Race Differences and Type II Errors: A Comment on Borkowski and Krause*

ARTHUR R. JENSEN
University of California, Berkeley

Magnitude comparisons of black–white differences on a variety of cognitive tests give a somewhat different picture of the results of a small-sample study by Borkowski and Krause (1983), who based their conclusions mainly on significance tests. Some of the critical variables in the study consisted of differences scores with unacceptably low reliability, inclining the results toward the favored hypothesis, namely, that the locus of the difference between blacks and whites in psychometric intelligence lies in metaprocesses, or the executive system, and not in the elementary cognitive processes. The results actually show fairly comparable black–white differences in measures of both types of processes.

Borkowski and Krause (1983) surely have the right idea by attempting to analyze the nature of the well-known black–white difference in psychometric intelligence in terms of various cognitive processes, both elementary cognitive processes and metaprocesses. However, there are problems with their data and their treatment of the data which predispose their results to Type II error (i.e., accepting the null hypothesis when it is false) and lead them to conclusions which are scarcely warranted. A careful second reading of the Borkowski and Krause article will reveal that their data actually demonstrate considerably less than one's initial impression on casual reading, and perhaps even something rather different.

Reduced to the simplest terms, the gist of their study is that the black–white difference observed on IQ tests is ascribable to differences in the higher-level metaprocesses, or the executive system, and not to differences in elementary cognitive processes (which Borkowski and Krause refer to as "perceptual efficiency"). In their own words:

> Our hypothesis is that racial comparisons of black and white children from about the same SES level, but who differ in IQ, will show limited differences on the tests of perceptual efficiency. But because of differences in early environmental influences on the development of word knowledge, metacognition, and control processes, sizable racial difference are predicted

*Correspondence and requests for reprints should be sent to Arthur R. Jensen, School of Education, University of California, Berkeley, CA 94720.

> for tests reflecting these components of the executive system. (Borkowski & Krause, 1983, p. 383)

More specifically, with reference to Jensen's theoretical position regarding the nature of black–white cognitive differences, they claim to have expected to find differences in Word Span, Choice Reaction Time (CRT), Choice RT – Simple RT (CRT – SRT) difference scores, and measures derived from the Posner task (Physical Identity – Name Identity, or PI – NI). They conclude:

> such differences did not materialize, nor did a composite index of efficiency show racial differences. . . . Overall comparisons of black and white children on components of efficiency and executive processing lead us to conclude that sources of differences in general intelligence between white and black children lie more in the executive system than in perceptual efficiency. (Borkowski & Krause, 1983, p. 392)

Future studies may ultimately prove this conclusion correct. But the present study by Borkowski and Krause lends it little support. The small sample sizes, the emphasis on testing differences for statistical significance rather than looking at the actual magnitude of differences, inadequate control of age, and low reliability of some of the ''perceptual efficiency'' variables all tend to stack the cards in favor of the B & K hypothesis.

SAMPLES AND AGE

The subjects (*S*s) were 29 white and 20 black children in grades 2 and 3; mean ages in grades 2 and 3 were 8.0 and 9.25, respectively. These samples are small for this type of study, making for a serious lack of power for statistical tests. For example, a difference between the racial group means, expressed in standard deviation units (σ), would have to be $.60\sigma$ to be significant at the 5% level of confidence—a difference equivalent to 9 IQ points. And a Pearson r has to be .29 or above to be significant at $p < .05$. The mean ages of the black and white groups are not given, but it appears that the black group may be older than the white, as 60% of the black group are in Grade 3 as compared with 55% of the white group. The direction of such an age difference would diminish the mean difference between the groups to some degree. It is irrelevant here whether the age difference is or is not significant. Either the groups should have been perfectly age-matched, or age (in months) should have been regressed out of all the test scores. Borkowski and Krause tested the significance of differences in a nonorthogonal (because cell frequencies were unequal) two-way (Race × Grades) ANOVA. The results of such an analysis are not strictly generalizable to populations having different proportional cell frequencies. Merely including grades in the ANOVA does not provide adequate control for age as a source of variance. But that is a minor issue.

RELIABILITY OF DIFFERENCE SCORES

A major point of the study is based on an overall composite index of "perceptual efficiency." The black and white groups do not differ significantly on this index. But neither do second and third graders differ significantly. (We are never told the magnitudes of these differences.) This may seem surprising, because with an age difference of 1.25 years between grades, we should expect an efficiency index to reflect some developmental trend. Two groups that differed as much in mental age as the 2nd and 3rd graders differ in chronological age would differ about one standard deviation unit in IQ. We need to take a close look at the precision of this "perceptual efficiency index." It comprises four scores, three of them based on intrasubject differences:

1. Word Span (i.e., memory span for simple words).
2. Difference between 2-choice RT and simple RT (i.e., CRT − SRT).
3. Difference (in number correct) between physical identity (PI) and name identity (NI) in the Posner Letter-Matching Test (i.e., PI − NI).
4. Intraindividual variability in SRT, as measured by the range of an individual's RTs over 22 trials. The *S*s slowest RT and fastest RT trials were eliminated, and the RT variability score consists of the second slowest RT minus the second fastest RT.

Each of these scores was converted to a z score; the sum of the four z scores is the "perceptual efficiency" index.

The reliability of the Word Span test is not reported. Being a memory-span test highly analogous to the Wechsler forward digit span test, except that it uses familiar words instead of digits, Word Span is a Level I test in Jensen's theory, and as such it would be expected to show a quite small black–white difference, as Jensen and others have found in numerous studies of the Level I–Level II theory (Vernon, 1981). (Level I comprises primary or short-term memory and recall of sequential input without intervening mental manipulation or transformation of it. Level II comprises mental manipulation and transformation of input, reasoning, and problem solving.) Hence the mean difference of 0.16σ between the white and black groups in Word Span is consistent with Jensen's repeated finding that blacks and whites differ little in Level I ability and differ about 1σ or so in Level II ability. (Throughout this paper, mean differences are expressed in standard deviation, or σ, units, where σ is the square root of the N-weighted mean of the variances of the two groups.)

It is well known in psychometrics that scores derived from intrasubject differences—so-called *difference scores*—are highly liable to unacceptably low reliability. Any reliability coefficient which is not significantly greater than zero would certainly be unacceptable. For the Borkowski and Krause study, with 48 degrees of freedom, a reliability coefficient of less than .25 is nonsignificantly different from zero at the 5% level (one-tailed test). (Note that the reliability is the squared correlation between true scores and obtained scores, and with 48 *df* a correlation of .50 is required for significance at the 5% level by a one-tailed test. A one-tailed test is required, because the reliability coefficient can only take positive values.) The higher the correlation between the two primary scores, the lower is the reliability of the difference between them. In other words, the more highly two tests correlate positively, the greater is the overlap between *S*'s true scores, and the larger is the proportion of their obtained-score differences which is error. The reliability of the difference between two standard scores A and B is given by Stanley (1971, Formula 23, p. 385) as follows:

$$r_{(A-B)(A-B)} = \frac{r_{AA}\sigma_A^2 + r_{BB}\sigma_B^2 - 2r_{AB}\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2 - 2r_{AG}\sigma_A\sigma_B}$$

It can be seen that the larger the correlation between A and B, the lower is the reliability of the difference score. (The reliability coefficient, being a proportion of the total variance, theoretically can never be less than zero.)

The reliability of difference scores which enter into the "perceptual efficiency" index must be examined accordingly.

The reliabilities of SRT and CRT are .87 and .61, respectively, and the correlation between SRT and CRT is .60. The estimated reliability of the CRT − SRT difference score is .31. Thus 69% of the variance in difference scores is measurement error. Such low reliability severely attenuates correlations and mean differences expressed in σ or standard score units.

The reliabilities of PI and NI are .91 and .78, respectively, but PI and NI are correlated .90, so *the reliability of the PI − NI difference score is .04,* which is nonsignificant. With a reliability of practically zero for the PI − NI score, there is no need for Borkowski and Krause's elaborate speculation (p. 393) of why their failure to find a correlation between PI − NI and Raven scores does not square with the results of another study which reported a significant correlation (Keating & Bobbitt, 1978).

Intraindividual variability in SRT is measured by the *range.* Because the range is based on only two measurements, it is notoriously unreliable. In the Borkowski and Krause study, however, RT trials were administered in two blocks of twelve trials each and the range was measured in each block; the final range score is the average of the two range measures. (I have taken this into account in calculating the following reliability estimates.) Given the reliability of SRT of .87, based on 24 RT trials, we can estimate, using the Spearman-Brown formula, the average reliability of a single trial, which is .22. If we assume that the correlation between the top and bottom values of the range of RTs (eliminating the two most extreme values) is one half of the reliability of a single RT, i.e., .11, then the *reliability of the range* would be .22, also nonsignificant. Very low or nonsignificant reliabilities such as these should never be used to correct correlations or mean differences for attenuation. If such unreliable scores show large correlations with other variables, it only means the correlations are probably flukes of sampling or measurement error and would not be replicated in a repeat study. A considerably better measure of intraindividual variability in RT than the range is the standard deviation of the *S*'s RTs over *n* trials.

To summarize, the composite index of perceptual efficiency comprises a type of test, memory span, which has long been known to show a smaller black–white difference than just about any other kind of cognitive test, in addition to three other variables (all differences scores) with exceedingly low or nonsignificant reliabilities. All is not lost, however, because some of the primary measures of perceptual processing have sufficiently adequate reliabilities to show possibly interesting correlations with race and psychometric intelligence.

MAGNITUDE OF BLACK–WHITE DIFFERENCES

Significance testing without explicit attention to the actual magnitude of differences can conceal or obscure potentially interesting and important findings, as seems to have happened in the Borkowski and Krause article. It is instructive to look at the black–white differences on all of the variables in this study when they are expressed in terms of a common metric. In Table 1 the black–white differences have been expressed in terms of the point-biserial correlation, r_{pb} (with race quantified as black = 0, white = 1, and the signs of the correlations reflected, where necessary, so that a positive correlation always indicates "better" performance by the white group), and in terms of the mean group difference in σ units ($\sigma_{\text{diff.}}$). (The r_{pb}, and $\sigma_{\text{diff.}}$ are, of course, almost perfectly monotonically related, the correlation between them being .98 for these data.) The point-biserial correlations and σ differences on variables for which the reliability coefficients were reported by Borkowski and Krause (or could be calculated by Stanley's (1971) Formula 23 for the reliability of a difference score) are also shown with correction for attenuation, which gives a more accurate indication of the true magnitude of the black–white difference. Unfortunately, reliabilities were not reported for most of the variables. The variables based on difference scores are grouped separately in Table 1; their exceedingly low reliabilities make them erratic and untrustworthy. The much more reliable primary variables, however, show interesting features.

The first thing we notice is that the r_{pb} and $\sigma_{\text{diff.}}$ for the perceptual efficiency tests, representing elementary cognitive processes, are not very different from the r_{pb} and $\sigma_{\text{diff.}}$ for the seven metamemory tests, intended to represent the higher-level executive or control processes. The mean r_{pb} of the five perceptual efficiency tests is .19 as compared with a mean r_{pb} of .21 of the seven metamemory tests. The mean σ difference for the perceptual tests is .43 as compared with .44 for the metamemory tests. The multiple correlation of the PI and NI scores of the Posner letter-matching task with race is .25, equivalent to a 0.51σ

TABLE 1
Black–White Comparisons in Terms of Point-Biserial Correlation (r_{pb}) and Mean Difference Expressed in *SD* Units ($\sigma_{\text{diff.}}$), and the Reliability of Variables (r_{xx})

Variable	r_{pb}[a]	r_{pb}[b]	$\sigma_{\text{diff.}}$[a]	$\sigma_{\text{diff.}}$[b]	r_{xx}
"Perceptual Efficiency" Tests					
Word Span	.07		0.16		
Letter Matching					
Physical Identity (PI)	.25	.26	0.51	0.54	.91
Name Identity (NI)	.21	.23	0.64*	0.72*	.78
Reaction Time (RT)					
Simple RT (SRT)	.29	.31*	0.62*	0.67*	.87
Choice RT (CRT)	.11	.14	0.22	0.28	.61
Difference Scores					
PI–NI	.26		0.55		.04[d]
CRT–SRT	−.17		−0.35		.31[d]
Intraindividual Range of SRT	.25		0.53		.22[d]
Intraindividual Range of CRT	−.21		−0.44		.07[d]
"Executive System" Tests					
WISC Information & Vocabulary (IQ)	.36*	.37*[c]	0.78*	0.81**[c]	
Strategy Use	.84**		1.59**		
Metamemory Tests					
Memory Elaboration	.21		0.44		
Planful Behavior	.15		0.31		
Organized Memory Search	.29		0.62*		
Memory & Task Difficulty	.19		0.40		
Memory & Interest Categories	.32*		0.67*		
Memory & Strategies	.24		0.51		
Memory Monitoring	.08		0.15		
Raven Matrices	.61**		1.55**		

[a] Positive value indicates white group had "better" score than black group.

[b] Corrected for attenuation.

[c] From data in WISC Manual, the reliability of the Information + Vocabulary composite is estimated to be .92, which is the basis for the correction for attenuation.

[d] Estimated reliability; see text for explanation.

*$p < .05$ by 2-tailed t test.

**$p < .01$ by 2-tailed t test.

difference—a quite remarkable finding. The multiple correlation of SRT and CRT with race is .31, equivalent to a .64σ difference. If the zero-order correlations are corrected for attenuation before they are entered into the calculation, the multiple correlation of SRT and CRT with race is .39, equivalent to a black–white difference of .84σ, which is quite comparable to the black–white difference on the WISC Information and Vocabulary score. Simple and Choice RT are totally without any learned knowledge content or skill; they measure only elementary processes. Yet together they discriminate between blacks and whites to the same degree as Information and Vocabulary, the two most culturally and educationally loaded subtests of the Wechsler Intelligence Scale. This is also a remarkable finding. (Other RT studies of black–white differences are reviewed by Jensen, 1980, pp. 204–206.)

In fact, the only two variables classified by Borkowski and Krause as ''executive system'' that show appreciably larger black–white differences than the ''perceptual efficiency'' tests are the Wechsler Information and Vocabulary score and Strategy Use (a composite of three scores on two complex tests which assess degree of organization and category clustering in the study and free recall of letters and pictures). The Wechsler tests were originally devised by Wechsler to be good measures of IQ, so to include them in a battery of metaprocess tests to predict intelligence differences is, in a sense, merely circular. The well-known black–white difference on Wechsler IQ, on Raven Matrices, and on other IQ tests is what needs to be explained; it cannot itself be part of the explanation. The Strategy Use variable is therefore much more interesting and merits further research and analysis. It shows a black–white difference fully comparable to that on the Raven Matrices, a well-established measure of fluid general intelligence. The sizes of the black–white differences on the Strategy Use and Raven variables (1.59σ and 1.55σ, respectively), however, seem too large to be representative of black and white differences in general. They are most likely exaggerated by sampling error, especially in view of the fact that the black and white groups in the Borkowski and Krause study are said to be of about the same SES level. A comparison by the author of large representative samples of black (N = 1,143) and white (N = 1,493) elementary school children in California on the same form of the Raven Matrices as was used by Borkowski and Krause, showed a black–white difference of only 1.07σ (Jensen, 1973).

REFERENCES

Borkowski, J. G., & Krause, A. (1983). Racial differences in intelligence: The importance of the executive system. *Intelligence, 7,* 379–395.

Keating, D. P., & Bobbitt, B. L. (1978). Individual and developmental differences in cognitive processing components of mental ability. *Child Development, 49,* 155–167.

Jensen, A. R. (1973). *Educability and group differences.* New York: Harper & Row.

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

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Vernon, P. A. (1981). Jensen's theory of Level I–Level II abilities: A review. *Educational Psychologist, 16,* 45–64.