Test Length and Validity

Richard Bell and James Lumsden
University of Western Australia

The effect of test length on predictive validity is examined empirically by successively omitting the poorest items and by calculating the correlations between the reduced test scores and the criterion. It was found, for four tests, that the curve of validity against test length had a very gentle slope for the longer tests and that all tests could be reduced by more than 60% without appreciable decreases in validity.

It has long been recognized that increasing test length will, all other things being equal, increase predictive validity because of the increased reliability of the test scores. The problem was first considered by Spearman (1910) and has subsequently been discussed in standard texts such as Gulliksen (1950) and Lord and Novick (1968, Section 5.11), who present a formula relating test length to validity for homogeneous tests as:

$$r_{nv} = \frac{r_{ov}\sqrt{k}}{\sqrt{1 + (k-1)\, r_{oo}}} \qquad [1]$$

where

r_{nv} = validity coefficient for the test of new length;

r_{ov} = validity coefficient for the test of original length;

r_{oo} = reliability coefficient for the test of original length; and

k = ratio of new length to the original length.

It is unlikely that any test presently in use for predictive purposes would meet the homogeneity requirement, and the authors have been unable to find any empirical tests of the formula.

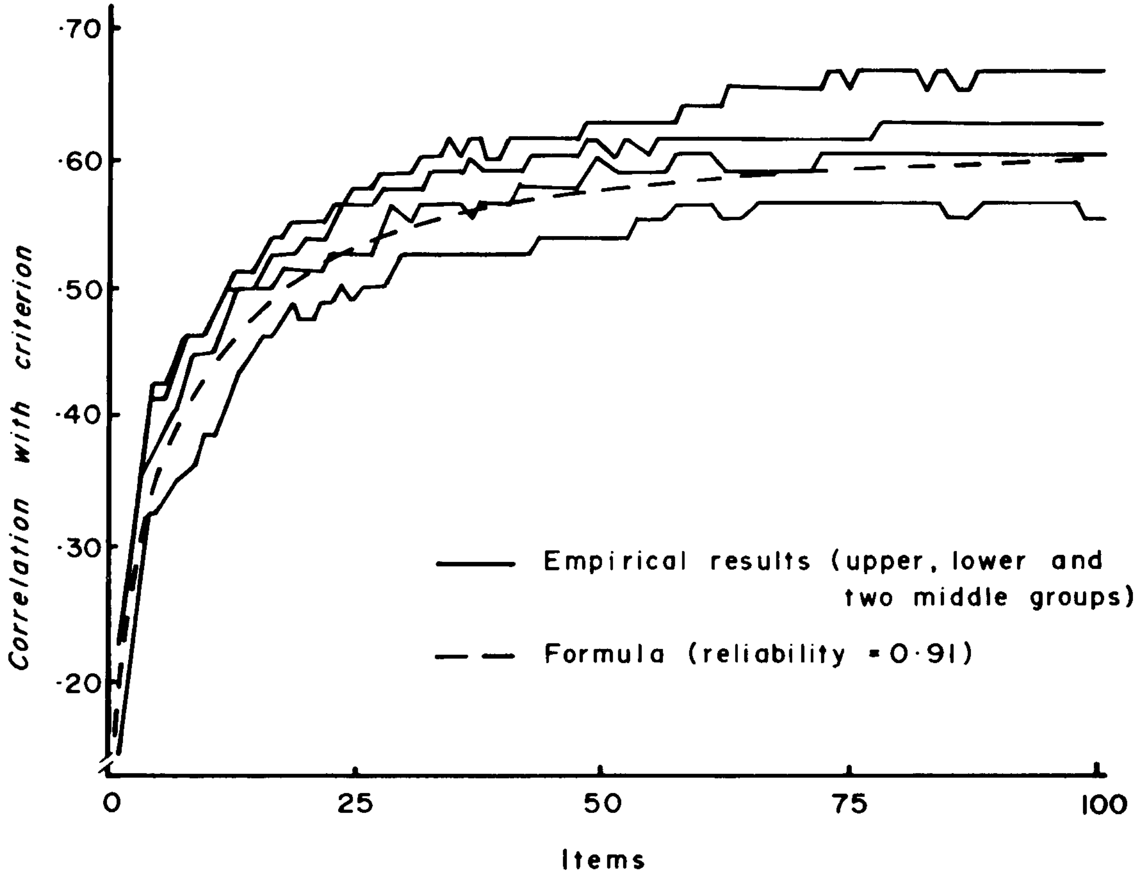
The problem remains of importance both theoretically (as in the attenuation problem) and practically (as in the problem of the optimally efficient test length). It is the purpose of this paper to describe a direct empirical approach that makes no assumptions about homogeneity.

The empirical approach has three steps:

1. A test is constructed and a validity coefficient (r_{T_nC}) is calculated for it.
2. The test is reduced in length by one item at a time and successive validity coefficients ($r_{T_{n-1}C}, r_{T_{n-2}C} \ldots$) are calculated. The item to be removed at each stage is determined by the considerations used in item selection for the original test, i.e., the "poorest" item is deleted.
3. The successive validity coefficients are plotted against test length.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 2 Spring 1980 pp. 165–170
© Copyright 1980 West Publishing Co.

Figure 1
Validity as a Function of Test Length for ASAT-B



These steps can be carried out conveniently and inexpensively from the data collected for the validation study for the original test.[1] If the curve of validity against test length remains steep throughout, then consideration may properly be given to increasing test length; there will be good reason to believe that the traits measured by the test and the criterion are significantly more strongly related than is suggested by r_{T_nC}. If the curve is flat in the longer test length region, then, clearly, lengthening the test by choosing further items from the pool is unlikely to be effective; and attenuation through unreliability is probably not a major effect. In this latter case consideration should be given to shortening the test.

Empirical Studies

Study 1

Test. The test used was the Australian Scholastic Aptitude Test (ASAT), Series B, a 100-item four-choice objective test, which purports to measure abilities relevant to the study of

[1] Copies of the computer program used to implement these calculations are available from the authors.

humanities, social sciences, mathematics, and science.

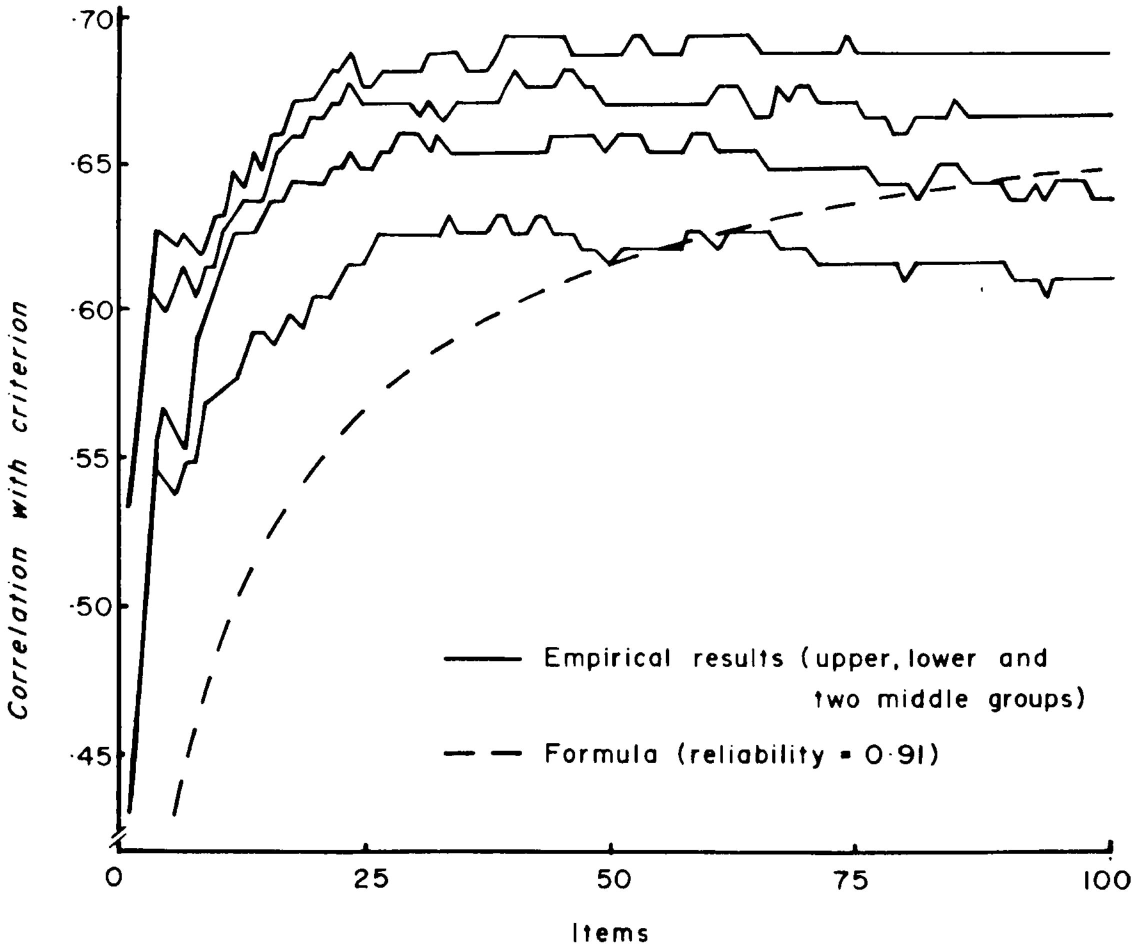
Criterion. The criterion was the matriculation marks aggregate, the sum of the five best scaled marks in the Tertiary Admission Examination, an achievement examination for 12th grade students intending to study at universities or other tertiary institutions in 1974.

Subjects. The subjects were nine samples of 500 students each, drawn randomly without replacement from a population of 6,479 students sitting for the Tertiary Admission Examination in Western Australia in 1973.

Method. Subtests were formed by successively deleting items with the lowest point-biserial correlation with total test score. The validity coefficients (product-moment) for the subtest totals and the criterion were then computed.

Results. The plots of the validity coefficients are shown in Figure 1, which also shows the plot

Figure 2
Validity as a Function of Test Length for ASAT-F



for estimated correlations computed by the Lord and Novick (1968) formula. There is only a slight decrease in validity down to about 40 items.

Study 2

Method. In this study the test used was ASAT, Series F, which is similar to the test in Study 1. The criterion was the same as for Study 1, but for students expecting to enter tertiary education in 1978. Subjects were nine samples of 500 students each, drawn as for Study 1 from 9,579 students sitting for the Tertiary Admission Examination in 1977. The method was the same as for Study 1.

Results. The plots of the validity coefficients and the formula estimates are shown in Figure 2. Again, reducing the length of the test to about 40 items would not appreciably reduce the validity.

Study 3

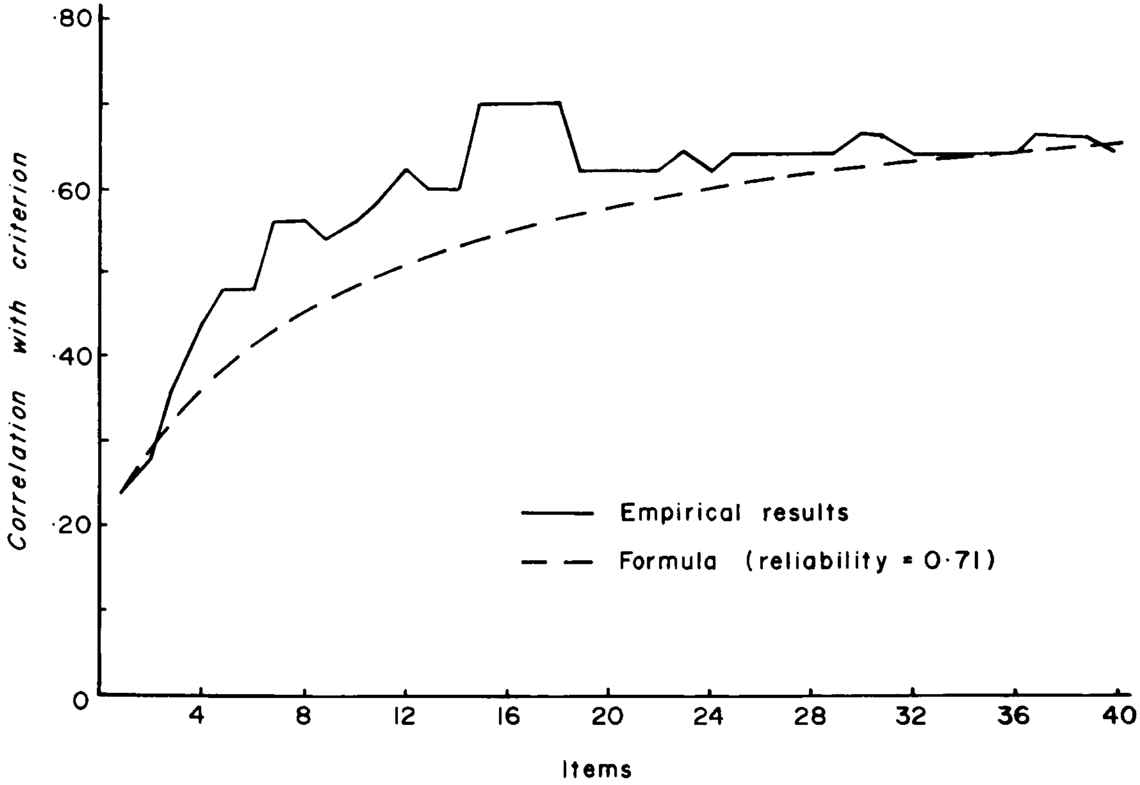
Method. For Study 3 the test was a 40-item four-choice objective test in economics. An examination of four essay questions in economics was the criterion. Subjects were 300 12th grade students. The method was as for Study 1.

Results. The plots of the validity coefficients and the formula estimates are shown in Figure 3. There was no substantial decrease in the validity down to about 15 items.

Study 4

Method. For Study 4 the test was a 40-item four-choice objective test in reading comprehen-

Figure 3
Validity as a Function of Test Length for an Economics Test



sion. The criterion was an examination in English consisting of two comprehension and comment essays and two essays drawn from a range of topics. Subjects were two samples of 750 12th grade students each. The method was as for Study 1.

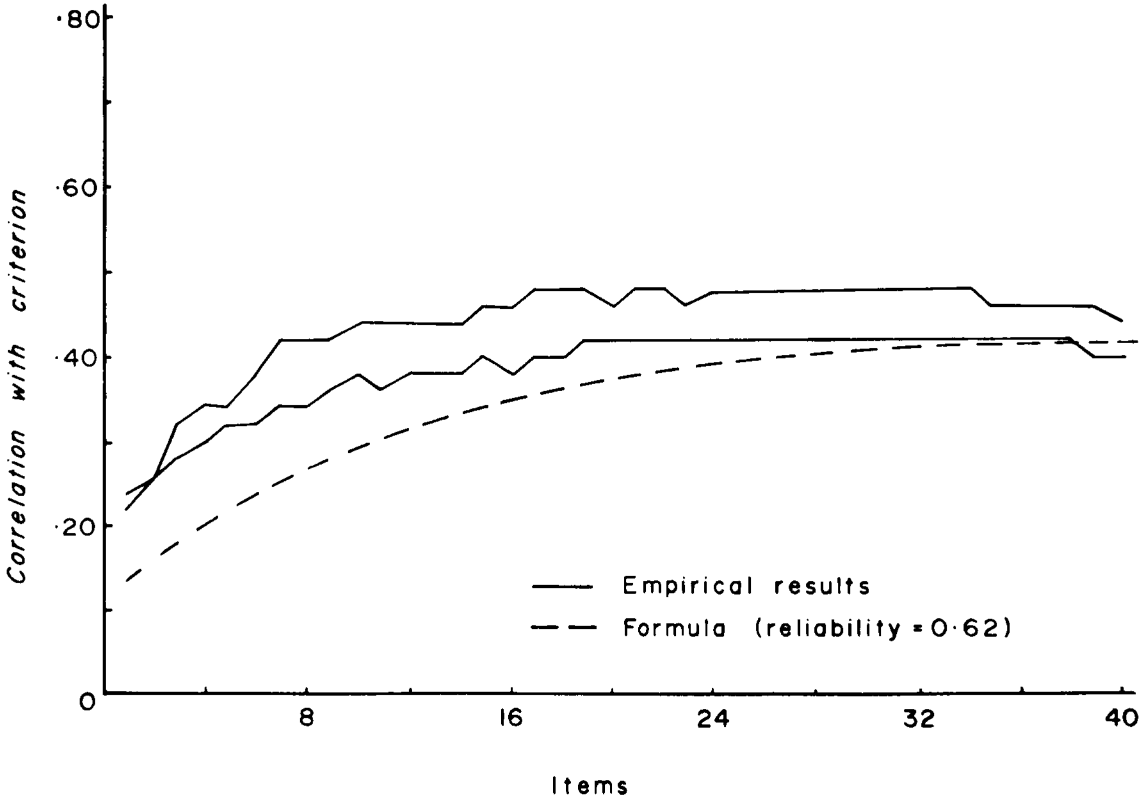
Results. The plots of the validity coefficients and the formula estimates are shown in Figure 4. The correlations were generally lower than in other studies, probably because the criterion was less relevant, but again substantial reductions in test length had little effect on validity.

Conclusions

Except for Study 1 the formula estimates did not match the empirical results. For Studies 2, 3, and 4 the slope of the formula estimates was gentler for the shorter test lengths and steeper for the longer test lengths than for the empirical validities. The tests used were manifestly not homogeneous, so that the results do not represent a fair challenge to the theoretical formula. They do indicate what may happen when the assumptions of the formula are not met.

The results suggest very strongly that all four tests were much above optimal length. Reductions of over 60% in testing time represent a highly valuable saving. It remains open whether other ability tests can be reduced so drastically without significant loss of predictive power. It is recommended that an empirical study of the relationship between test length and validity be carried out routinely and reported for any test for which predictive validity is claimed.

Figure 4
Validity as a Function of Test Length for an English Comprehension Test



References

Gulliksen, H. *Theory of mental tests.* New York: John Wiley, 1950.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

Spearman, C. Correlation calculated with faulty data. *British Journal of Psychology,* 1910, *3,* 271-295.

Acknowledgments

The assistance of the Tertiary Institutes Service Centre, in making the data available, is gratefully acknowledged.

Author's Address

Richard Bell, Research Unit in University Education, University of Western Australia, Nedlands, Western Australia 6009.