New Zealand Psychologist, 1982, *11*, 64-67

A Bayesian Approach to Weighting Progressive Achievement Test Scores

David C. Hughes and Brian Keeling

Education Department, University of Canterbury

Following a general discussion of Bayesian thinking as applied to the problem of estimating unknown quantities, a Bayesian approach is applied to the problem of weighting Progressive Achievement Test scores gathered in two consecutive years to obtain more reliable estimates of subjects' true scores. Adoption of such procedures is advocated.

The Bayesian approach to statistical inference and hypothesis testing has its roots in Bayes' theorem which dates back to the 18th century, but it is only over the past fifteen years or so that a resurgence of interest in Bayesian thinking in the social sciences has been noted. Recent texts illustrating the use of Bayesian statistics in psychological and educational research include those by Phillips (1973) and by Novick and Jackson (1974). The essence of the Bayesian approach lies in the revision of prior opinion regarding the probability of events in the light of new information. Using new data and Bayes' theorem (or some analogue of it), prior opinions are weighted with new information to yield posterior opinions. This process can be iterative: posterior opinions may in turn become new prior opinions which may be modified to form new posterior opinions following the gathering of new data, the re-application of Bayes' theorem, and so on.

Bayesian thinking has a wide range of applications in the social sciences, including hypothesis testing, inferences about relationships and differences between samples and populations, and the estimation of unknown quantities. Of course all of these are also the province of orthodox statistics and, indeed, so far as statistical *techniques* are concerned, there is little to distinguish Bayesian from orthodox statistics. The distinctive element is the particular way in which Bayesian thinking pervades the use of orthodox statistical techniques. The following example illustrates the special features of the Bayesian standpoint in the particular area of application with which this paper is concerned—the problem of estimating unknown quantities.

Consider a hypothetical situation which is reasonably close to the experience of many classroom teachers. A teacher wishes to assess the level of reading comprehension of his Standard 4 pupils using the nationally standardised Progressive Achievement Test of Reading Comprehension (Elley and Reid, 1969). In the course of his/her experience in teaching these pupils, the teacher will already have gathered much data on each pupil's reading level and would probably have little hesitation in forming estimates of their likely performances on the PAT. Some of these estimates may be crude and rather tentative—for example, "Mary should score at an above-average but not at an outstanding level", or "John is limited and he won't do very well". Bayesians would not be satisfied with gross qualitative estimates of this nature; they would require teachers to be more conscious of these predictions and to quantify them rigorously. For example, Bayesians would press the teacher to specify for each pupil a median score estimate (expressed probably as a within-age percentile in the case of the PAT) where higher or lower scores are equally likely. This is a demanding but not impossible task for the teacher, and it is made easier by also requiring the teacher to specify a particular "credible interval" around that median score estimate: the wider the credible interval, the less the risk of an incorrect estimate. Thus, in terms of prior probability, the teacher may formulate the following estimate of Mary's likely PAT performance—"Mary should score at around the 70th percentile but the odds are nineteen to one that her score will lie between the 50th and 85th percentiles. These odds of nineteen to one represent the '95% credible interval' for Mary's obtained score on the PAT. Median score estimates and associated credible intervals would be created for all the other pupils. The PAT test is then administered and scored following which the results are compared with the teacher's estimates.

At this point it is necessary to remind readers that this illustration is concerned with the application of Bayesian thinking to the problem of estimating unknown quantities, in this case estimation of the "true" level of each pupil's reading ability. Both the teacher's estimates and the PAT results can be regarded as indicators of this true level but it is certain that neither of them is infallible. Teachers' estimates are inevitably based on limited contact with each pupil's reading behaviour and may reflect, to a greater or lesser degree, sundry biases and irrelevancies. PAT scores, though technically very reliable, are based on just a single sample of reading passages and test items administered at one particular point in time. Through an analogue of Bayes' theorem, Bayesians demonstrate how both these imperfect estimates of true reading ability may be combined to yield posterior estimates which are more accurate and dependable than either alone. In other words, Bayesians show how probabilities (the teacher's estimates) are modified by new data (PAT scores) to yield posterior probabilities which are more accurate estimates of unknown quantities ("true" reading ability levels). In Bayes' theorem these posterior probabilities are derived from the weighted sum of the standard error of estimate (teacher's estimates) and the standard error of measurement (PAT scores); the greater the standard error the less weight that measure receives in deriving posterior probabilities.

The next section of this report gives a practical demonstration of the application of Bayesian thinking (and of Bayes' theorem) to the estimation of unknown quantities. It differs from the example just given in that prior probabilities are derived from earlier test scores not from teachers' estimates.

Table 1

Intercepts, Regression Coefficients, Standard Errors of Estimates and Standard Errors of Measurement (Form II) for the three PAT Tests.

	a	b	δ_{est}	δ_{meas}
RC	3.26	.90	5.1	2.8
RV	7.58	.87	5.9	3.3
LC	7.82	.72	4.5	2.8

Method

The subjects were the 638 pupils from three intermediate schools in Christchurch used in the study of difference scores by Hughes and Tuck (1978). The study included all pupils from these three schools for whom there were complete data on the PAT: Reading Comprehension (RC), Reading Vocabulary (RV) and Listening Comprehension (LC) tests administered by the schools in 1975 and 1976. Details of the data collection and spot-checking of the answer sheets may be found in Hughes and Tuck (1978).

Results

The regressions of the Form II raw scores on the Form I raw scores were calculated separately for RC, RV and LC using regression equations of the form $\tilde{Y} = a + bX$ (X = Form I raw score; $\tilde{Y}$ = estimated Form II raw score). The equations were then used to estimate pupils' Form II performances on all three tests from their performances on the equivalent Form I tests. Table 1 shows the values of the obtained intercepts (a), regression coefficients (b) and standard errors of estimate (δ_{est}). Table 1 also shows the standard errors of measurement (δ_{meas}) for each of the tests at the Form II level as given in the relevant PAT manuals. The best estimate of the pupils' true ability at at the time of testing in Form II is obtained by weighting the actual Form II scores and the estimated Form II scores together in inverse proportion to the squares of the respective standard errors (Thorndike, 1980). The standard errors of these weighted estimates is given by the formla

$$\sqrt{\frac{1}{\frac{1}{\delta^2_{est}} + \frac{1}{\delta^2_{meas}}}} \quad \text{(Thorndike, 1980)}$$

Table 2

Weights and Standard Errors of Weighted Scores for the three PAT Tests

	Weight	δ_{weight}
RC	3.32 : 1	2.45
RV	3.20 : 1	2.88
LC	2.58 : 1	2.38

Table 2 shows the weights obtained for each of the tests and the standard errors of the weighted scores (δ_{weight}). For example, to find the weighted scores for RC the actual and estimated Form II scores are weighted in the proportion 3.32 to 1.0 respectively.

A real example from the data shows how the procedures operate. In 1976 'Patrick' obtained the scores shown in Table 3. Table 4 shows Patrick's Form I scores.

Entering Patrick's raw scores in the regression equations using the intercepts and regression coefficients shown in Table 1 gives estimated Form II scores of 18.56, 45.86 and 23.66 for the RC, RV and LC tests respectively. Applying the weights in Table 2 to the actual and estimated Form II scores gives the weighted scores shown in Table 5.

Two points can be made about the weighted scores in Table 5 in comparison with the Form II scores in Table 3.

1. The standard errors of the weighted scores (see Table 2) are lower than the standard errors of measurement of the Form II scores (see Table 1) indicating that the weighted scores are more reliable than the Form II scores.

Table 3

Scores for Patrick on the three PAT Tests in Form II

	Raw Score	Level Score	Percentile Rank
RC	20	7B	45
RV	40	7B	56
LC	35	9C	86

Table 4

Scores for Patrick on the three PAT Tests in Form I

	Raw Score	Level Score	Percentile Rank
RC	17	6C	39
RV	44	7C	71
LC	22	5C	27

Table 5

Weighted Scores for Patrick on the three PAT Tests at Form II

	Raw Score	Level Score	Percentile Rank
RC	20 (19.67*)	7B	45
RV	41 (41.40)	7A	59
LC	32 (31.83)	8B	71

*Scores in parentheses are the weighted scores correct to two decimal places.

2. The pattern of weighted scores is different from the pattern of Form II scores. Hughes and Tuck (1978) have discussed a number of methods to test for significant differences between test scores. However for purposes of illustration it seems appropriate to select the method suggested by the test authors in the Listening Comprehension manual (Elley and Reid, 1971, p. 15). This 'rule-of-thumb' method involves comparisons between level scores; a difference of two or more levels on pairs of tests suggests the possibility of a real difference in ability. When this method is applied to the Form II data in Table 3 the LC score is found to be significantly higher than the other two scores. However, the weighted scores in Table 5 are not significantly different.[1]

In other cases, of course, the weighted scores will confirm a significant difference found between the Form II scores. For example, 'Charles' showed a significant difference between his Form II listening and reading scores (LC = 10E, RC = 7B and RV = 7D). Because his Form I scores showed a similar pattern his weighted

[1] The level score method of detecting differences is based on the standard errors of the differences between scores. The size of the standard error of a difference is related to the reliabilities of the tests involved. Given that the reliabilities of the weighted scores are greater than the reliabilities of the Form II scores the standard errors of the differences of the weighted scores will be slightly less than those of the Form II scores. Thus, using the same criterion with the weighted scores as is used with the Form II scores increases the probability of making a Type II error. However, given the rule-of-thumb nature of the level score method this is not important in practice.

scores changed very little (LC = 10F, RC = 7B and RV = 7D). One could therefore be confident that a real difference in ability existed.

Conclusion

The weights used above are based on the administration of Form A in Form I and Form B in Form II. However, it seems likely that similar weights would be found with the administration of the test forms in the reverse order. If more data could be gathered to allow the calculation of weights for other PAT tests (either RC, RV and LC at other class levels or other pairs of tests, e.g., PAT Mathematics at Form I and II) we would have a relatively simple procedure for using existing data to better estimate the abilities of pupils. Given the substantial amount of time and money invested in PAT testing the modest additional time required to compute the more reliable weighted scores would seem fully justified.

References

Elley, W. B., & Reid, N. A. (1969). *Progressive Achievement Tests: Reading Comprehension and Reading Vocabulary.* Wellington: N.Z.C.E.R.

Elley, W. B., & Reid, N. A. (1971). *Progressive Achievement Tests: Listening Comprehension.* Wellington: N.Z.C.E.R.

Hughes, D. C., & Tuck, B. F. (1978). The stability of scores and difference scores on the Progressive Achievement Tests of Listening and Reading Comprehension. *New Zealand Journal of Educational Studies,* 13, 67-77.

Novick, M. R., & Jackson, P. H. (1974). *Statistical Methods for Educational and Psychological Research.* New York: McGraw-Hill, Inc.

Phillips, L. D. (1973). *Bayesian Statistics for Social Scientists.* London: Thomas Nelson and Sons Ltd.

Thorndike, R. L. (1980). The application of Bayesian thinking to educational measurement problems. *A.C.E.R. Bulletin for Psychologists,* N28, 9-16.