AGE STANDARDISATION AND SEASONAL EFFECTS IN MENTAL TESTING

By H. GOLDSTEIN and K. FOGELMAN

(*National Children's Bureau, 8 Wakley Street, London, ECIV, 7QE*)

Summary. It is pointed out that there are two distinct types of age standardisations for tests of school attainment. One is concerned with allowing for differences in age between children tested at the same time, and the other is concerned with allowing for average changes in test scores with time of year. Standardisation procedures do not normally take account of this distinction. It is also shown that there is no increase in attainment scores between April and July among 11-year-olds in the last year of primary school. This 'seasonal' effect creates further difficulties in providing satisfactory age standardisations and it is suggested that full age-time standardisations should be carried out by selecting standardising samples over whole age ranges at different times of year.

INTRODUCTION

Between 1962 and 1972 the National Foundation for Educational Research increased its sales to local authorities of standardised ability and attainment tests from 2 to 3·5 million.

Whatever the reasons for this increase, it does indicate an increasing use of, and presumably reliance upon, such tests. Whether these tests are used for research purposes or to allocate individual children to schools or groups, it is clearly important that the standardisation of the tests should be soundly based. The purpose of the present article is to point out that, in two respects, present standardisation procedures may give rise to misleading results.

Since any one test may be applied to children in an age range within which the average score changes, an adjustment for age is necessary if the scores obtained by children of different ages are to be compared. In addition to such an 'age standardisation' it is common to 'normalise' the raw scores by transforming them to give a distribution with a mean of 100 and a standard deviation of 15. For present purposes, however, we are only concerned with the age standardisation, and the use of a normalising transformation will not affect our arguments.

The procedure which is usually followed in constructing norms for a test involves the administration of the test to a large representative sample of children at a given point in time (usually over a period of a few weeks). The change in average score with age is then estimated from this sample, and it is this estimate which is subsequently used to adjust or standardise for age.

This procedure, however, is inadequate since the test score will depend not only on the *relative ages* of a population of children, that is the differences in age at a given point in time, but also on the *time during the school year* when testing takes place. This is shown formally in the appendix using a mathematical model, and may be illustrated as follows.

If we consider a child tested at different times during the school year (the discussion will be restricted to one school year) then his score will be

expected to change with time. Suppose that the expected or average change for all children is x score points per month. Such change will be referred to as ' longitudinal ' change and if x is known it may be used to ' adjust ' children's scores to a common time of year—a ' longitudinal ' adjustment.

As has been pointed out, however, in the standardisation procedure which is usually adopted it is not x which is estimated but y, say, where y is the average change in score with age for children of different ages measured at a single point in time, a ' cross-sectional ' adjustment. Although both the adjustment for age and that for time during the school year are measured in the same units, they are logically distinct. (In practice, standardising samples often span more than one year group so that ' longitudinal ' changes are also included and the final adjustment will, therefore, be an average of the ' longitudinal ' and ' cross-sectional ' changes.)

It is reasonable to suppose that average scores on attainment tests will be closely linked to the teaching programme. For example, towards the end of primary schooling all the children in a class or year group will have been exposed to very nearly the same amount of teaching and as a result the average test score *might* be expected to vary more with calendar month than with month of relative age difference between children, that is, we should expect x to be greater than y. One result of this would be that children tested early during a school year would tend to be penalised. However, the next section demonstrates that x does not remain constant throughout the school year and this should be taken into account when undertaking standardisation. It may also be true that the value of y depends on time of year.

The implications of this are clear : namely, that purely ' cross-sectional ' age standardisation will, in some circumstances, be inadequate and that ideally what is needed is a standardisation which also takes account of the change in score as children progress through the school year ; what might be termed an ' age-time ' standardisation. We now discuss the way in which the average test score actually does change during the school year.

NON-LINEAR AGE TRENDS AND SEASONAL EFFECTS

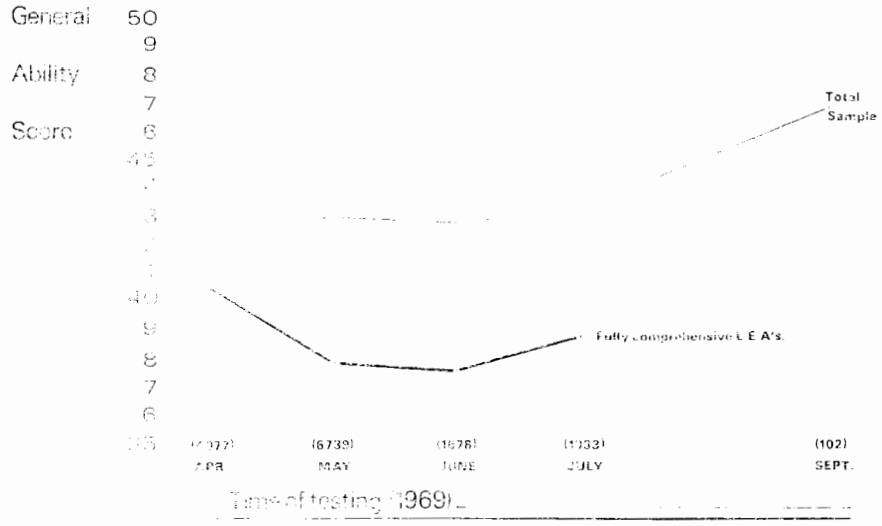
The equations presented in the appendix assume a linear relationship between test score, age and time. On the face of it, it seems reasonable to assume that the later in the school year a child is tested the greater will be his/her knowledge and, therefore, the greater will be the score obtained on an appropriate test.

However, data from the National Child Development Study (Davie *et al.*, 1972) do not support this.

As part of this study, 13,659 children born during the same week, namely, 3rd—9th March, 1958, were given tests of general ability, reading and mathematics between April and September, 1969. The general ability test was provided by the NFER and consisted of verbal and non-verbal items. The reading test was a parallel test to the Watts-Vernon. The mathematics test was a modified version of the Vernon graded arithmetic test. Since all these children are born in the same week there are no ' cross-sectional ' age effects. Figure 1 shows the mean general ability score (out of a total of 80) by month of testing. Similar results were found for the other two tests.

FIGURE 1

NATIONAL CHILD DEVELOPMENT STUDY: MEAN GENERAL ABILITY BY MONTH OF TESTING.
(Numbers of children in parentheses.)



During the three months of the Summer term from April to July there is a striking absence of an increasing trend, although according to available norms (NFER, 1972) a mean increase of about 1·6 score points would have been expected. This implies that if an age standardisation based on an expected increase is carried out, in contrast to what would be expected on the basis of the argument in the preceding section, those children tested *later* will tend to be penalised. Consider, for example, a child who is actually at the 50th centile with a score of 43·2 measured in July. Assuming an average gain of 1·6 points between April and July, and if April were chosen as the base month for test standardisation, then such a child would actually be classified as being at the 46th centile. In addition it should be noted that the estimate of 1·6 points is based on a 'cross-sectional' age standardisation and not as required on a longitudinal standardisation. For an individual child, such a misclassification is unlikely to lead to serious consequences. On the other hand, an average misclassification of this order may be important when applied to large groups of children, such as all those in one school. Furthermore, if the mean scores before April also do not increase linearly with time, an adjustment over a longer period than three months might lead to greater errors of misclassification than that given in this example.

It should be recognised, however, that in practice the biases which may be introduced into the standardisation of tests for national use by variations between local education authorities, regions, etc., will normally be larger than biases arising from any failure to take account of the effects described above. On the other hand, *within* one school or local authority, the former biases will be constant and biases arising from seasonal and 'longitudinal' effects could become important.

There is also evidence from studies in the United States which tend to confirm the present findings. For example, Beggs and Hieronymus (1968)

show a smaller than average increase in score on tests of arithmetic and language between May and September, for children in the 3rd–6th grades.

Several explanations for the present results are possible.

First, it is possible that the time of testing the children in this sample is associated with other factors influencing test scores. The dates of testing depended basically on the administrative convenience for the schools and local education authorities concerned. The possibility exists, therefore, that, say, the most co-operative and enthusiastic schools and local authorities, where the children would be tested earlier, would also tend to contain the most able children. It has not been possible to test this directly, but the date of testing has been related to the child's social class, the size and type of school, and the region of the country. None of these factors was associated with the date of testing.

A second possibility is that these figures reflect an actual lack of development during this period. To put it bluntly, perhaps children do not learn anything between April and July of their last year of primary schooling. Undoubtedly, in many schools there is a lessening of the academic demands made on their pupils towards the end of the school year, and this may be more marked in the year in which the pupils are about to change schools. (Not all those tested were about to change school. 10 per cent were in Scottish schools and, therefore, not due to transfer to secondary schools until the following year. Also 4 per cent were in independent schools and many of these would stay until the age of 13). However, it seems extremely unlikely that such an effect would begin to appear so early in the school year as to account for a lack of increase in score between April and May. Furthermore, the large increase between July and September would remain to be explained. This increase is more easily reconcilable with a third explanation. It may be that the skills which the tests are designed to measure continue to develop during this period, but the motivation to display them in the testing situation decreases. Perhaps the desire to do well on such tests fades as the end of the year approaches. This seems slightly more plausible than the previous explanation but it still does not seem adequate to explain the appearance of the levelling-off so early in the school year.

Each of the explanations so far considered has assumed that scores are in some way 'artificially' decreased towards the end of the period considered. There is, of course, a second possibility—that scores have been 'artificially' increased at the beginning of the period. In areas where tests are still administered as part of the selection procedure, this is usally done in February or March. For most children this will have been preceded by practice on tests of attainment and general ability similar to those used in the Study. This practice will, to a limited extent, have raised scores on the tests (see Yates, 1953, and Vernon, 1954). However, this effect will decrease once the practice ceases and this would lead to the results reported here.

The bold line in Figure 1 represents those children (475 in all) in local authorities which had fully comprehensive education and where, therefore, we can be reasonably sure that the children did not take an '11+' examination. It can be seen that the same pattern persists. It would be wrong to assume, however, that these children had not experienced any tests of this kind, since even within a comprehensive system it is common for such tests to be used at the time of transfer to secondary schools. To test this hypothesis properly we should need to know, for each child, the extent of his familiarity with tests and testing situations. This information, unfortunately, we do not have.

CONCLUSIONS

Whatever may be the explanation, the data reported in this paper offer strong evidence for a seasonal effect with children of this age on the tests used. Given the diversity of tests (general ability, reading and mathematics) it seems likely that this ' seasonal ' pattern would be found in many tests of attainment. Further research would be necessary to show whether similar results would appear for children of different ages.

There are important implications here for the production and use of normative data for tests of attainment. Present procedures assume not only that (cross-sectional) age gains are independent of the date of testing, but that age gains based on such ' cross-sectional ' data can also be applied ' longitudinally.' We have tried to indicate that these assumptions are questionable.

It follows that a single age adjustment is inadequate and that standardisation procedures should, where possible, take account of the time of year of testing as well as the age of the child at that time. The age-time relationship for any given test would need to be determined by selecting standardising samples over the whole age range at different times of year.

ACKNOWLEDGMENTS—This work was partly supported by grants from the Social Science Research Council and the Department of Health and Social Security to the National Children's Bureau. We would like to thank Mr. M. J. R. Healy, Dr. R. Sumner, Miss J. Tarryer and our colleagues at the National Children's Bureau for their valuable comments.

REFERENCES

BEGGS, D. L., and HIERONYMUS, A. N. (1968). Uniformity of growth in the basic skills throughout the school year and during the summer. *J. educ. Measurement*, 5 91-97.

DAVIE, R., BUTLER, N. R., and GOLDSTEIN, H. (1972). *From Birth to Seven*. London : Longman.

NELDER, J. A. (1968). Regression, model-building and invariance. *J. R. statist. Soc.*, A, 131, 303-329.

NFER (1972). Personal communication from Miss Jill Tarryer.

VERNON, P. (1954). Symposium on the effects of working and practice in intelligence tests. *Br. J. ed. Psychol.*, 24, 57-63.

YATES, A. (1953). Symposium on the effects of working and practice in intelligence tests. *Br. J. ed. Psychol.*, 23, 147-154.

(*Manuscript received 27th June*, 1973)

APPENDIX

It is convenient to consider a time span restricted to one year corresponding to the school year. During this year the school year group (e.g., children in the last year of junior school) will be taken as defining the population. Thus, at any one point in time the age range of the children will also (normally) be one year.

Suppose that the relationship between test score and age for an individual child is linear, i.e.,

$$y = \alpha + \beta x \qquad (1)$$

Since the rate of change of score with age will vary from child to child, β will be a random variable over the population of children. In addition, if children are classified by other factors, for example by sex, or by region of the country, then it will be necessary to modify equation (1) by introducing further terms to describe these classifications. The particular classification of interest in the present paper is the birth date of the child or, equivalently, the age of the child at a given point in time.

Suppose that equation (1) holds except that α depends on birth date, or for convenience, the age of the child at the beginning of the school year (t), and suppose also that this relationship is linear. This gives

$$y = \alpha^1 - \beta^1{}_1 t + \beta x \qquad (2)$$

Here $\beta^1{}_1$ is a fixed coefficient but since there is variability among children with the same birth date, α^1 will be a random variable.

We may write equation (2) for the i^{th} child, as

$$y_i = b_{oi} + \beta_1 x_{1i} + b_{2i} x_{2i} \qquad (3)$$

where

b_{oi} is a random coefficient $E(b_{oi}) = \beta_o$
b_{2i} is a random coefficient $E(b_{2i}) = \beta_2$
β_1 is a fixed coefficient $= \beta^1{}_1 + \beta_2$
x_{1i} is the difference between the age of the i^{th} child and the mean age for the year group at a given point in time ($-0{\cdot}5 \leqslant x_{1i} \leqslant 0{\cdot}5$).
$x_{2i} = x - x_{1i}$ i.e., the mean age of the year group at the time of testing, or simply calendar time of year measured from the start of the school year.
Var $(b_{oi}) = \alpha_{oo}$ Var $(b_{2i}) = \alpha_{11}$, cov $(b_{oi}, b_{2i}) = \alpha_{o1}$

Assume also that b_{oi}, b_{2i} are normally distributed. Equation (3) is a regression model of the second kind as discussed by Nelder (1968) and uses Nelder's notation.

The parameters β_1, β_2 represent the average change of score with relative age and with time of year, respectively. For a given value of x_{2i}, i.e., at a fixed time of year, the score will depend on x_{1i} only and β_1 may be used to ' adjust ' an individual's score to a common value. This is referred to as ' cross-sectional ' adjustment and is the method by which attainment test scores are commonly standardised, as explained in the main text. It is not, however, appropriate to use β_1 for adjusting the scores of children measured at different points in time, since the score will then depend also on β_2. For a given value of x_{1i}, i.e., children of the same relative age measured at different points in time, β_2 is used for adjustment. This is referred to as ' longitudinal ' adjustment. It follows that where age standardisation is required over a time period for children with different relative ages, knowledge of both the longitudinal and cross-sectional adjustments is necessary and a standardising sample should be adequate to estimate the size and nature of these and possible interactions between them. The linearity assumptions should also be critically examined in view of the ' seasonal ' effects described in the main text.

Example.

A sample of 239 children, 160 in the last year of primary school and 79 in the first year of secondary school, were given a mathematics test which had been used for 11-year-olds in the National Child Development Study. These school classes were selected with the intention of providing representative samples of their age groups, and all the children were tested at the same time (end of September, 1972).

Equation (3) is appropriate since there are both ' cross-sectional ' (within year groups) and ' longitudinal ' (between year groups) effects within the sample. However, since the longitudinal effect is based on only 2 times, the model (3) is under-identified, since it turns out that there is one more parameter to be estimated than there are estimating equations, and an assumption must be made about the values of the variances and covariance. One assumption which satisfies the likelihood equations is to set $\alpha_{11} = 0$, which gives (3) as the ordinary fixed effects model. Hence the results in Table 1 are, for simplicity, presented in terms of a fixed effects model.

TABLE 1

REGRESSION COEFFICIENTS AND ANALYSIS OF VARIANCE TABLE.

x_2 is measured from the start of the last year of primary school (mean age of class = 10·5 years).

Coefficient	Fitted Value	Standard error	Mean square ratio
β_0	15·9		
β_1	1·8	1·5	1·3
β_2	4·7	1·1	18·4*
Total variance = 66·3 Residual variance = 61·3			

Significance levels : * $P < ·001$, otherwise $·05 < P$.

It will be seen from Table 1 that the estimate of β_2 (4·7) is greater than that of β_1 (1·8) which is not statistically significant. However, the difference $\beta_2 - \beta_1$ is not significantly different from zero (χ^2 (1 df) = 2·1) so that this small sample does not provide sufficient evidence on which to make a reliable inference concerning this difference. Since the estimate of β_2 is based on two occasions a year apart, it represents an average 'longitudinal' effect, and as pointed out earlier there are 'seasonal' variations during the school year.