for school-age children. But the corresponding average correlations of Gestalt Closure with the achievement tests are .44 and .34, despite the fact that there is nothing in Gestalt Closure which resembles scholastic achievement. The fact that Gestalt Closure correlates more highly with the achievement tests than with the mental processing tests can only mean that the achievement tests are better measures of *g*, not that they depend upon uniquely scholastic content. Thus, in accordance with Spearman's early observations, the K-ABC achievement tests in general are more highly correlated with the K-ABC mental processing tests than the latter are correlated among themselves, as is shown in Table 2.

The *IM* (p. 2) also emphasizes the purported virtue of minimizing the role of language and verbal skills in the mental processing subtests. It is indeed possible to measure *g* by means of nonverbal tests. Raven's Progressive Matrices test is an excellent example of such measurement. But greater ingenuity is required to devise nonverbal tests which measure *g* as well as do many verbal tests, such as vocabulary, similarities, verbal analogies, and the like. Again, it is not the verbal content per se which favors *g*, but the fact that the verbal medium lends itself so well to the processes of relation education, which processes, in turn, evince the strongest manifestation of *g*. If verbal content per se were so important, we should expect the nonverbal mental processing tests to correlate much more highly with the WISC-R Performance IQ (which has no verbal content) than with the WISC-R Verbal IQ. Yet the mean correlation of the eight mental processing tests with WISC-R Performance IQ is .36 and with Verbal IQ, .37, a trivial difference. The average correlation between the mental processing tests and the highly verbal Stanford-Binet IQ is .37. The average correlation among the eight mental processing tests themselves is only .37. These mental processing tests correlate as much with verbal

TABLE 2
CORRELATIONS AMONG AND BETWEEN MENTAL PROCESSING AND ACHIEVEMENT TESTS

	Mean Correlation			
Correlated Variables	Preschool		School-Age	
Among achievement tests	.60		.62	
Among mental processing tests	.34	**	.37	*
Between achievement and mental processing	.42		.42	

*Difference significant at $p < .05$.
**Difference significant at $p < .01$.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

tests as with nonverbal tests, because *g* is the major agent of correlation among all the tests. The Mental Processing Composite as a whole is correlated .74 with the Achievement Scale as a whole (*IM,* p. 90, Table 4.9). WISC-R Full Scale IQ and Stanford-Binet IQ are both correlated .78 with the K-ABC Achievement Scale in school-age samples—not all that different from the Mental Processing Composite. Yet consider the following statements from the K-ABC *IM:*

> The finding that WISC-R Full Scale IQ correlated more highly with K-ABC Achievement than with the K-ABC Mental Processing was anticipated because of the heavy weight given to verbal ability and factual knowledge in determining a child's global IQ on the WISC-R. This result gives credence to our contention that conventional IQs are to a large extent measures of children's school-related accomplishments. (p. 111)

> Acquiring knowledge . . . is so dependent on educational opportunities, environmental background, motivation, and often nonintellective variables, that it seems unreasonable to equate these achieved skills with intellectual functioning. Thus they have been kept entirely separate from the processing scales. (p. 33)

How similar is the *g* of the K-ABC to the *g* of the WISC-R and Stanford-Binet? The ideal way to determine this would be to obtain the correlations between *g* factor scores derived from each of the tests in a representative sample of children. As such ideal data are not available, we can use a more indirect method with the data provided in the *IM:* We can compare the *g* factor (unrotated first principal factor) loadings of the 13 K-ABC subtests for school-age children with the correlations between each of the subtests and the WISC-R Full Scale IQ and with the Stanford-Binet IQ. It is certain that the global IQ of each test would be very

TABLE 3
K-ABC SUBTEST FACTOR LOADINGS ON *g* AND CORRELATION WITH WISC-R IQ AND STANFORD-BINET IQ

		Correlation	
K-ABC Subtest	g Loading	WISC-R	Stanford-Binet
Sequential processing			
Hand Movements	.54	.31	.33
Number Recall	.55	.40	.47
Word Order	.64	.39	.46
Simultaneous processing			
Gestalt Closure	.47	.23	.10
Triangles'	.65	.59	.40
Matrix Analogies	.62	.57	.48
Spatial Memory	.56	.42	.32
Photo Series	.67	.48	.39
Achievement			
Faces & Places	.69	.55	.52
Arithmetic	.82	.65	.66
Riddles	.78	.66	.68
Reading/Decoding	.79	.45	.48
Reading/Understanding	.71	.62	.65
Average value	.67	.47	.47

Note. Correlations from K-ABC *Interpretive Manual,* p. 116.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

highly correlated with the g factor scores derived from each of these tests. But how well can we predict the correlations of the K-ABC subtests with the WISC-R and Stanford-Binet IQs, from a knowledge of the g factor loadings of the K-ABC subtests? Table 3 shows the g loadings and correlations. The correlations between the columns reflect the degree to which the g of the K-ABC resembles the g of the WISC-R and Stanford-Binet. These correlations are shown below. In parentheses are shown the congruence coefficients.[1]

g-Loadings × WISC-R	$r = +0.79$ (.99)
g-Loadings × Stanford Binet	$r = +0.83$ (.98)
WISC-R × Stanford-Binet	$r = +0.86$ (.99)

These correlations and congruence coefficients are so high as to suggest that all three tests measure very much the same g. The correlation of WISC-R × Stanford-Binet approaches the magnitude of correlation found for equivalent forms of the same test. Hence, we are justified in averaging the correlations in these two columns to obtain a better estimate of the K-ABC subtest correlations with the common g factor of the WISC-R and Stanford-Binet. The correlation between the profile of correlations of the K-ABC subtests with the mean WISC-R and Stanford-Binet and the K-ABC subtests' g loadings is $+0.84$. In other words, the correlations of the g loadings of the various K-ABC subtests predict their correlations with the general factor of the WISC-R and Stanford-Binet with a high degree of accuracy, as shown graphically in Figure 1 by the regression of the K-ABC subtests' mean correlations with WISC-R and Stanford-Binet IQs upon the subtests' g loadings. But there is an apparent anomaly: The regression constant, or intercept, is -0.26, rather than 0, indicating that the subtests' correlations with WISC-R and Stanford-Binet IQs are considerably lower, overall, than the subtests' g loadings. This discrepancy is reflected also in the average correlations seen in Table 3, where the average of the g loadings is 0.20 greater than the average correlations (i.e., 0.67 vs. 0.47). Why should this be the case if the g of the K-ABC is essentially the same as the g of the WISC-R and Stanford-Binet? Fortunately, this is not a crucial issue in the present context, as no definitive answer is possible from the available data. The g loadings shown here are based on the total school-age standardization sample ($N = 1500$), obviously a more heterogeneous group than the different and smaller (Ns = 182 and 121, respectively) samples used in comparing the WISC-R and Stanford-Binet with the K-ABC. The K-ABC variances are substantially smaller in these comparison samples than in the K-ABC

[1] The coefficient of congruence, r_c, is an index of factor similarity on a scale of 0 to ± 1. Unlike the Pearson r, which, being based on standardized variates, reflects only the degree of similarity between the profiles (of factor loadings) per se, the congruence coefficient also reflects differences in the absolute values of the factor loadings. A value of r_c above ± 0.90 is the usual criterion for concluding identity of factors, although some experts set a more stringent criterion at $+0.95$. The congruence coefficient is computed as follows:

$$r_c = \frac{\Sigma ab}{\sqrt{\Sigma a^2 \Sigma b^2}},$$

where a and b are the homologous factor loadings obtained on a given factor in groups A and B.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

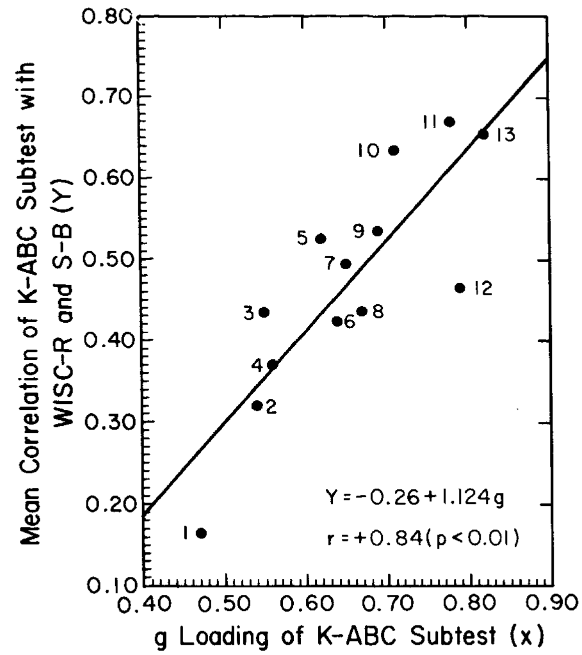

Figure 1. Regression of mean correlations of K-ABC subtest with WISC-R Full Scale IQ and Stanford-Binet IQ upon the *g* loading of the K-ABC subtests. K-ABC subtests: 1. Gestalt Closure, 2. Hand Movements, 3. Number Recall, 4. Spatial Memory, 5. Matrix Analogies, 6. Word Order, 7. Triangles, 8. Photo Series, 9. Faces and Places, 10. Reading/ Understanding, 11. Riddles, 12. Reading/Decoding, 13. Arithmetic.

standardization sample. Greater variance, of course, magnifies the overall size of the *g* loadings. Hence, if the WISC-R and Stanford-Binet IQs were correlated with the K-ABC subtests in the standardization sample, the correlations would most probably approximate the subtests' *g* loadings even more closely, especially in their overall magnitude.

It is also instructive to look at the *g* factor derived from the eight K-ABC mental processing tests in relationship to reading comprehension, which is generally the most highly *g*-loaded subject in the school curriculum. The *g* in this case is extracted from the matrix of correlations among only the eight mental processing tests, so as to preclude any influence by the *g* factor on the K-ABC achievement tests. The *IM* (p. 127, Table 4.25) shows the correlations of the mental processing tests with the Passage Comprehension score of the Woodcock Reading Mastery Tests. How closely do these correlations resemble the *g* loadings of the mental processing tests? The two sets of variables are shown in Table 4. The correlation between the two columns is .79; the congruence coefficient is .99. In other words, the *g* factor of just the mental processing tests is quite similar to the ability measured by a test of reading comprehension, an ability generally regarded as a good measure of Spearman's *g* among school-age children. The subtests' *g* loadings

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

TABLE 4
THE *g* LOADINGS OF THE K-ABC MENTAL PROCESSING TESTS AND THE TESTS' CORRELATIONS WITH READING COMPREHENSION

K-ABC Test	*g* Loading	Correlation with Passage Comprehension[a]
Sequential processing		
Hand Movements	.57	.38
Number Recall	.59	.40
Word Order	.65	.48
Simultaneous processing		
Gestalt Closure	.47	.28
Triangles	.69	.42
Matrix Analogies	.63	.48
Spatial Memory	.62	.34
Photo Series	.71	.47
Average correlation	.62	.41

[a] Woodcock Reading Mastery Tests. *N* = 550.

average .21 larger than the subtests' average correlation with the Passage Comprehension test, most likely because reading comprehension is not a pure measure of *g*, but also reflects other factors such as verbal ability. Even within the K-ABC standardization sample, the average of the correlations between the K-ABC reading/understanding achievement tests and each of the mental processing tests is only .38; this is .24 less than the average *g* loading of the mental processing tests.

The complexity dimension and g

Not every *g* is an equally good *g*, of course, because the *g* of any given battery of tests is determined by the nature and combination of the tests in the battery. Remarkably, however, some tests rather consistently show higher *g* loadings than others, almost regardless of the batteries among which they are factor analyzed, provided there is some degree of diversity among the tests in the battery (i.e., provided they are not all verbal tests or all spatial tests). Raven's Matrices, for example, show a high *g* loading in almost any battery in which they are analyzed. Even in a restricted college population, the Raven correlates more highly with the Wechsler Adult Intelligence Scale (WAIS) Full Scale IQ (r = .72) than does any of the WAIS subtests, even when there is no correction for contamination of subtest correlations with Full Scale IQ due to inclusion of the subtest in calculating the IQ. Also, the Raven has a larger loading (.80) on the *g* factor of the WAIS than does any of the WAIS subscales (Vernon, 1983). On the other hand, certain other tests, such as Digit Span (especially digits forward) consistently show fairly small *g* loadings in whatever battery they are included.

One single feature of tests that seems to be most consistently related to *g* is *cognitive complexity.* Tests involving more complex information processing, regardless of the informational content per se, show larger *g* loadings than less complex tasks. Hence, backward digit span, for example, is consistently more *g* loaded than forward digit span, and problem arithmetic is more *g* loaded than arithmetic computation. Rule-inferring tasks are more *g* loaded than rule-applying tasks.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

Guttman's (1954) radex model, a type of multidimensional scaling of any given collection of tests based upon the degree of inter-battery correlations, represents the tests' locations as points inside a circle. The complexity dimension of tests, which is directly related to a test's degree of correlation with all other tests in the collection, is scaled in terms of proximity to the center of the circle, the most complex tests being nearest the center. The contents of tests, designated as verbal, numerical, figural, etc., are represented in different sectors of the circle. In the radex model, the dimension of proximity to the center of the circle corresponds to *g*, and the various sectors correspond to what are termed *group factors* in common factor analysis. The outer circumference of the circle represents *specificity* (i.e., sources of variance not shared in common among any of the tests in the collection).

A recent study by Marshalek, Lohman, and Snow (1983) applied radex scaling to 34 highly varied cognitive tests given to 241 high school students, with the results shown in Figure 2. Cognitive complexity of the tests (and their *g* loading) is greatest in the innermost circle and steadily decreases as we move toward the periphery.

It would be informative to see the K-ABC tests scaled in this fashion, among a much larger collection of some 20 or 30 tests (including the WISC-R and Stanford-Binet), such as we see in Figure 2. It is my conjecture that none of the K-ABC mental processing tests would fall into the innermost circle of complex tests, and that most would fall into the outer area of simple tests. Probably only one or two of the mental processing tests, such as Triangles and Matrix Analogies, would fall into the intermediate area. The mental processing global score would also most likely be located in the intermediate area. A few of the tests included in Figure 2 tend to resemble some of the K-ABC mental processing tests, and it is worth noting their locations. Very near the periphery there are Digit Span Forward (Number Recall), Auditory Letter Span (Word Order), Street Gestalt and Harshman Gestalt (Gestalt Closure), Film Memory (Spatial Memory), and W-Picture Arrangement (Photo Series). Hand Movements and Face Recognition are also tests of short-term memory and would probably fall close to the memory span (MS) area in the lower left sector. In the area of intermediate complexity, there is Paper Form Board (Triangles). There is no test in the inner circle, the area of highest complexity, which closely resembles any of the K-ABC mental processing tests. The Matrix Analogies was specially devised to be simpler and less abstract than Raven's Matrices, and, as Spearman originally noted, *g* is related to abstraction as well as to complexity. Despite its simplification, however, Matrix Analogies shows a higher correlation with Stanford-Binet IQ and with Woodcock Reading Comprehension than any other subtest; it also shows the second highest correlation of any K-ABC subtest with WISC-R Full Scale IQ. The fact that Matrix Analogies ranks only fourth in *g* loading when factor analyzed among the eight mental processing tests suggests that most of these tests are not strong measures of *g*, but reflect in addition some other common factor which would not be separable from the present *g* factor unless more different tests, such as all the WISC-R subtests and the Raven, were included in the factor analysis.

In summary, then, whereas only four of the Wechsler subtests fall into the area of

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

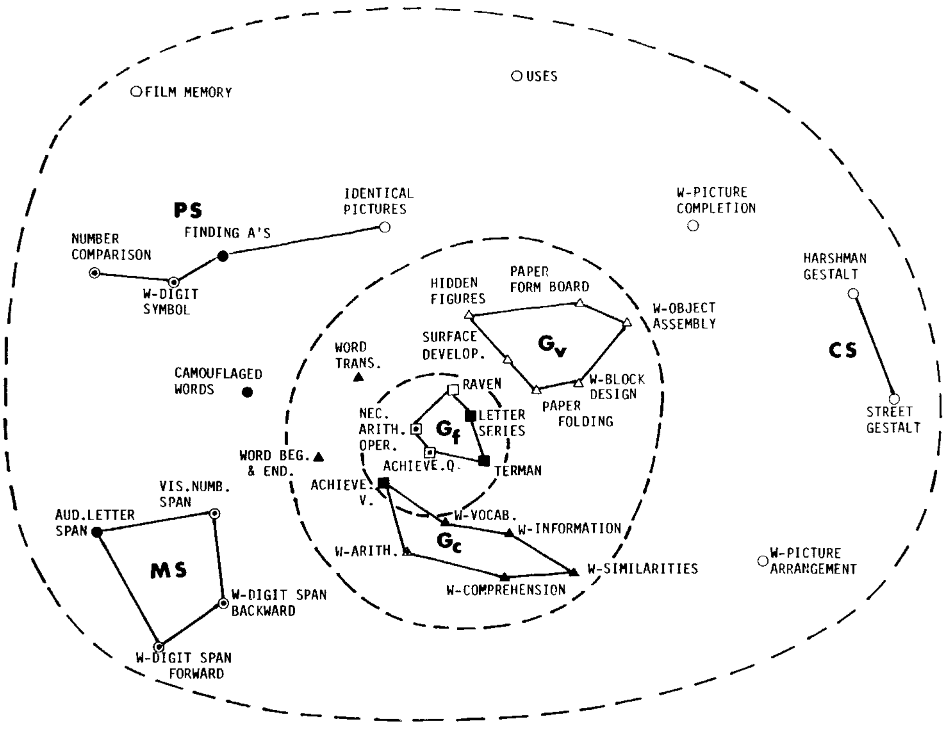

Figure 2. Multidimensional scaling of 34 tests, showing three levels of complexity ("concentric" circles) and three content areas. Complex, intermediate, and simple tests are indicated as black (verbal), dotted (numerical), and white (figural-spatial) squares, triangles, and circles, respectively. "W" stands for Wechsler Adult Intelligence Scale, G_f–fluid *g*, G_c–crystallized *g*, G_v–spatial-visualization reasoning, PS–perceptual speed, MS–memory span, CS–closure speed. (From "The complexity continuum in the radex and hierarchical models of intelligence" by B. Marshalek, D. F. Lohman, and R. E. Snow, 1983, *Intelligence, 7,* p. 107. Copyright 1983 by Ablex Publishing Corporation. Reprinted by permission.)

low complexity and seven are in the area of intermediate complexity, seven or eight of the K-ABC mental processing tests are of the type that would fall into the low complexity area and only two or three would fall into the area of intermediate complexity. This seems to me a reasonable conjecture based on my examination of the tests themselves and of their correlations with other tests. Of course, this conjecture remains to be empirically validated by someone in a position to perform the suggested analysis. My prediction, however, is that the K-ABC Mental Processing Composite will be found to be a relatively weak measure of *g* in terms of the complexity criterion, as compared with, say, Raven's Matrices, WISC-R Full Scale IQ, or Stanford-Binet IQ.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

NATURE OF THE BLACK-WHITE DIFFERENCE

Spearman's hypothesis

I have dwelled on the *g* factor of the K-ABC because the black-white difference is importantly related to *g*. Spearman (1927, p. 379) originally suggested that the black-white difference is mainly a difference in *g*. He had noticed that the size of the mean black-white difference, in standard score units, varied from one cognitive test to another, with the differences ranging from near zero to slightly more than one standard deviation in the same black and white samples. He had also noticed that the black-white differences on various tests were directly related to their *g* loadings. Elsewhere (Jensen, in press, a), I have examined Spearman's hypothesis, comparing data across 11 large-scale studies on intelligence. In each of these studies, some 6 to 13 diverse cognitive tests had been administered to black and white samples, from preschool children to adults. Hence, it was possible to test Spearman's hypothesis on the mean black-white differences on 121 tests and their corresponding *g* loadings. Results in this large-scale analysis showed a substantial and significant correlation between the mean differences and the *g* loadings. Moreover, Spearman's hypothesis is borne out in each of the 11 studies. Thus, the average black-white difference on diverse mental tests appears to be chiefly a difference in the general factor common to all of the tests rather than a difference in the more specific sources of test score variance associated with any particular informational content, scholastic knowledge, specific acquired skill, or type of test.

Intrinsic versus extrinsic differences

It is also instructive to examine the 13 subtests of the K-ABC in relation to Spearman's hypothesis. First, however, we should look at the summary data in the *IM* (pp. 149–150) concerning the black-white difference on the K-ABC and how they compare with the black-white difference on conventional tests. On the K-ABC mental processing scales, it is claimed, "differences between black and white children are approximately half the size of discrepancies typically reported for IQ tests" (Kaufman & Kaufman, 1983, p. 15). Reference is made to Table 4.36 in the *IM* (p. 152), comparing black-white differences on the K-ABC and WISC—R, reproduced here in Table 5. This tabled comparison is insufficiently informative, however, and may even be seriously misleading. It may come as a surprise that nowhere in the *IM* (nor anywhere else, as I am aware) is there an *intrinsic* comparison of the black-white difference on the K-ABC with the black-white difference on any conventional tests, such as the WISC-R or Stanford-Binet. All existing comparisons are *extrinsic*. What I mean by these terms is this: An *intrinsic* comparison of black-white differences on various tests reflects psychometric properties of the tests themselves, as regards item content, factor composition, or validity. A proper intrinsic comparison can be conducted only by administering the two (or more) comparison tests to the identically same black and white age-matched samples, and then, using the *raw scores,* dividing the mean group difference by the mean within-groups standard deviation. Thus, an intrinsic comparison does not in the least reflect differences between the standardization samples of the two tests. No intrinsic comparisons have been made of the K-ABC with other tests. An *extrinsic* com-

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

TABLE 5
BLACK-WHITE DIFFERENCES ON THE K-ABC AND WISC-R

	Blacks		Whites		
Intelligence Scale Score	*N*	Mean	*N*	Mean	Difference
K-ABC standard score	807		1569		
Sequential Processing		98.2		101.2	−3.0
Simultaneous Processing		93.8		102.3	−8.5
Mental Processing Composite		95.0		102.0	−7.0
WISC-R IQ	305		1870		
Verbal		87.8		102.0	−14.2
Performance		87.2		102.2	−15.0
Full Scale		86.4		102.3	−15.9

Note. From K-ABC *Interpretive Manual,* p. 152.

parison, on the other hand, is characterized by one or both of two features: (1) comparisons are based on different samples of uncertain comparability, and (2) the group difference on each test is expressed in standard deviation units derived from the standardization sample for each test. Unfortunately, the comparison of black-white differences on the K-ABC and the WISC-R shown in Table 5 (Table 4.36 of the *IM*) (Kaufman & Kaufman, 1983, p. 152) is based on the separate standardization samples of the K-ABC and the WISC-R; in addition, black-white differences are expressed in standard score units based on the distinct normative groups used in the standardization of each of these tests. This means that the observed black-white differences on the K-ABC and the WISC-R are confounded with variation that is totally irrelevant to the intrinsic psychometric properties of the tests themselves. Indeed, the comparisons in Table 5 show yet another source of extrinsic variation: The K-ABC sample consists of children ranging in age from 2½ through 12½ years, whereas the children in the WISC-R sample range from 6 through 16 years. This is a critical difference, because the black-white difference during the preschool years is smaller on conventional tests, and the difference, in σ units, does not become asymptotic until after children reach school age. This difference in age distribution between the two standardization samples, therefore, had the effect of spuriously minimizing the black-white differences on the K-ABC relative to those on the WISC-R. Also, notice in Table 5 that the black-white difference is almost three times greater on simultaneous processing than on sequential processing (−8.5 vs. −3.0). This disparity raises a crucial, but as yet unconsidered, question concerning the proper weights that should be assigned to simultaneous and sequential processing in arriving at the Mental Processing Composite. From the way the scores are combined, the resultant weights seem quite arbitrary, and I find nothing in the discussion of this issue in the *IM* which would contradict this impression.

In brief, I can find no data anywhere in the *IM,* or elsewhere, that permit a direct or intrinsic comparison of the K-ABC with any other tests concerning the size of the black-white difference. The difference is quite likely smaller on the K-ABC mental processing tests, but just how much smaller has not yet been properly determined.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

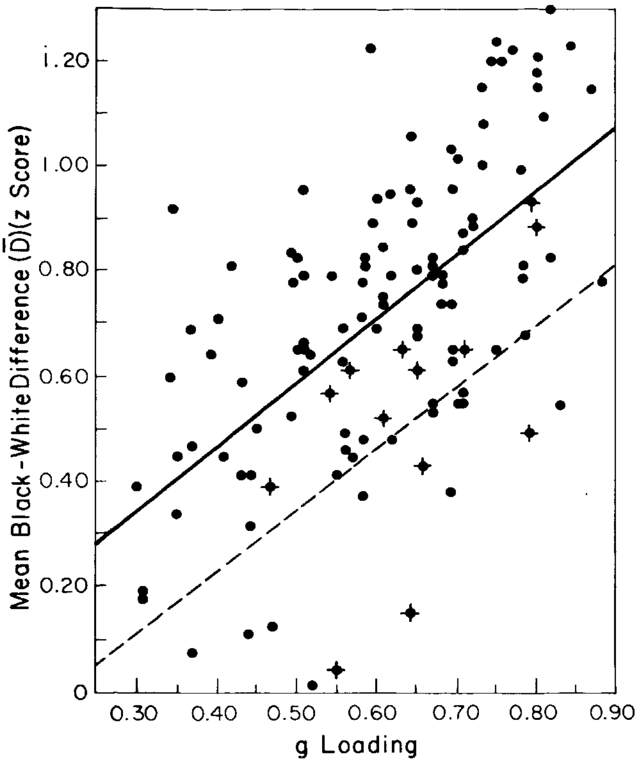

Figure 3. Correlation scatter diagram of *g* loadings and mean black-white differences (in standard score units) for 121 tests in 11 studies. The data points for the 13 tests of the K-ABC (based on the school-age standardization sample) are indicated by crosses. The regression line for all 121 tests is shown as a solid line; the regression for the 13 K-ABC tests is a dashed line.

The K-ABC and Spearman's hypothesis

Spearman's hypothesis predicts a positive correlation between the *g* loading of tests and the magnitude of the mean black-white difference. As noted earlier, this relationship has been substantiated for a large number of diverse tests (Jensen, in press, a). But does it hold for the subtests of the K-ABC? Figure 3 shows the scatter diagram for the relationship of the mean black-white difference, $\bar{D}$, in σ units, to the *g* loadings of 121 tests in 11 studies, including the 13 subtests of the K-ABC, for school-age children. The *g* loadings (first unrotated principal factor) were obtained from the factor analysis of the tests used within each of the 11 studies, based on different samples. Inevitably, there is a considerable amount of "noise" in such data, which attenuates the correlation. Once again, however, Spearman's hypothesis is borne out. As seen in Figure 3, the overall correlation (*r*) between *g* and $\bar{D}$ is +0.59 ($p < .01$). The 13 subtests of the K-ABC are not out of line with Spearman's hypothesis. The scatter diagram for the K-ABC battery is shown as crossed

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

dots in Figure 3; the correlation between g and $\bar{D}$ for just these 13 K-ABC subtests is $+0.58$ ($p < .05$)—nearly the same as the correlation of $+0.59$ for all 121 tests. By this criterion, then, the K-ABC tests cannot be regarded as at all atypical; they conform to Spearman's hypothesis at least as well as many other tests.

There is also one notable difference, however: The regression line for the K-ABC tests (indicated in Figure 3 by a dashed line) falls significantly below the regression line (solid line) for all 121 tests. The regression equations are as follows:

$$\text{All Tests:}\quad \bar{D} = 1.21\,g - .024$$
$$\text{K-ABC Tests:}\quad \bar{D} = 1.29\,g - .343$$

It is seen that the slopes are very similar, but the intercepts differ by .319. That is, the K-ABC tests show considerably smaller black-white differences than would be predicted from their g loadings. This phenomenon poses what may be the major puzzle of the K-ABC.

One possible explanation is that the g factor of the K-ABC is spuriously inflated, either by the greater heterogeneity of the K-ABC standardization sample or by some common source of variance in all the K-ABC subtests that is not the same as Spearman's g but which cannot be distinguished from Spearman's g except by factor analyzing the K-ABC tests among a diverse collection of other cognitive tests. How closely does the K-ABC battery conform to Spearman's hypothesis if we take as the subtests' g loadings their correlations with the WISC-R Full Scale IQ? What happens, in fact, is that the correlation between g and $\bar{D}$ increases slightly, from $+0.58$ to $+0.62$, and the regression equation changes markedly: $\bar{D} = 1.08\,g - .027$. The regression line under this condition is much closer to the regression line for all 121 tests, mainly because the intercepts now differ so slightly. But this finding is rendered somewhat ambiguous by the fact that the correlations between the K-ABC and WISC-R IQ are based, not on the K-ABC standardization sample, but on a composite sample from other studies that appears to have smaller variances of K-ABC scores than the standardization sample itself (Kaufman & Kaufman, 1983, Tables 4.19 and 4.20).

Spearman's hypothesis has been found to hold not only for tests factor analyzed within a given battery, but also for the mean g loadings of various tests batteries (Jensen, in press, a). It is instructive to look at the K-ABC Mental Processing Composite (MPC) and Achievement (Ach) scores in this manner, as well. Figure 4 shows the scatterplot for the mean g loading and mean black-white difference of the MPC and Ach batteries, along with the corresponding bivariate data points of test batteries in ten other studies. The correlation between mean g loadings and mean black-white differences, omitting the two K-ABC scales, is $+0.75$ ($p < .05$). (With the total K-ABC included, the correlation drops to $+0.62$, $p < .05$.) As can be seen in Figure 4, both of the K-ABC scales fall markedly below the regression line based upon the other ten batteries. This finding seems especially surprising in the case of the achievement tests. Although they have the largest average g loading of any battery in this collection, the mean black-white difference on the achievement tests is on a par with that for batteries having much lower mean g loadings. The mystery persists. Figure 4, however, contains another clue, related to the fact that

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

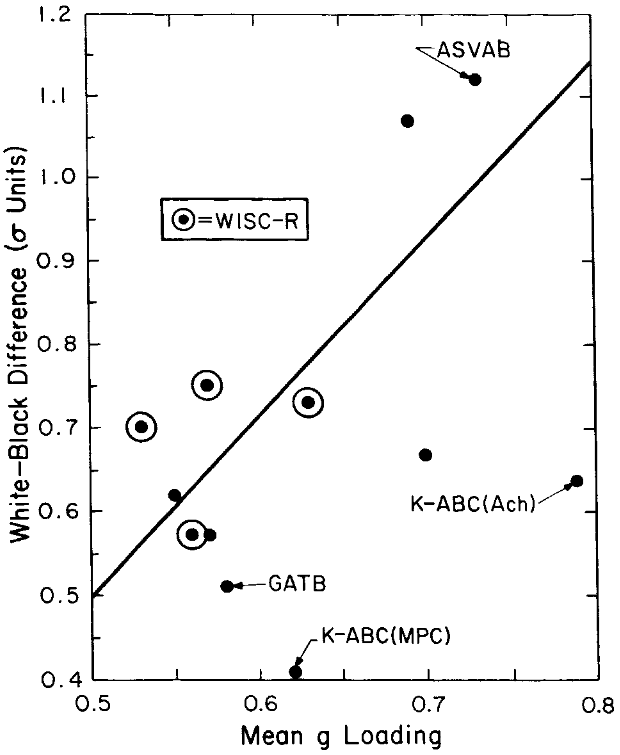

Figure 4. Mean black-white differences in 12 test batteries as a function of the mean *g* loading of the tests in each battery. The K-ABC scales were omitted in calculating the regression equation in order to see how well they fit the trend established by the other tests. For the ten data sets (other than the K-ABC), the regression of the mean black-white difference ($\overline{D}$) on mean *g* loadings is $\overline{D} = -0.57 + 2.14\, g$. The correlation between *g* and $\overline{D}$ *is* +0.75 ($p < .05$). (Some of the other test batteries are the General Aptitude Test Battery [GATB], the Armed Services Vocational Aptitude Battery [ASVAB], and the WISC-R.)

there are four studies involving the WISC-R. We see that there is considerable variation in the WISC-R data points for different samples in different studies. As already noted, the heterogeneity of samples affects the magnitudes of *g* loadings and group differences when expressed in σ units. Heterogeneity or range of talent in the sample increases *g* loadings and reduces intergroup differences. Could this be the explanation for the outlying positions of the K-ABC's MPC and Ach scales in Figure 4? We know that these data points would much more closely fit Spearman's hypothesis were it not for certain influences which are irrelevant to the hypothesis. The amount of *variation* among the *g* loadings and among the mean black-white differences in a given battery will affect the mean *g* as well as the mean black-white difference. When these two sources of variation, entirely irrelevant to Spearman's hypothesis, are partialled out of the correlation between the mean *g*

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

loadings and the mean black-white differences, the second-order partial correlation is +0.93 without the two K-ABC scales, and +0.86 with them (as compared with the zero-order correlations of +0.75 and +0.62, respectively). Even with these corrections, however, the MPC and Ach bivariate data points remain outliers, with the mean black-white differences on each score falling about 0.20 σ below the values predicted from their *g* loadings. There is no doubt about the fact that the K-ABC scales show smaller black-white differences for the size of their *g* loadings than do the WISC-R and other widely used test batteries. The crucial question is, *Why?*

HYPOTHESES ABOUT THE DIMINISHED BLACK-WHITE DIFFERENCE ON THE K-ABC

The apparently smaller black-white difference on the K-ABC invites several possible explanations, all of which at present can be regarded only as hypotheses. One might give better odds to some hypotheses than to others, in deciding priorities for future research, however.

Culture bias

No evidence has been presented that the K-ABC is less culturally biased in favor of white children than any other test. For example, in nationally representative samples of black and white children, Raven's Matrices test, one of the most purely *g* loaded of all tests, shows a mean black-white difference of 1.1 σ. Yet the several main objective criteria of predictive bias and of item bias have failed to detect any sign of black-white bias in the Raven. Why, then, does the K-ABC Mental Processing Composite show a mean black-white difference which is only about half as large as that found with the Raven? To support a hypothesis of differential culture bias, one would first have to show that the well-known tests which show larger black-white differences than the K-ABC are, in fact, culturally biased with respect to black-white comparisons, and, second, one would also have to show that the K-ABC is less culturally biased than the other tests. Neither of these has been done. Various objective criteria for the detection of bias have failed to reveal black-white cultural bias in most of the widely used standardized tests of cognitive ability or scholastic aptitude (Jensen, 1980a; Reynolds, 1982; Wigdor & Garner, 1982), and, to date, no objective data or arguments based thereon have been advanced which can explain any part of the observed black-white difference on standard tests in terms of cultural bias. Therefore it seems a most unpromising hypothesis that any smaller black-white difference manifested on the K-ABC should be attributed to its being even *less* culturally biased than conventional tests.

Biased selection of subtests in the K-ABC

A more likely partial cause of the diminished black-white difference is that several of the tests composing the K-ABC mental processing battery were specially selected, in part, because of their past record for showing smaller black-white differences than other tests. This effect could be due to the presence of either specific factors or some unidentified common factor on which black-white differences are small or even reversed. We know, for example, that black children score higher than white children on the factor of short-term memory in the WISC-R (Jensen & Reynolds, 1982). The Digit Span subtest is the most highly loaded on this memory factor, and as Digits Forward is less *g* loaded than Digits Backward, Digits Forward is the purest measure of the short-term memory factor in the WISC-R. The black-white difference is only about half as large on Forward as on Backward Digit Span (Jensen & Figueroa, 1975).

The K-ABC *IM* (Kaufman & Kaufman, 1983, p. 42) states that forward digit span (labeled Number Recall in the K-ABC) was selected for the mental processing battery specifically because, in previous studies, it had been found to yield no significant differences between blacks and whites when socioeconomic status (SES) is controlled. Backward digit span, in contrast, shows a substantial black-white difference even when SES is controlled. The K-ABC, however, does not include backward digit span, which involves more complex cognitive processes and is hence more highly *g* loaded than forward digit span. The Word Order test of the K-ABC is also essentially a forward memory span task; it is correlated +0.61 with Number Recall in the school-age standardization sample.

Similarly, certain other K-ABC tests were expressly selected because previous studies had shown especially small black-white differences: Gestalt Closure ". . . has produced equal means for blacks, Hopi Indians, and whites" (Kaufman & Kaufman, 1983, p. 41); and Face Recognition ". . . no cultural differences between the performance of 10- and 11-year-old American children and deprived Guatemalan Indian children of the same age; . . . rural 10- and 11-year-olds from Kenya performed almost as well, even though there were no black faces in the test" (p. 38).

I can find no psychometric or theoretical justification for selecting items or tests on the basis of their degree of discrimination between different populations. The construct validity of intelligence tests includes no stipulations about minimizing (or maximizing) the magnitudes of any particular population differences. If a population difference is to be viewed as a truly *dependent* variable in the measurement process, prior knowledge of such difference must not be considered in the construction of the measuring instrument itself.

Greater heterogeneity of the K-ABC standardization sample

As mentioned previously, an increase in the heterogeneity of ability produces two statistically inevitable effects: (1) inflation of *g* loadings, and (2) shrinkage of group differences expressed in standard score, or σ, units. Because the K-ABC tests show smaller black-white differences than would be predicted from their *g* loadings, it is a reasonable hypothesis that there is greater heterogeneity of the K-ABC standardization sample, as compared with the samples used for the standardizations of the many other tests which show larger black-white differences.

Several bits of evidence in the *IM* suggest that this heterogeneity hypothesis is especially worth considering. Unlike most other test standardizations, special consideration was given to including representative proportions of exceptional children in the K-ABC standardization sample (Kaufman & Kaufman, 1983, p.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

11). The inclusion of mentally retarded, learning disabled, speech and language impaired, severely emotionally disturbed, and other handicapped children, as well as gifted children, could very well make for greater heterogeneity and consequently larger raw score variance than typically exists in other standardization samples. Added to these sources of heterogeneity is another: the greater representation of minorities—blacks, Hispanics, Native Americans, Asians—in the K-ABC sample. I am not at all criticizing the K-ABC sampling procedure on these grounds, but only suggesting a possible explanation for its discrepancy with other tests.

The *IM* cites various studies in which various samples have been tested on both the K-ABC and on either the WISC-R, Stanford-Binet, or Peabody Individual Achievement Test (Kaufman & Kaufman, 1983, Tables 4.19, 4.21, 4.22). On each of these tests, standard scores are scaled to a mean of 100 and a standard deviation of 15 (or 16 on the Stanford-Binet), based on each tests's own standardization group. Hence, if the K-ABC standardization group is more heterogeneous than the others, we should predict that if one and the same group is tested on the K-ABC and on one of the other tests which has been standardized on a less heterogeneous sample, the K-ABC should show the smaller standard deviation *(SD)* when the *SD*s of both tests are expressed in terms of their own standard scores. Among those instances in which the appropriate comparison could be made (given the evidence of the *IM*), only six (or 21%) of the studies show a *larger SD* on the KABC than on the comparison test, while 22 (or 79%) of the studies show a *smaller SD* on the K-ABC than the comparison test—a highly significant (χ^2 = 9.14, 1 *df*, $p < .01$) difference, favoring the heterogeneity hypothesis. As I have previously suggested, the best way to test the heterogeneity hypothesis with respect to its effect on the black-white difference would be to compare the black-white difference on the K-ABC and on a comparison test (e.g., WISC-R) in terms of σ units derived from the raw scores obtained by testing the very same black and white age-matched groups on both test instruments, with the order of test administration counter-balanced.

Minor sampling artifacts

Two features of the selection process for the K-ABC standardization sample could also contribute in some part to the diminution of the black-white difference; at least, it is highly improbable that these features would have the opposite effect of increasing this difference.

Although the percentage of minority children in the standardization sample closely matches their percentage in the latest census of the nation's population, the *IM* (Kaufman & Kaufman, 1983, p. 15) indicates that a disproportionate number of minority children were selected from urban areas. It is well known that urban populations, black or white, generally score several points higher than rural populations on conventional IQ tests. Hence, if rural children are underrepresented in the black standardization sample, the effect would be an underestimation, to some presently unknown degree, of the size of the mean black-white difference on all the scores derived from the K-ABC. This effect would be somewhat lessened, but not entirely eliminated, by controlling for socioeconomic status (SES) as indexed by

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

the parents' level of education. Persons from urban and rural backgrounds matched on educational levels still differ, on average, in IQ.

The *IM* (Kaufman & Kaufman, 1983, p. 63) indicates that the parents' permission had to be obtained for all children tested in the standardization sample. This condition could also have biased the sample to some unknown degree. Whenever there is self-selection or parental selection for participation in a testing program, the most common effects, in our experience, have been that (1) a larger proportion of children from lower SES homes fail to return parents' permission forms, for whatever reason, and (2) parental permission is somewhat less apt to be granted for children showing low ability or poor adjustment to school. Both of these biasing effects would predictably affect the black and white populations differentially and would tend spuriously to minimize the average black-white difference.

Scale artifacts

There is an unusually large pile-up of raw scores of zero on some of the K-ABC subtests, with as many as 10 to 20% of the children in a given age group failing to pass a single item in the subtest (Kaufman & Kaufman, 1983, pp. 158–159). Scores of zero are especially common at the younger age levels. Obviously, some of the subtests do not have sufficient "bottom," in the range of item difficulty, to allow reliable discrimination of individual differences throughout the full range of ability. If the true distribution of ability approximates the normal curve in both the black and white samples, and if the black mean is below the white, then inevitably there will be a larger proportion of black children with raw scores of zero than of white children, and the true mean level of ability of the black children who obtained raw scores of zero would be lower than the true mean ability level of the white children who scored zero. Such failure to discriminate true ability differences among black and white children who score zero could only have the effect of diminishing the overall mean black-white difference. By the same reasoning, if there is a ceiling effect on high scores (and this is not discussed in the *IM*), the net effect of such a ceiling would also be to diminish the black-white difference. Whereas the lack of sufficient "bottom" in item difficulty would tend to diminish the black-white difference at the younger age levels, the ceiling effect, because of insufficient top, would tend to diminish the black-white difference at the older age levels. Although the biasing effects of these scale artifacts would not be very large, there is no doubt that they would to some degree work against the full expression of the true black-white difference in whatever ability is measured by the K-ABC.

Thus, the smaller black-white difference found with the K-ABC, as compared with previous tests, may well be attributable to any or all of several potential sources, including greater heterogeneity of the standardization sample, artifacts of subject sampling, and even certain scale artifacts. It surely has *not* been demonstrated that the apparently diminished group difference is a result of more valid or less culturally biased measurement of children's intelligence by the K-ABC.

IMPLICATIONS FOR FUTURE TEST CONSTRUCTION

Recent theoretical developments and empirical research on the nature of intelligence should make it possible to produce a better test of general intelligence than

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

the Stanford-Binet, the Wechsler scales, the K-ABC, or other tests of that type. The question of what intelligence tests of the future will (or should) be like is so open-ended as to allow virtually unlimited expression of opinion and conjecture. A small, but fairly representative, sample of rumination on this topic can be found in a symposium on "intelligence testing in the year 2000" (Sternberg & Detterman, 1979). My purpose here is necessarily much more limited. I will mention, as briefly as possible, just those points about future test construction that have been especially brought home as a result of my study of the K-ABC and which have not been emphasized in previous speculations on this topic. I do not assume that we are now in full possession of a psychometric technology that needs only to be applied to the task of constructing a better test than now exists. The technical means for achieving a test that would incorporate the theoretical desiderata I envisage will still require considerable research and development. But the increasing pace and quality of work in this field augurs well for the future of intelligence testing. The foreseeable obstacles are social and political, rather than scientific.

A single-purpose test

A good test of general intelligence will probably be a single-purpose instrument, just as the clinical thermometer and sphygmomanometer are single-purpose instruments. The Wechsler tests have set the unfortunate precedent for attempting to divine more diverse kinds of psychological information from an "intelligence test" than is statistically or psychometrically warranted. The K-ABC, unfortunately, follows in the same footsteps. The practice of basing inferences on the single scores of various brief subtests or on certain profile deviations among these subtests usually amounts to a scarcely supportable kind of "clinical psychologizing," bordering on crystal-ball gazing. If it is deemed important to measure traits other than general intelligence, or *g*, then separate and special tests should be devised that can be justified, theoretically, factorially, and empirically, as valid measures of the specific traits in question.

Measurement of g

The *g* factor is inescapable in any kind of cognitive test which allows persons to pass or fail some objective standard of performance. (By "cognitive" test, I mean simply that individual differences in sensory and motor capacities per se contribute negligibly to the total variance in test scores.) Therefore we should try to devise as good a measure of *g* as possible. This is indeed a big order, technically as well as theoretically, despite the virtually unlimited range of instrumentality implicit in Spearman's theorem of "the indifference of the indicator." Recall that not every cognitive test or battery of various cognitive tests yields an equally good *g*. With a limited time for testing, therefore, it becomes important to select techniques that will provide the best measurement of *g* in the most efficient manner possible. This can be done by means of computerized testing, which permits "zeroing in" very quickly on the subject's ability level with respect to a specific type of task, so that the maximal proportion of the testing time is spent on just those "items" that are the most optimally discriminatory at any given level of ability.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

Minimize prior knowledge

Although items involving specific kinds of knowledge and practiced skills can often be excellent measures of *g* when combined in a test, the interpretation of individual or group differences in *g* so measured puts too great a premium on the uniformity of opportunity for the acquisition of such knowledge and skills. We want our test of *g* to "read through" most of the variation in people's scholastic knowledge and acquired skills. In this respect, the intent of the K-ABC Mental Processing Composite is properly aimed.

But I believe we can go much further in this direction. And we can do so even while making use of verbal materials, which are so expressly avoided in the K-ABC mental processing tests. I suggest that this could be accomplished by measuring the average response latency, or "reaction time," to very simple items, the knowledge content of which is fully within every subject's grasp, as could be shown by the fact that there would be no reliable individual differences on the "test" if it were administered without a time limit and the total score were based on the number of correct responses. Various tests of this nature have already been found to measure *g* without necessarily measuring the group factors commonly found in conventional psychometric tests, such as verbal, numerical, and spatial abilities factors (Jensen, 1982a, 1982b, in press, b; Jensen, Schafer, & Crinella, 1981; Vernon, 1983). Detailed descriptions of these methods and how they can be further developed is beyond the scope of this paper. The Semantic Verification Test (SVT) now being used experimentally in my own laboratory is one example of how exceedingly simple verbal materials, which have been highly over-learned by virtually all persons having more than a third-grade education, can be used to measure fluid *g* at least as well as such highly complex verbal tests of vocabulary and verbal reasoning as Terman's Concept Mastery Test. All that is needed is familiarity with the three letters A, B, C, knowledge of the words *before, between, first, last,* and *not,* and an understanding of the concepts of "true" and "false." Subjects demonstrate the prerequisite skills for this test by performing every item type in the whole test as an untimed paper-and-pencil test. Among college students, the modal number of errors on this test is 0, the variance is scarcely greater than 0, and the reliability of individual differences is 0. Yet the average response latency to 80 of these items measures individual differences with a split-half reliability above .90. The items are so very easy that response latencies average less than one second per item. The subject sees each item on a display screen and responds by pressing either one of two buttons labeled *True* or *False.* First, some permutation of the letters A, B, C is shown for 2 seconds. This is immediately followed by a statement about the order of the letters, to which the subject responds either "True" or "False," by pushing the appropriate button. The statements are always of the following form:

B *after* A
B *before* C
A *first*
C *last*
B *between* A and C

All of these statements are also presented in the *negative* form (e.g., B *not after* A). The various forms involve different degrees of complexity of cognitive processing,

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

so that a slope measure can be obtained for each individual, indicating the linear rate of increase in response latency as a function of task complexity. The mean response latency of individuals and the intraindividual variability of the latencies on this test among college students are correlated about -0.5 with scores on the Raven Advanced Progressive Matrices, a highly *g*-loaded nonverbal test of figural reasoning. If corrected for the severe restriction of range in our highly selected university sample, the correlation is close to -0.7. This is only one of many possible tests for measuring *g* in terms of speed of cognitive processing in a variety of basic cognitive tasks, each of which reflects *g*. It should be noted that this is not a "speeded" test, in the sense that the subject must work against a time limit. The presentation of each item in the test is completely self-paced by the subject with each item appearing only after the subject has pressed a "home" button, situated a short distance below and between the *True* and *False* buttons.

A considerable variety of brief tasks is needed to measure *g* in this way, for the same reason that a number of different items is required in conventional tests. Every task or item also measures task-specific variance as well as *g* or other common factors; because many different tasks are employed, however, these uncorrelated specificities are, in effect, minimized in the total variance.

An external criterion for g

As the *g* factor is not identical from one battery of tests to another, we must appeal to some criterion outside of factor analysis for deciding which battery yields the better *g*. I suggest *complexity* of cognitive processes as the major criterion. In studies based on factor analysis or multidimensional scaling, what appears to be a continuum of cognitive complexity required for the successful performance of various tasks is virtually identical to the *g* factor of the tasks (Marshalek et al., 1983). The *g* factor, of course, is not a subjective property of tests. And although we are capable of making reliable subjective judgments of task complexity, it is also possible to define complexity objectively, both in terms of the number of elements or "cognitive manipulanda" that can be built into a task and in terms of the average response latency for a task. Complexity can be objectified more readily on relatively simple tasks, that is, tasks that are so simple that the only reliable measure of individual differences is response latency. Although such simple tasks, observed singly, are not themselves very good measures of *g*, their differences in complexity reflect corresponding differences in *g*. Such tasks can be used, therefore, as reference measures in a factor analysis, along with scores on a number of other, less objectively analyzable tests. A collection of psychometric tests that yields a good *g* factor would be recognized by the degree to which the *g* loadings increase on the reference tests, when these are ordered from lesser to greater degrees of objectively defined complexity. The *g* of different test batteries can also be compared by means of contrasting high- and low-scoring groups in terms of the degree to which they are discriminated by the complexity-ordered reference tests. The test battery with the better *g* can be identified by the fact that the high- and low-scoring groups show a greater rate of divergence in their respective mean levels of performance as the complexity-ordered reference tests increase from lesser to greater complexity; the test battery with the *g* of lower quality, concomitantly, will evince a lower rate of divergence.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

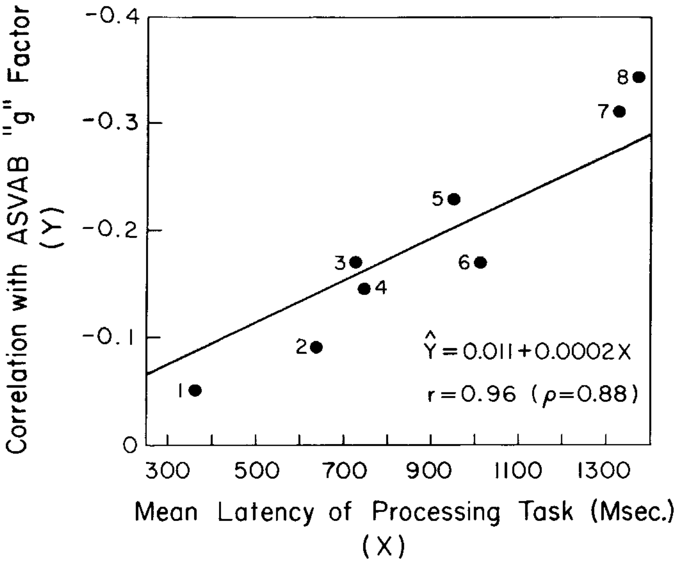

Figure 5. Correlation of eight cognitive processing tasks with the ASVAB general factor score as a function of task complexity as indicated by mean response latency on each task in a vocational college sample (*N* = 106). (From "Individual and group differences in intelligence and speed of information processing" by P. A. Vernon and A. R. Jensen, 1984, *Personality and Individual Differences.)*

Another method for assessing the *g* factor of a test battery is to correlate the mean response latency of the reference tasks, as an index of their complexity, with the degree to which the tasks are correlated with the *g* factor of the test battery in question. Vernon and Jensen (1984) have done this in a sample of 106 vocational college students, using the eight cognitive processing tasks of varying complexity as the reference tests, and their correlations with the *g* factor scores derived from the Armed Services Vocational Aptitude Battery (ASVAB). The Pearson correlation between (a) the mean response latencies of the eight cognitive processing tasks and (b) the correlations of the processing tasks with the *g* factor scores of the ASVAB is +0.96, as shown in Figure 5. We should expect a lower correlation for psychometric test batteries that yield a less satisfactory *g*. Also, using vocational college and university students as criterion groups that differ in average level of psychometric *g*, it was found that the correlation between the complexity of the eight processing tasks, as indexed by their mean response latencies, and the mean differences between the vocational and university students in response latency is +0.97, as shown in Figure 6.

A purely neuropsychological correlate of *g*—the average evoked potential—is another promising candidate for further research and development as an objective criterion of *g*. The most relevant research in this vein has been reviewed by Eysenck and Barrett (1985). Correlations in the range of .70 to .80 are reported between measurements derived from the average evoked potential (AEP) and

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

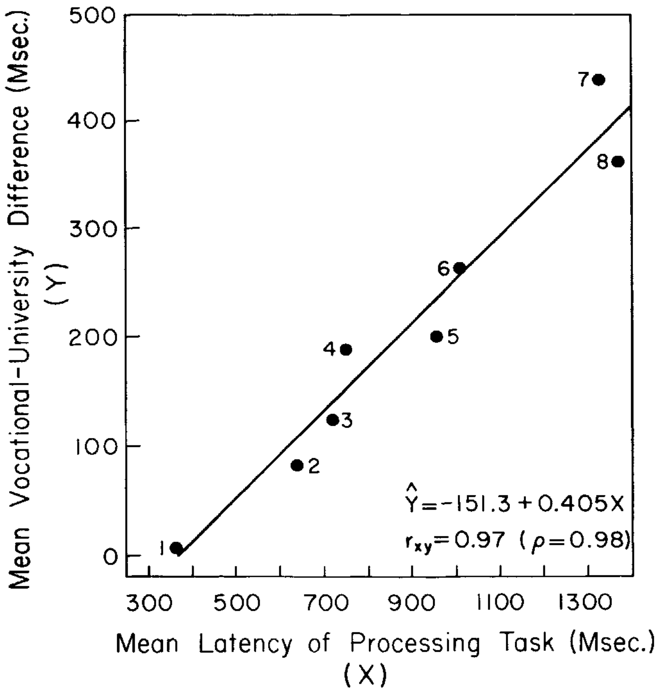

Figure 6. Mean difference (in msec) between vocational college students (*N* = 106) and university students (*N* = 100) on eight cognitive processing tasks as a function of task complexity as indicated by mean response latency on each of the tasks in the vocational college group. (From "Individual and group differences in intelligence and speed of information processing" by P. A. Vernon and A. R. Jensen, 1984, *Personality and Individual Differences.*)

Wechsler Adult Intelligence Scale (WAIS) IQ. Moreover, the correlation appears to be attributable mainly to the *g* factor of the WAIS. A most interesting and important discovery is that the *g* loadings of the 11 WAIS subtests are directly proportional to the various subtests' correlations with the AEP. The rank order correlation between (a) the subtest *g* loadings and (b) the correlations of subtests with the AEP is +0.93 ($p < .01$). This finding, of course, calls for replication, not only with the WAIS, but also with other test batteries. Replications of the same general finding would put the AEP in a powerful position as a reference measure for *g*. Its concurrence with the other criterion measures of *g* based on response latencies to cognitive tasks of varying complexity would also require investigation. We have found considerable overlap between reaction time and the AEP in their respective correlations with *g* (Jensen, Schafer, & Crinella, 1981).

Factor analysis of tests between and within families

The *g* factor in an intelligence test battery should be influenced as little as possible by differences in cultural and educational background. This is not to say that

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

cultural and educational background should not be correlated with cognitive test scores or with *g* factor scores, but rather that the definition and identification of what is a good *g* factor should not itself be influenced by population differences in ethnic, cultural, and educational background. One way to minimize these sources of variance in examining the factor structure of a test battery is to perform the factor analysis on data from highly homogeneous groups with respect to ethnic and social class background, thereby helping to ensure that differences in the relative magnitudes of the factor loadings will not reflect to any appreciable degree the particular mix of different backgrounds represented in the sample. The factor loadings of tests tend to be larger, overall, in an ethnically and socioeconomically heterogeneous sample than in a comparatively homogeneous sample. The important question, however, is whether the heterogeneity of the sample in any way changes the character of the factors. That is, does the heterogeneity affect the rank order of the magnitudes of the factor loadings on any factor? It should not do so for a good *g* factor.

An extreme example serves to illustrate my point. In India, literacy is highly related to the urban-rural distinction. If a battery consisting of several verbal and nonverbal paper-and-pencil tests were to be given to a mixed sample of urban and rural Indians, and if all the test intercorrelations were then factor analyzed, the resulting factors would largely reflect a demographic feature of the sample and would give a distorted picture of the structure of abilities—a picture determined mainly by the relative proportions of rural and urban subjects in the sample. Because the average correlation between the verbal and nonverbal tests would be quite low as compared with the correlations among the verbal tests and among the nonverbal tests, respectively, the *g* factor, if it emerged at all, would be quite small, and the magnitudes of the loadings on the various tests would give a poor indication of what the relative magnitudes of the *g* loadings would be in a more homogeneous population. In the highly heterogeneous sample, most of the common factor variance would be divided between a verbal factor and a nonverbal factor.

The most efficient means for investigating the effect of background heterogeneity on factor structure is to compare the results of factor analyses performed separately on *between*-family correlations and on *within*-family correlations. By definition, all of the variance that is attributable to ethnic, cultural, and socioeconomic factors exists *between* families. None of this variance exists *within* families (i.e., among full siblings reared together in the same family). By obtaining test data on pairs of full siblings in a large number of families, it is possible statistically to resolve the intercorrelation between each pair of tests in the battery into two component correlations: a *between*-families correlation and a *within*-families correlation. (The rationale and methodology for this procedure have been fully explicated elsewhere [Jensen, 1980b].) The sibling means on each test are the basis for calculating the *between*-families correlations; the signed sibling differences are the basis for the *within*-family correlations. The matrices of between- and within-family correlations are factor analyzed separately, and the resultant factors can be compared for degree of similarity by means of the congruence coefficient. (An example is given by Jensen, 1980b.) The *g* factor derived from the within-families correlation matrix, being free of extraneous variance due to demographic heterogeneity of the popula-

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

tion sample, is the preferable basis for assessing the *g* quality of the various tests. Other things being equal, moreover, the preferable test battery is the one which shows the least discrepancy between the factors derived from the two types of analysis (i.e., the between and within). The lack of such discrepancy would define one of the characteristics of a nonbiased test, with respect to the various subpopulations represented in the sample under analysis.

As noted previously, one of the problems in comparing the black-white difference on the K-ABC with the black-white difference on other tests is that the K-ABC standardization sample is suspected of greater heterogeneity than, say, the WISC-R sample. This means that a σ unit of difference on the K-ABC may represent a larger actual difference in ability than a σ unit of difference on the WISC-R; it also means that comparisons based on standard scores are likely to show a spuriously smaller black-white difference for the K-ABC than for other tests. This problem highlights the value of partitioning the total variance into between- and within-family components, using sibling data, as a means of comparing the composition of the total variance in the standardization samples of different tests. For example, if my conjecture concerning the greater heterogeneity of the K-ABC sample (as compared with the WISC-R sample) is correct, one would predict that the ratio of between-family to within-family variance would be greater for the K-ABC than for the WISC-R, owing to the fact that ethnic, cultural, and social class differences affect mainly the between-families variance. In my experience with this type of analysis (e.g., Jensen, 1974, 1977), the within-family variance is much more stable across different population samples than is the between-family variance. The reason for this greater stability is that within-family variance is affected little, if at all, by the particular mix of heterogeneous background factors in any given sample. This fact suggests that the within-family variance or its square root (i.e., the standard deviation) provides a more invariant unit for the expression of individual or group differences, and for the derivation of standard scores. Standard scores on different tests, standardized on somewhat different populations, would then be more directly comparable.

References

Carlson, J. S., & Jensen, C. M. (1982). Reaction time, movement time, and intelligence: A replication and extension. *Intelligence, 6,* 265–274.

Eysenck, H. J., & Barrett, P. (1985). Psychophysiology and the measurement of intelligence. In C. R. Reynolds & V. Willson (Eds.), *Methodological and statistical advances in the study of individual differences.* New York: Plenum.

Galton, F. (1883). *Inquiries into human faculty.* London: Macmillan.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarfeld (Ed.), *Mathematical thinking in the social sciences.* Glencoe, IL: Free Press.

Jensen, A. R. (1974). Cumulative deficit: A testable hypothesis. *Developmental Psychology, 10,* 996–1019.

Jensen, A. R. (1977). Cumulative deficit in IQ of blacks in the rural south. *Developmental Psychology, 13,* 184–191.

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

Jensen, A. R. (1980b). Uses of sibling data in educational and psychological research. *American Educational Research Journal, 17,* 153–170.

Jensen, A. R. (1982a). Reaction time and

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014

psychometric *g*. In H. J. Eysenck (Ed.), *A model for intelligence*. Heidelberg: Springer-Verlag.

Jensen, A. R. (1982b). The chronometry of intelligence. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1). Hillsdale, NJ: Erlbaum.

Jensen, A. R. (in press [a]). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences.*

Jensen, A. R. (in press [b]). Methodological advances in the chronometric analysis of mental ability. In C. R. Reynolds & V. Willson (Eds.), *Methodological and statistical advances in the study of individual differences.* New York: Plenum.

Jensen, A. R., & Figueroa, R. A. (1975). Forward and backward digit-span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology, 67,* 882-893.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences, 3,* 423-438.

Jensen, A. R., Shafer, E. W. P., & Crinella, F. M. (1981). Reaction time, evoked brain potentials and psychometric *g* in the severely retarded. *Intelligence, 5,* 179-197.

Kaufman, A. S., & Kaufman, N. L. (1983). *K-ABC: Interpretive Manual.* Circle Pines, MN: American Guidance Service.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7,* 107-127.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The Handbook of School Psychology,* New York: Wiley.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* London: Macmillan.

Sternberg, R. J., & Detterman, D. K. (Eds.). (1979). *Human intelligence: Perspectives on its theory and measurement.* Norwood, NJ: Ablex Publishing Corp.

Vernon, P. A. (1983). Speed of information processing and general intelligence. *Intelligence, 7,* 53-70.

Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences, 5,* 411-423.

Wigdor, A. K., & Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies. Part I: Report of the committee; Part II: Documentation section.* Washington, D.C.: National Academy Press.

Downloaded from sed.sagepub.com at DALHOUSIE UNIV on June 26, 2014