Journal of Vocational Behavior 29, 332–339 (1986)

The Role of General Ability in Prediction

ROBERT L. THORNDIKE

Teachers College, Columbia University

Several data sets were analyzed to compare the prediction possible from a uniform general factor score with that produced by a separately tailored set of regression weights when those weights are applied to a new cross-validation sample. Double cross-validation designs were used. When regression weights were derived from large groups, they provided an increase of 10–15% in the prediction over that possible from a uniform general factor measure. However, with smaller samples, of the size typical of industrial personnel research, the uniform general factor score was clearly superior. © 1986 Academic Press, Inc.

In its beginnings, ability testing focused on providing a measure of general cognitive functioning—of something approximating Spearman's *g*. The Binet test in its various forms and adaptations provided a single score that was viewed as providing a general predictor of academic competence and of ability to function effectively in living and work. The early adaptations to paper-and-pencil, group-administered formats by Otis and others, appearing most dramatically in the Army Alpha and Army Beta of World War I, were likewise single-score tests oriented toward the appraisal of a general cognitive ability.

But there soon developed a movement toward more specialized ability tests, and tests of more limited cognitive functions. An early formulation of a doctrine of specialized ability tests appeared in Clark Hull's 1928 book entitled *Aptitude Testing*, that put forth the rationale and procedures for combining the results of specialized tests, using the statistics of multiple regression, to generate tailored batteries for each job. There followed shortly afterward Thurstone's development of multiple factor analytic procedures that seemed to dispense with any central factor of intellect, dividing ability up completely into a number of distinct and more limited abilities and implying that these more specialized abilities provided the key both to the theoretical understanding of human cognition and to the practical prediction of success in training and work.

Requests for reprints should be sent to Robert L. Thorndike, Teachers College, Columbia University, 525 W. 120th Street, Box 219, New York, NY 10027.

0001-8791/86 $3.00

Following up on the theoretical emphasis on specialized abilities, tests of such special abilities began to multiply, and aptitude test batteries began to replace general ability tests. Thus, the work of the U.S. Employment Service from 1945 on led to the production of the General Aptitude Test Battery (GATB). During World War II the Army Air Force Aviation Psychology Program produced the Aircrew Selection Battery, and after that war all the Armed Services introduced batteries for the selection and classification of entering enlisted personnel. The Psychological Corporation published the *Differential Aptitude Tests* in 1947, and these were followed by a number of other batteries developed by test publishers for civilian use in guidance and job placement. In all this enthusiasm for differential and specialized tests, the role of a central, general cognitive ability was at least deemphasized, if not completely forgotten.

But testing specialists have always realized, if they have not emphasized the point, that the ability factors of factor analysis are *not* completely independent, but rather are correlated. And even more so, tests that have been designed to measure distinct ability factors *do* show pervasive positive correlations. The notion of general ability sneaks back in through the back door in the pervasive correlations among the wide variety of special ability tests. So one is led to ask; To what extent is the effectiveness of these special ability tests attributable to their unique characteristics, and to what extent is it due to the core of general ability—some pervasive *g*—of which they all partake? What role does a broad general cognitive ability play in the prediction that we get from a test or a test battery? How does this compare with the effectiveness of a specially tailored set of tests and test weights designed for some specific educational or training program, or some specific job? This paper describes and summarizes several analyses that I have carried out in order to throw some light on this question.

It is important to realize that when we obtain validities for the tests in a battery of predictors and combine them in a weighted composite in such a way as to maximize the prediction of some criterion of job or training success, the prediction that we are able to obtain in the original experimental group capitalizes on the idiosyncrasies of the specific sample of cases on which the weights were determined. Our *real* interest is in how valid the procedure will be for some new, some independent, sample. We must always expect *some* slippage from the original sample to new and different sets of cases. An unbiased estimate of the validity of a particular choice of predictor tests and of weights for combining them always requires a cross-validation design in which a new and independent sample is obtained, and the validity of the weighted composite is determined for this new sample. It is with the validity in this cross-validation sample that the validity of a single universally applied measure of *g* should be compared. I have sought out several extensive data sets for which two

groups have been tested so that cross-validation is possible, and have compared the validity of tailored sets of weights for specific tests with a single score based on the common core running through all the tests of the battery—an approximation to a measure of *g*.

The first data set was drawn from the manual of the Differential Aptitude Tests (DAT) (Bennett, Seashore, & Wesman, 1966)—a battery of eight tests of verbal, quantitative, spatial, mechanical, and clerical abilities. The manual gives the validities of these tests for school grades in different school subjects for small groups of persons in a substantial number of schools, but also reports the median values for these sets of validity coefficients. Correlations among the tests are also reported, so it is possible to determine a set of regression weights and corresponding multiple correlations based on the reported intercorrelations and median school grades correlations.

However, the tests of the DAT are far from independent, but show correlations with one another that average about .50, so one can identify from the eight an underlying common factor, and determine a common factor score. Since validity coefficients were reported separately for girls and for boys, it was possible to determine regression coefficients on one sex and cross-validate those weights by applying them to the other sex. This was done separately for each of six subject areas, cross-validating girls' weights on boys and vice versa. The essence of the results is shown in the first section of Table 1.

For *this* data set, the shrinkage in validity upon cross-validation is quite small. The criterion variance accounted for by opposite sex weights is 95% as great as that with own sex weights—an outcome that reflects the fact that each set of weights is based on the pooling of a *large* number of cases. In the data sets that we will look at presently, the shrinkage becomes a considerably more serious matter.

The validity of the general factor score is substantial, but it *is* less than that for the cross-validation weights. One loses about 15% in one's predictive effectiveness by using a uniform general factor score, the same for all subjects and both sexes, rather than a set of weights tailored to each subject. It is of some interest that whereas score for the *first* general factor accounts, on the average, for 28% of the variance in school grades, a *second* factor, orthogonal to the first, accounts for less than 3%. It is the first factor that predominantly carries the load of prediction.

This first data set was limited to a school setting and to secondary education. The second data set relates to prediction of performance (again grades) in a wide variety of technical training schools in the U.S. Army. For this data set, I am indebted to Jack Hunter, who tracked it down in a vintage Army technical report (Schmidt & Hunter, 1978). The tests were 10 tests of the Army Classification Battery. Validity data were available for this set of tests for each of two classes in each of 35 Army

TABLE 1
Average Predictive Validity (R) of and Variance Accounted for (R^2) by Tailored Predictor Batteries vs the General Cognitive Ability Factor

Predictions based on	Average predictions			
Data Set 1: Predictor: Differential Aptitude Tests, criterion: grades in six high school courses				
	R (girls)	R (boys)	R (both sexes)	R^2 (both sexes)
Own sex weights	.621	.573	.597	.356
Opposite sex weights	.605	.560	.582	.339
First-factor score	.565	.499	.532	.283
Data Set 2: Predictor: Army Classification Battery, criterion: grades in Army technical training schools				
		R	R^2	
Own sample weights		.748	.560	
Cross-sample weights		.701	.491	
Uniform first-factor score		.668	.446	
Uniform second-factor score		.168	.028	
Data Set 3: Predictor: General Aptitude Test Battery (cognitive tests only), criterion: job performance				
		R	R^2	
Own sample weights		.458	.210	
Cross-sample weights		.318	.101	
Uniform first-factor score		.348	.121	
Data Set 4: Predictor: AAF Aircrew Classification Battery, criterion: pass–fail in pilot training				
	Experimental group		Prescreened group	
	R	R^2	R	R^2
Pilot stanine	.64	.410	.48	.230
Unweighted sum of 10 printed tests	.59	.348	.375	.141

Technical Training Schools. In a double cross-validation design, regression weights were determined for each class in each school and were applied to that class and to the other, or cross-validation, class. Sample size varied from school to school, but averaged about 280 students in each class. A distillation of these analyses is presented in the second section of Table 1.

For this data set, with samples averaging just under 300, the shrinkage in criterion variance accounted for as one goes from the original to the cross-validation sample is about 12%, as compared with the 5% for DAT samples. Here again, the single, uniform common factor score accounts for about 80% as much variance as do the regression weights applied to

their *own* sample. However, in comparison with the cross-validation sample, the single, common factor score provides *over 90%* of the predictive effectiveness. Here again, a second general factor orthogonal to the first accounts for less than 3% of the criterion variance, though the criterion correlations for this second factor are rather consistent from one class to the other in direction and size. Taken together, the two factors account for almost 97% as much criterion variance as can be accounted for by the cross-validated regression weights. Inspection of the loadings of the tests on this second factor, and of the validity coefficients for different schools, leads to the confident identification of this as a mechanical vs clerical dimension—a dimension that is meaningful but of relatively limited practical importance.

It must be recognized that the first general factor that we have extracted is not *g* in any sense of fundamental psychological theory. It is just what is common to *this* particular test battery. To the extent that the military establishment is heavily loaded with mechanical and technical types of training programs, the test battery is likely to have a slant in a technical or mechanical direction. However, I suspect that the correlation between the first-factor score from any one aptitude test battery and any other would be decidedly high, so that each would constitute a fairly good approximation to an underlying *g*. As more and more specific types of test task are pooled, the underlying general factor emerges more and more clearly, and the communality among batteries will become greater. Thus, different batteries tend to converge on more and more nearly the same *g*. This is illustrated in an Air Force research battery of 65 tests. Here, scores on the sum of subsets of eight tests from the battery, in which the average correlation for single tests was about .25, correlate to the extent of .67 on the average. Subsets of 16 show an average intercorrelation of .81, while Tests 1–32 correlate .87 with Tests 33–64. The increase is very nearly that called for by the Spearman–Brown prophecy formula relating reliability to test length, so it is much as if one were getting an increasingly reliable measure of the same underlying variable. One might argue from this that with almost any sufficiently large and varied set of measures, it is basically the same *g* that emerges.

Both of the data sets reported on so far have been concerned with success in school—academic or technical. We would like to get comparable evidence on the role of general ability in the prediction of on-the-job success. Unfortunately, validity data are not easy to come by in which criterion measures of on-the-job performance have been obtained for two or more separate samples of persons working in the same job. The only convenient compilation that I have been able to find appears in the Technical Manual for the USES *General Aptitude Test Battery* (U.S. Employment Service, 1970). Though some validity data are reported here for more than 400 different occupations, there were only 28 of these that

met the criterion of providing job (as distinct from training) criterion information for two or more samples each containing at least 50 cases. In most of these cases the samples were relatively small—composed of 50 to 100 individuals.

My first analysis was of the five tests that could be considered primarily cognitive in nature. A first-factor score was obtained, based on these five tests, and was compared with the self-sample validity of the score based on regression weights, and with the cross-sample validity based on those same weights as applied to the other sample. The results are summarized in the third section of Table 1. Here the shrinkage from own sample to cross-validation sample is a distressingly large 50%. Furthermore, the *uniform general factor* score shows up as an appreciably *better* predictor than the cross-validated regression weights. Apparently, with samples of this size one does better to completely forget about the elegancies of ad hoc weighting for a specific job, and to fall back on a measure of a general ability factor. Possibly, one might consider pooling all validity data for all the jobs, and calculate a kind of universal set of predictor weights that would be applied to all jobs. But I suspect that this would come quite close, both in its nature and in its effectiveness, to the general factor score.

Incidently, I carried out a similar analysis of the three motor-manipulative tests of the GATB. Here again, a simple sum of the scores on the three tests outperformed the regression weights when these were applied to the cross-validation sample.

One final analysis undertaken was to examine how consistent the ratios were of the validity of the regression weighted score to that of the general factor score across the different occupations for which data were available. In the case of the Army Technical Schools, out of the 35 schools, 22 showed ratios between 0.8 and 1.0 with 0.9 being about the typical ratio. They went as low as 0.6, and as high as 1.3, leading to the interesting question of which schools showed the highest ratio for g and which schools showed the lowest. The results were that the two for which the ratio of g to the regression weights was highest were Track Vehicle Chassis Repairman and Medical Technician, and the one for which the g ratio was lowest was Stenographer. In the case of the GATB data, the spread of ratios was greater because the Ns were small and everything was more erratic. The ratios ranged from 0.4 to 3.5. The two occupations for which the g factor was most effective were Camp Counselor and Electrician, and the two for which it was lowest were Manager of a Retail Food Store and Meat Cutter. Logical explanations for these patterns are not yet evident.

The results from these three data sets suggest that it is quite fruitless to develop special weighting systems for predicting job performance when one's validity data are based on samples of the size that is typical of

(and often all that is feasible in) industrial personnel selection research. Somewhere as the sample size gets up to 100–200, one appears to reach a break-even point where regression weights will do *as well as* a general factor score, and beyond that specialized weighting systems may begin to do better than a common factor score by a modest amount.

Note that I do *not* propose any common *single test* with a uniform type of item as a measure of the general ability. *No* single-test-item format can provide a really satisfactory measure of a general *g* factor. *Any* one item type carries a burden of measurement error, possibly of some group factor of intermediate width, and *certainly* of variance that is specific to the item type. The specific factor variance is likely to represent, in part at least, the degree to which the individual has developed strategies—"tricks of the trade"—that are effective for that particular test, and that have no relevance or validity for nontest situations. The good measure of *g* will be one that samples widely from a number of different tasks—8 or 10 or a dozen. In this respect, some of the early test makers (e.g., the authors of the Army Alpha or of the Kuhlmann–Anderson) showed good judgment, although numerous item types introduce compensating problems in the need for careful timing, time wastage on multiple sets of instructions, and possibly an undesirable emphasis on speed.

One reason that a measure of cognitive ability sometimes does not show up so favorably in relation to other more specialized tests, or in relation to noncognitive measures, is that prior test, educational, or life hurdles have already screened out those low in *g*, who would have been likely to fail because of limited cognitive ability. We do not often get a clear demonstration of this phenomenon, so it may be worth while to remind the reader of one such demonstration, stemming from the Aviation Psychology Program of World War II (Flanagan, 1948).

During that war, the standard procedure in the AAF was to screen all applicants for aircrew training with the AAF Qualifying Examination, with a cutoff designed to be comparable to the level achieved by the average college sophomore. Those passing the qualifying test were later given a 2-day battery of tests to determine their assignment to pilot, navigator, or bombardier training, or their consignment to the limbo of gunnery school. As the war rolled on, and the supply of aircrew in training became quite abundant, it was possible to sell to the "top brass" the idea of letting one "experimental group" of 1000 applicants go into training, waiving *all* the normal prerequisites. The men in this "experimental group" were not screened by the qualifying test, and were assigned to pilot training no matter how badly they scored on the relevant tests of the classification test battery.

With this unscreened group, the regular classification battery, using the standard set of regression weights, was quite an effective predictor

of training success (or, more often, of failure). Correlation of the Pilot Stanine, the regression-weighted composite of classification tests, with pass–fail in pilot training was .64. But in this group a simple unweighted sum of 10 paper-and-pencil tests, a number of which had been included in the test battery to predict navigator rather than pilot success, was .59. If we think of this sum as one more approximation to a measure of *g*, we have one more instance in which *g* provided, in a large sample, 85% of the prediction achieved by a regression-weighted battery with weights arrived at after several years of research on large groups of trainees.

By contrast, however, we may look at the relative validity of the *g* measure in groups that *had* been screened with the Qualifying Examination, an examination that correlated .78 with the *g* measure in the unscreened group. The values appear at the right in the section at the bottom of Table 1. Of course, the prescreening resulted in lower validity for *both* the Pilot Stanine and the *g* measure. But now *g* is only 61% as effective as the stanine. Prescreening had weeded out most of those who would have failed because of limited general cognitive ability, leaving other components as relatively more important. Thus, where weights are based on *large* groups, and where screening on *g* has already taken place, a specialized battery can add significantly to predictive effectiveness. But where these conditions are *not* met, *g* appears to be what makes the predictive wheels go around.

The general conclusion of this paper is that it is easy to fool personnel administrators and perhaps even oneself by doing fancy regression weightings of tests in a prediction situation. But unless the cross-validation study that is implied by my remarks is completed, conclusions regarding predictive efficacy will be useless. And with samples of the size that are usually encountered in personnel work, *g* may be the best predictor available.

REFERENCES

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1966). *Differential Aptitude Tests manual* (4th ed.). New York: Psychological Corporation.

Flanagan, J. C. (1948). The aviation psychology program in the Army Air Forces. *The experimental study of a thousand applicants sent into pilot training.* (Rep. No. 1, pp. 78 ff). Washington, DC: U.S. Govt. Printing Office.

Schmidt, F. L., & Hunter, J. E. (1978). Moderator research and the law of small numbers. *Personnel Psychology,* **31,** 215–232.

U.S. Department of Labor (1970). *Manual for the USTEF General Aptitude Battery Test,* Section III, Development. Washington, DC: U.S. Government Printing Office.

Received: July 1, 1986.