Higher Education 16: 33–36 (1987)

The use of correlation coefficients in test validation

URSULA A. HAUG* & D.H. IRVINE

School of Management Studies, Polytechnic of Central London, 35 Marylebone Road, London NW1 5LS, UK

Abstract. Although the Pearson Product Moment Correlation Coefficient, r, is commonly used in student selection, it is argued here, that it is frequently more appropriate to use the Bi-serial Correlation Coefficient r_{bis} for this purpose.

Introduction

Validating a selection test frequently involves correlating the scores of individuals taking the test with their criterion measures of performance such as examination marks. When both are available a Pearson Product Moment Correlation Coefficient, r, is generally calculated. As Anastasi (1982, p. 158) points out, this particular statistic "assumes that the relationship is linear and uniform throughout the range" and that "these conditions are generally met." Why then should we not use r?

Exam marking and exam boards

Unless standardised objective examination questions are being used, exam marks suffer from unreliability and other defects (Vernon, 1940 Ch. II) and the greater the subjectivity of judgements being made by the examiners the greater is the unreliability likely to be. Thus essay type answers are likely to be allocated marks which are less reliable than those given for quantitative answers. Various ways of dealing with this problem have been suggested (Vernon op. cit.) and are commonly used. For example the person marking scripts may choose to mark question by question rather than by marking all questions on the script, so that comparability is more easily achieved. He or she may also place scripts or questions in rank order and then recheck the marks. An internal moderator may also be used and, finally scripts may be further assessed by an external examiner and by an exam board. It is argued here, however, that in practice all these people are mainly concerned with fairness.

* Now working in Clinical Psychology.

They want to ensure that students are, as accurately as is humanly possible, divided into those who pass and those who fail. Sometimes other classifications may be made, such as grade of first degree or the award of "Distinction". If we take, for example, many Polytechnic courses, students are divided into those who pass (with perhaps Distinction) and those who fail.

Examiners are, therefore, concerned with paying special attention to students on the borderline, and it is their marks which are accorded close scrutiny and which are most likely to be given reliable marks at the conclusion of a board. Board members pay much less attention to marks awarded to students in the upper-middle and bottom or definite fail ranges, because a few marks here or there will make no critical difference to the outcome. Thus the concern of the exam boards to be fair is met by paying close attention to the marks of those on the borderlines. Because students achieving an award of "Distinction" are usually relatively few we can, for statistical purposes, treat them as a "Pass".

The purpose of selection in education

Selection tests, if used at all, are used in order to predict outcomes (Predictive Validity), but it would be unreasonable to expect a single test or even a test battery to predict a future exam mark. The best that we can expect is that the test will help in predicting those who will pass and the rest. Bearing in mind the manner in which examination answers are marked, how external examiners and boards operate, and what we hope to achieve from a selection test, it is argued that the biserial correlation coefficient is a more appropriate statistic for calculating the validity of a selection test. With this statistic selection test scores are correlated with a simple Pass/Fail criterion.

An example

Dangers associated with the use of a test on a population on which it has not been standardised are well known. Because of legislation on race and sex discrimination the topic has been given fresh prominence (Runnymede Trust/British Psychological Society, 1980). The example to be quoted is from research by the authors who were concerned with the particular education difficulties found by students in the United Kingdom whose native language is not English.

Research by Irvine (1977) using the Answer-in-Sentence (AIS) Test, Irvine (1974) suggested that non-native English students (nnEs) had a major problem in understanding spoken rather than written English. It seemed reasonable,

therefore, to expect that a test of speech intelligibility would be of use in predicting the likelihood of a non-native English-speaking student passing or failing the examination course. Haug (1981) undertook an extensive study to validate Test AIS for selection of non-native English speaking students, and to examine their particular learning problems (Haug, op. cit., and Haug and Irvine, 1983).

Research workers do not enjoy having their pet hypotheses and convictions proved wrong or in having difficulty in proving that they are right, Leapman (1980). It was a painful experience therefore, to find that, by using the Pearson r coefficient the expected and hoped for results were not always being obtained. Data was obtained from non-native English speaking students taking courses in Mathematics, Mechanical Engineering, Life Sciences, Management, English as a Foreign Language and other subjects. After much deliberation the authors concluded that Pearson's r was an inappropriate statistic for the reasons outlined earlier, and that a Bi-serial coefficient r_{bis} was more appropriate. Cheerfulness returned when the use of r_{bis} produced significant correlations. A single example taken from Haug, op. cit., is given as an illustration. The scores on the AIS Test were correlated with exam results of non-native English speaking students studying for the HND in Computer Studies.

$$N = 31 \qquad r = 0.091 \text{ p.n.s.}$$
$$r_{bis} = 0.408 \; p < 0.05$$

Thus when exam marks are used as the criterion there is an insignificant correlation, but when a Pass/Fail criterion is used the correlation becomes large enough to be significant, indicating that Test AIS has some value in predicting whether a non-native English speaking student is likely to pass or fail on this course, but not the mark which is likely to be obtained.

There is also the possibility that the relationship between AIS and results on some exam courses is heteroscedastic (Anastasi, op. cit., p. 158). Under this condition a non-native English student above a certain level of achievement in understanding spoken English is likely to succeed on his course. A higher level of achievement will not have much effect. Further collection and analysis of data using AIS should reveal whether the data are homoscedastic or heteroscedastic.

References

Anastasi, Anne (1982). *Psychological Testing.* London: Collier-MacMillan.

Haug, Ursula A. (1981). Student Selection with particular reference to applicants whose native language is not English. Ph.D., P.C.L. School of Management Studies.

Haug, Ursula A. & Irvine, D.H. (1983). "Assessment of Spoken English Language problems of non-native English speakers", in S.H. Irvine & J.W. Berry (eds), *Human Assessment and Cultural Factors.* New York and London: Plenum.

Irvine, D.H. (1974). "A new type of Speech Intelligibility Test", *Ergonomics* 17: 783–788.

Irvine, D.H. (1977). "The Intelligibility of English Speech to non-native English Speakers", *Language & Speech* 20: 308–316.

Leapman, M. (1980). "Diary of Impermeable Prose", *The Times*, 9th June, reprinted in *Bulletin of the British Psychological Society*, May 1982, pp. 215–216.

Vernon, P. (1940). "The Measurement of Abilities". University of London Press.