Journal of Vocational Behavior 29, 438–444 (1986)

Comments on the *g* Factor in Employment Testing

ROBERT L. LINN

University of Illinois, Urbana–Champaign

The papers in this special issue of the *Journal of Vocational Behavior* document several important facts about cognitive ability tests. A key and undisputed fact is that in a large and heterogeneous sample, positive correlations among ability tests are pervasive. Although the specificity and the pattern of correlations with a set of marker tests may vary a great deal from one test to another, all cognitive tests share some degree of common variance. Whether the *g* factor is defined in terms of the first principal component, the first principal factor, or the highest order factor in a hierarchical factor analysis, cognitive ability tests have positive loadings on the factor.

Of course, the *g* defined by one battery of tests or a particular factoring technique is not identical with the *g* defined by another battery or another factoring technique. Jensen (1986) acknowledges that the *g* factor is "not invariant across different samples of tests," but goes on to show there is a substantial degree of congruence provided each battery of tests is reasonably diverse in coverage.

Except for a carefully selected battery of tests, on the other hand, a single factor will not explain all of the common variance. This fact is acknowledged, at least implicitly, in the Hunter (1986), Jensen (1986), and Thorndike (1986) papers, but as these authors indicate, the lower order ability factors are correlated. Jensen is most explicit in his argument that good, simple structure requires oblique factors which lead logically to second- or third-order factors eventually culminating in *g*.

The ability to abstract a general factor from a diverse set of cognitive ability tests and the similarity of the definition of *g* from one test battery to another are interesting results, but, alone, do not demonstrate that *g* is of great practical or scientific significance. The primary contribution of the set of papers in this special issue is in the evidence that they provide on the questions of practical and scientific significance of *g* and

Requests for reprints should be sent to Robert L. Linn, Department of Educational Psychology, 210 Education Building, 1310 South Sixth Street, University of Illinois, Champaign, IL 61820.

0001-8791/86 $3.00

in the exploration of the implications of these results by Gottfredson (1986) and by Gottfredson and Crouse (1986).

In the space allotted for my comments, it would be impossible to do justice to all of the analyses and discussion in the papers. Hence, I will focus on just two issues: the practical utility of *g* and some of the policy implications of the results reported in these papers.

PRACTICAL UTILITY

Predictive validity has long been one of the principal bases for demonstrating the practical utility of tests. The evidence that tests of general cognitive ability have a useful degree of predictive validity for a wide range of jobs seems indisputable. The work of Hunter (1986) and his colleague, Frank Schmidt, provides an impressive array of evidence to support this conclusion. The conclusion of Hunter (1986) and of Jensen (1986) that specific cognitive ability tests add essentially nothing to the predictive power of *g* is more debatable, however.

Although Hunter acknowledges that psychomotor tests improve prediction for low-complexity jobs, he concludes that general cognitive ability is sufficient for complex jobs. Jensen summarized his closely related conclusion as follows: "virtually all test validity would be drastically reduced, usually to a level of practical uselessness, if the *g* factor were partialed out of the reported validity coefficients in all categories of test use." Thorndike's results provide only partial support for these conclusions.

Thorndike's (1986) analyses provide a convincing demonstration of the instability of multiple regression in small samples. With small samples, which are all that can often be obtained in practical work, a composite measure of general ability will provide better prediction than the unstable composite formed by using multiple regression weights. His results also show, however, that with large samples, multiple regression provides better prediction than a single general ability composite. These results are consistent with expectations based on the sampling variance of regression weights for correlated predictors.

Thorndike's (1986) results for the pilot training data are also worthy of note before accepting the idea that if you know *g* you need look no further for purposes of prediction. As Thorndike notes, the pilot stanine was substantially more effective than the *g* measure for the prescreened group. Again, a large sample is needed for the regression-weighted composite to outperform the measure of *g*, but given a large enough sample and prescreening on *g*, "a specialized battery can add significantly to predictive effectiveness" (Thorndike, 1986).

The correlations and path diagrams in Hunter's Fig. 5 provide support for the use of job knowledge tests as criterion measures as well as for the value of general cognitive ability as a predictor of job performance. I will not dispute either point, but I do think that more attention should

TABLE 1
Intercorrelations of Criterion Measures and Multiple Correlations of Tests with Criterion Measures in the ETS Civil Service Study[a]

Group	N	Intercorrelations			Multiple correlations		
		SR–JK	SR–WS	JK–WS	SR	JK	WS
			1. Medical technicians				
Black	166	.45	*b*	*b*	.16	.40	*b*
White	290	.25	*b*	*b*	.17	.52	*b*
			2. Cartiographic technicians				
Black	101	.28	.14	.47	.32	.68	.30
Hispanic	101	.42	.37	.50	.28	.59	.33
White	241	.36	.19	.55	.35	.66	.49
			3. Inventory management specialists				
Black	114	*c*	*c*	.24	.35	*c*	.51
Hispanic	74	*c*	*c*	.09	.28	*c*	.54
White	205	*c*	*c*	.24	.41	*c*	.39

[a] Based on Tables IV-2, IV-4, IV-9, VII-2, VII-5, and VII-8 from Campbell, Crooks, Mahoney, and Rock, 1973. The criterion measures are denoted SR for overall supervisory rating, JK for job knowledge test, and WS for work sample measure.

[b] The attempt to create a work sample for medical technicians was unsuccessful.

[c] Job knowledge tests were not administered to the inventory management specialists.

be given to the adjustments that were made in the correlations, because they influence estimates of the magnitude of impact, which is a key issue. To illustrate that the adjustments for attenuation and range restriction are nontrivial, I went back to the original ETS study (Campbell, Crooks, Mahoney, & Rock, 1973) that provided a substantial part of the civilian data summarized in Hunter's Table 5.

Uncorrected correlations for the three occupations included in the Campbell, Crooks, Mahoney, and Rock report are summarized in Table 1. As can be seen, the uncorrected correlations between job knowledge tests and the work sample measure are all well below Hunter's adjusted summary figure of .80. The multiple correlations of the tests with the criterion measures are also all well below Hunter's adjusted summary validity coefficients. Since the multiple correlations are based on the calibration samples, i.e., are not cross-validation results, it is clear, as Thorndike's (1986) results show, that even the validities in Table 1 are inflated.

The point of presenting the results in Table 1 is not to argue that adjustments should not be made. Indeed, I have argued elsewhere (Linn & Dunbar, in press) that such adjustments are needed for certain purposes. However, correlations that are changed dramatically by adjustments should

always be viewed with caution. They do not, in my opinion, justify the sweeping conclusions that Hunter draws from the summary results in his Table 5. At the very least, it should be emphasized that the adjusted correlations in that table are substantially larger than their unadjusted counterparts.

Crouse's (Gottfredson & Crouse, 1986) conclusion that achievement tests would work as well as the Scholastic Aptitude Test (SAT) for predicting college performance is consistent with the notion of *g* that is presented by Jensen. Like the other authors in this issue, Crouse accepts the fact that a measure of *g* provides reasonably good prediction of important criteria (in his case college performance). However, his main argument hinges on his conclusions that the SAT adds too little to the predictive power of high school grades to be of practical utility and that achievement tests, which would be expected to predict as well as the SAT, have greater educational value and therefore should be preferred.

There is, as Crouse has shown, a high correlation between predicted freshman grade point averages based on high school records alone and the predicted values from a combination of high school records and the SAT. Nonetheless, the prediction is consistently improved by the addition of the SAT. Crouse acknowledges that the SAT provides some improvement in prediction, but concludes that the gain is too small to be of practical utility, especially in comparison to the costs.

The typical increment in validity provided by the SAT is .07 or .08. Somewhat larger increments are typical at selective colleges. Ramist (1984), for example, reported an average increase of .10 for 51 colleges where the sum of the SAT Verbal plus Mathematical scores was between 1100 and 1199. The corresponding average increment for 22 colleges with a sum of 1200 or more was .11. Crouse prefers to consider "correct admissions decisions" in evaluating the increment due to the SAT, and by this approach finds that the increment in correct decisions is only about 1 to 3 per 100 applicants.

Although I would judge either set of figures to be of greater value than Crouse does, it is clear that competent judges can disagree on the magnitude of the increase that is needed to be considered worthwhile. The value of tests in the admissions process is not limited to increments in predictive validity, however. As I have argued elsewhere, "tests provide students with an alternative means of demonstrating academic ability" and they "provide a measure that is comparable across schools and across time" (Linn, 1982, p. 284). Crouse's argument would be considerably stronger if grades were comparable from one school to another. But, in fact, they are not. Without the common yardstick that tests provide, students who attend schools with stringent grading standards and stiff competition for grades would be placed at an unfair disadvantage, especially if admissions officers were unfamiliar with the school.

Crouse is probably correct in concluding that a set of subject matter achievement tests could contribute as much to prediction as the SAT, or, I would add, the general achievement tests provided by the American College Testing Program. It is unlikely, however, that such tests would yield larger increments in validity, and the difficulty of obtaining the needed consensus on the contents of the tests and the increased costs that would be involved in such as system would be nontrivial. Nonetheless, I think that this suggestion deserves more serious study and debate than it has received so far. The motivational value that Crouse foresees and the possibly positive effect that such a system could have on the rigor of high school courses are appealing.

POLICY ISSUES

Hawk (1986) appears enthusiastic about the gains that could be achieved by the proper use of *g* in personnel selection. Indeed, he goes so far as to suggest an astronomical figure of $80 billion a year as the potential gain in productivity. It is hard to take this figure seriously, in part because the people with the highest predicted performance on one job are the same people who have the highest predicted performance on other jobs, and in part for the reasons that are clearly explicated by Gottfredson (1986). As she indicates, the conflict between the competing goals of equality and efficiency are real. The differences between the ratios of blacks and whites with IQs in the recruitment ranges for the jobs in her Table 1 and the ratios of the percentage of the male populations actually employed in each occupation are quite striking.

Similar differences could be demonstrated for admissions rates to colleges and professional schools by comparing the ratios that would be obtained by a strict reliance on previous grades and admissions test scores with the actual ratios at many colleges. At least such an outcome is consistent with the results reported by Willingham and Breland (1982) who found that "minority status has, in most colleges, a large positive effect on selection" (p. 84).

The proper balance between efficiency and equality is, of course, a social policy issue rather than a scientific one. It is not an issue that can be resolved by evidence that greater reliance on tests of general ability could improve productivity, though such evidence is certainly relevant to the debate. Differences in group performance on general ability tests are also relevant to the social policy debate. The relative weights that should be given to these two sources of evidence depend on one's values. Whatever one's values, however, the social consequences that Gottfredson has brought to the fore demand consideration.

The question of heritability of general ability has long been a volatile issue, especially when juxtaposed with the magnitude of group differences. In light of this and of previous criticisms of Jensen's implicit linking of

the two (e.g., Cronbach, 1975), it is worth emphasizing one of Jensen's clearly stated conclusions about heritability and mean differences between populations. "A phenotypic difference per se affords no basis for inferring the degree to which either genetic or environmental factors contribute to the difference."

This conclusion is sound. But whatever the actual role of heredity, Americans place great stock in the role of education and hard work. Furthermore, social policies that are intended to improve the environmental conditions that may influence general ability are much more palatable than ones that rely on genetics. The question is whether there is any reason for optimism that environmental changes can be effective in improving the general ability of large segments of the population.

There are no easy answers to this question. Certainly, our efforts at compensatory education have fallen far short of the needs or the expectations of many reformers. There are two sources of evidence, however, that provide at least some basis for optimism. The first of these is the large increase in general ability of soldiers between the first and second World Wars and the second comes from several recent reports of trends in achievement for blacks and whites.

Humphreys (in press) recently summarized the results reported earlier by Tuddenham (1948) that showed that the median tests scores of soldiers in 1942 would have ranked above the 80th percentile in 1917. This presumably represents a large and important increase in *g* from one generation to the next. Humphreys (in press) argues that the large increase in the average amount of education during this time period provides the most plausible explanation of the dramatic increase in average performance on tests of general ability. There was a gain of roughly 3 years of schooling on the average during this period that saw a gain of about 1 standard deviation in test scores. As Humphreys indicated, the cost of such a large increase in schooling was high, but the apparent gains in the general ability of the population were well worth the cost. The papers in this issue provide ample evidence of the value of such a large gain in *g*.

Additional increases in average performance were achieved following World War II until the early 1960s, when the much publicized decline in SAT scores began (Congressional Budget Office, 1986). The decline, which was relatively steady for more than a decade, ended sometime in the late 1970s, and modest increases have been achieved on a variety of tests during the past few years (Congressional Budget Office, 1986). Although some have dismissed the decline on the grounds that the tests measure little of importance, the relationships of general ability to other important variables that are reviewed in the papers of this issue clearly indicate the fallacy of this sanguine attitude.

The gap between black and white performance on tests of general ability remains large, and for the reasons given by Gottfredson (1986)

the gap has serious social consequences. However, the educational achievement trend data that were recently summarized by the Congressional Budget Office (1986) suggest that some small progress has been made in closing the gap between black and white scholastic achievement. As summarized in that report, "the average scores of black students: Declined less than those of nonminority students during the later years of the general decline; Stopped declining, or began increasing again, earlier; and rose at a faster rate after the general upturn in achievement began" (p. 75). These results are of some encouragement, but the gap remains large. Given the significance of the societal consequences of *g* that Gottfredson has articulated, massive efforts that may be required to make more significant progress in closing the gap are still gravely needed.

REFERENCES

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in prediction of job performance. A six-year study.* Final Project Report (PR-73-37). Princeton, NJ: Educational Testing Service.

Congressional Budget Office (1986). *Trends in educational achievement.* Washington, DC: The Congress of the United States, Congressional Budget Office.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist,* **30,** 1–14.

Gottfredson, L. S. (1986). Societal consequences of the *g* factor. *Journal of Vocational Behavior,* **29,** 379–410.

Gottfredson, L. S., & Crouse, J. (1986). Validity versus utility of mental tests: Example of the SAT. *Journal of Vocational Behavior,* **29,** 363–378.

Hawk, J. (1986). Real world implications of *g*. *Journal of Vocational Behavior,* **29,** 411–414.

Humphreys, L. G. (in press). Intelligence: Three kinds of instability and their consequences for policy. In R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy.* Champaign, IL: Univ. of Illinois Press.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior,* **29,** 340–362.

Jensen, A. R. (1986). *g:* Artifact or reality? *Journal of Vocational Behavior,* **29,** 301–331.

Linn, R. L. (1982). Admissions testing on trial. *American Psychologist,* **37,** 279–291.

Linn, R. L., & Dunbar, S. B. (in press). Validity generalization and predictive bias. In R. A. Berk (Ed.), *Performance assessment: The state of the art.* Baltimore, MD: The Johns Hopkins Univ. Press.

Ramist, L. (1984). Predictive validities of the ATP test. In T. F. Donlon (Ed.), *The College Board technical handbook for the Scholastic Aptitude Test and achievement tests.* New York: College Entrance Examination Board.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior,* **29,** 332–339.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist,* **3,** 54–56.

Willingham, W. W., & Breland, H. M. (1982). *Personal qualities and admissions.* New York: College Entrance Examination Board.

Received: August 1, 1986.