J. Child Psychol. Psychiat., Vol. 21, pp. 363 to 369.
Pergamon Press Ltd. 1980. Printed in Great Britain.

CRITICAL NOTICE*

Fifteen Thousand Hours. MICHAEL RUTTER, BARBARA MAUGHAN, PETER MORTIMORE and JANET OUSTON. Open Books, London, 1979, pp. 279. £3.50.

Review of Fifteen Thousand Hours

I THINK that this is a very important, and I hope, seminal book. Major British educational research projects are pitifully scarce. The strengths of this one are that, as one would expect from the Rutter stable, it was meticulously planned and executed; a lot was known about the children before they entered secondary school, so that it is in a sense a longitudinal study; systematic observations were made within the school, and related to other measures; and the authors addressed themselves to very significant questions. Their first message—that the characteristics of a school powerfully affect the behaviour and achievements of its pupils—should bring comfort to any teacher beset by self-doubt. True, as parents all of us know this to be true. But as teachers or social scientists we tend to be more influenced by the evidence of the effect of social class and I.Q. on school performance. "Hard" evidence that within a social class and I.Q. group schools *do* count is therefore heartening.

It should, perhaps, be pointed out that some doubt must remain about the extent to which the study did in fact establish that differences in the children's achievements reflected school rather than family influences. Although the authors have statistically equated the schools for fathers' occupation and children's school records at age 10, other possibly important family characteristics were not taken into account. For example, some working-class parents with children of average ability are more knowledgeable about and interested in education than others. If these families select a secondary school with a "good" reputation and thereafter give their children more educational support, the children's school career will depend to a greater extent than the authors allow on parental as well as school characteristics.

That said, the study certainly shows that children of the same social class and similar test scores at age ten fare very differently in different secondary schools. The authors' choice of "outcome" variables (public examination results, orderly classroom behaviour, truancy and delinquency rates) reflects the aspirations of most parents. Whatever our social class or race, we usually hope our children will get "O" levels, and keep out of trouble with the school authority and the police.

Nevertheless, these "outcomes" are by any standard limited—no school, surely, would state as its *aims* that the children should get a few "O" levels, and keep out of trouble. Wouldn't most teachers and parents want to add, at the least, such aspirations as that the child should enjoy school, learn to relate tolerantly and co-operatively to other children, be helped to feel good about himself, feel able to tackle the difficulties and problems he meets, develop wide interests, and so on? And shouldn't we be able to state academic aims for the 30% of children who are not going to get any "O" levels or CSE passes?

Still the authors can hardly be blamed for not addressing these issues, when there is so much general confusion about them. On the one hand some heads act as though they were running grammar schools, and stake their reputation on the number of "O" and "A" levels the school clocks up. On the other hand, radical teachers who argue that the function of schools is to validate and perpetuate the class system find it difficult to articulate positive aims for the school. What *are* secondary schools aiming at? And how are we to assess their success?

*Requests for reprints: Readers wishing to obtain reprints of *all four* papers in this Critical Notice should write to any *one* of the authors listed below.

Dr. B. Tizard, Thomas Coram Research Unit, 41 Brunswick Square, London WC1N 1AZ, U.K.

Professor H. Goldstein, Department of Statistics and Computing, Institute of Education, Bedford Way, London WC1H 0AL, U.K.

Professor M. Rutter, Department of Child and Adolescent Psychiatry, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF, U.K.

The authors' second, and more important message, is that differences between schools in children's achievements and behaviour can be shown to be related to differences in school and classroom characteristics. Again, this message is of a practical interest to parents—when we come to select a secondary school we would like to be able to identify the characteristics of the school which will meet our child's needs. For educationalists and social scientists, an attempt to measure the processes within the school which result in particular outcomes is of considerable importance.

As the authors freely admit, their "Process variables" were a very rum collection of forty-six items. They ranged from whether the children's work was displayed on the wall (related to examination successes and delinquency rates, but not to classroom behaviour or truancy) to the number of the teacher's disciplinary interventions (related to classroom behaviour but not to examination success).

In a thought-provoking discussion, the authors conclude that these items were, in fact, measuring aspects of the school's "ethos". This concept brings us back to the consideration of what are the expectations within the school for both teachers and children, and by what mechanism these expectations are transmitted. That the measures chosen are crude, and the mechanisms mostly guesses, is undeniable. What matters is that the authors did succeed in measuring aspects of school life often thought to be unmeasurable, and that the study does address itself to the crucial question of how a school's functioning affects children. It is to be hoped that it will provoke further studies in this area. (A small caveat: nowhere do the authors present a list of the data about each school. The importance of preserving anonymity is clear, but one or two identifying variables could have been omitted—e.g. size of school, voluntary or maintained status—and the readers would still have been able to assess and re-analyse much of the data for themselves.)

BARBARA TIZARD

Fifteen Thousand Hours: *A Review of its Statistical Procedures*

Introduction. Few educational research studies are so entirely free from methodological weaknesses that a sufficiently determined and diligent critic cannot find enough material with which to mount a plausible attack. The measure of a good critique, however, is not its achievement in exposing all the weaknesses of a study, but its illumination of those weaknesses which really matter. It will be useful to expand a little on this topic before discussing the present study.

Suppose it was discovered that the authors of a report had calculated a standard deviation incorrectly. At one extreme this error might be so serious that its correction would also involve the reversal of several important conclusions. Clearly such an error should be publicised so that unjustified inferences are not drawn. If, however, the error were not so serious and if the replacement of the incorrect by the correct value were to change nothing of substance, we might take one of two views. Firstly we could recognize that mistakes will always occur, satisfy ourselves that this was not a serious one and say nothing more. Alternatively, we might feel that the authors of the report should have been more careful, and we might then wish to treat the mistake as evidence contributing to an existing opinion of serious incompetence. Examples of both types of response can readily be found, and each has some legitimacy. There is a third type of response which takes whatever errors of such kind as can be detected, and uses them in an attempt to undermine the credibility of a study, even though the sum total of the errors could not be said seriously to threaten any conclusions.

Having made the above distinctions, the real difficulty is to recognise them in practice. In particular it is extremely important to distinguish minor blemishes from major and potentially catastrophic ones. A failure to do this has often lead to a debate losing its audience, who either cannot follow its detail or do not feel that there would be profit in doing so. As in all good criticism, the point is to evaluate the importance of any defects and then to communicate their essence to the intended audience in terms which are intelligible. I shall attempt to do this in the following comments, bearing in mind that the typical reader will not be a practising statistician. I should also emphasise that this review is concerned with the statistical and design methodology of the study and not directly with such matters as the choice of measures, etc.

A brief outline of the study. A "cohort" of approximately 2000 children were followed from before their entry to secondary school until their first public examinations. For these children, who attended 12 Inner London secondary schools, there were "outcome" measurements of attendance, delinquency, behaviour and exam results and the basic analyses compared average values for the 12 schools. One set

of analyses then studied the extent to which differences between schools could be accounted for by various characteristics of the school, after making allowance for possibly differential intakes at 11 years. A second set of analyses, this time using the school rather than the child as the unit of analysis, also looked at the way in which various characteristics of the schools were associated with the outcome measures.

Adjusting for intake differences. The authors rightly point out that causal inferences from survey data concerning "outcome" differences between schools, are greatly strengthened if appropriate adjustments can be made for pre-existing differences between the intakes to the schools. Thus, comparatively good exam results at 16 years might simply reflect an academically selective intake at 11 years and if it were possible to "equate" such intake factors between schools, and if the outcome differences are little changed by this, then we would feel more justified in attributing these differences to other measured factors. The principal difficulty with this approach lies in ensuring that we have been able to measure *all* the relevant intake factors. Rutter *et al.* use, principally, verbal reasoning score and occupation group prior to intake, justifying these on the grounds that they were the best predictors of the outcome variables. Previous research, however, has shown that there are many other variables, family size, subject attainments, etc., which are also at least as good predictors, and are also associated with selective intake to schools. It is curious that the authors do not even discuss the possibility that the variables they use may only make a partial adjustment, and it remains an open question as to how much of the subsequent differences between schools could be attributed to additional variables not used. This seems to be a substantial criticism in view of previous research and implies extra caution in interpreting the results of the study. An incidental technical point here is that it is the *within-school* correlation of intake variables with outcome which it is appropriate to study and not the overall correlation used by the authors.

If we turn to one of the statistical analyses which uses the intake adjustment procedure, we discover some technical inadequacies which seem to cast further doubt upon some of the conclusions and which also seem to contribute to a general picture of a less than fully competent technical expertise. Table 5.9 in Appendix G reports an analysis of variance which shows that verbal reasoning (V.R.) score accounts for 28.6% of the variance in the outcome examination score and that there are significant remaining differences between schools after adjusting for V.R. The obvious next question is how many of these school differences can be accounted for by, say, the "process" or "physical" or "ecological" factors measured during the study. This is done in another analysis given on page 171 in Table 9.4. Here, however, V.R. score only accounts for 14.5% of the variance (thus performing a less adequate adjustment), and indeed the total variance is only 25.9% and this includes that taken up by parental occupation, process score, etc. Needless to say, in this analysis, the process, etc. factors had significant effects after adjusting for V.R. and occupation. There arises the question of whether, given the full adjustment of which V.R. is capable, as in Table 5.9, the process score, etc. would still be related to outcome. The authors, however, seem unaware of the relationship between the analyses in Tables 5, 9 and 9.4 and the fact that they could easily have used V.R. in Table 9.4 in just the way it is used in Table 5.9. (It would incidentally be informative if we could be given the sizes of the differences between schools, etc., rather than just the percentage of variance accounted for.) Admittedly, the authors do mention that there are some technical problems associated with using linear models (analysis of variance and regression analysis) but these are not as unsurmountable as they seem to suggest. Instead, they prefer to introduce "log-linear" models which utilise the outcome measures in terms of "percentage of good attenders", "good exam results", etc. rather than "mean attendance score", "mean exam score", etc. The log-linear models, however, suffer from many of the deficiencies already mentioned and some more seriously. For example, the adjustment for V.R. uses V.R. group in only three categories as opposed to seven for the previous analyses, with half the children falling into the middle group. In this respect it is an even more inadequate adjustment than before. While, in many ways, these log-linear model analyses are the most interesting in the book, they actually do not carry as much useful information as the previous analyses might have done had they been better executed. There is also a technical deficiency in the presentation of the results of these log-linear analyses, in that the G^2 statistic cannot really be considered a "measure of importance" since it is a test statistic which depends on sample size, and its components cannot sensibly be interpreted separately (page 169).

There are other deficiencies involved with the execution and interpretation of the analysis involving

intake adjustment, but the above should indicate that much is left to be desired and they do not encourage the reader to place a great deal of confidence in the authors' results.

School ethos. Some of the most widely publicised findings of this study concern the so-called "school ethos" factors. The authors define "ethos" in terms of 39 "process" variables. These are chosen out of a large number of 46 on the basis of having statistically significant correlations with at least one outcome variable. It is not very surprising that using a composite score based on these 39 variables they found high correlations with exam score and behaviour. By choosing the significant and hence larger correlations only, they are capitalising on chance and by choosing a large number to go into a composite score they are virtually guaranteeing reasonably high correlations, especially with a sample size of only 12 schools (of which more below). It is not difficult to obtain similar results with purely random data (Preece, 1979).

There is no real attempt to provide a definition of school ethos which has an educational basis and the authors do not seem to appreciate the need to provide one. In what claims to be a major educational report this seems an important oversight.

Units of analysis and sample size. As explained earlier, some of the analyses in this book (such as that in Table 5.9) use individual children as units of the analyses whereas others use the schools. The trouble with the latter (more numerous) analyses is that the sample size is only 12. Apart from the fact that these schools are not really a random sample so that an inference to any other population of schools is problematical, the authors place far too great a reliance on the results of significance tests, by tending to dismiss non-significant relationships as unimportant. A striking example occurs on page 99 where Local Authority schools have an average ranking on attendance which is twice that of voluntary aided schools and where boys' schools have an average rank twice that of girls' schools for behaviour and exam results. Such differences are very large but, not surprisingly, they are non-significant since the sample size is only 12. At the bottom of the page the authors conclude that the sex differences are of negligible importance. Not only is that conclusion unjustified, but it results from a fairly elementary statistical misunderstanding and diverts the authors from further consideration of factors which have an obvious potential for explaining school differences.

Conclusions. The above comments have been almost entirely negative so that it will not be surprising if I conclude that the study results should be treated with caution if not scepticism. Nevertheless, the basic *idea* behind the study is a useful one and there is here some *prima facie* evidence that genuine school differences may exist which are related to measurable school factors. Undoubtedly, there should be further research along these lines which avoids the deficiencies of the present study. As for educational practice, it would seem wise to hesitate before applying any of the results of this study too literally and the quite strong conclusions drawn by the authors in Chapter 10 need to be viewed rather cautiously.

Reference

PREECE, P. F. W. (1979) Fifteen Taus and Rhos. *Br. Educ. Res. Ass. Newsletter,* August.

HARVEY GOLDSTEIN

School Influences on Pupil Progress: Research Strategies and Tactics

We welcome these attempts to examine critically our findings and our conclusions in order to determine just what has and what has not been established. Any piece of empirical research is necessarily limited in what it can accomplish so that there are the parallel dangers of dismissing studies because of the methodological imperfections present in all pieces of research (if the findings run counter to our prejudices) or of acting as if they provided the whole truth needed for policy decisions (if the results please us). Inevitably the reality is more complicated and it is necessary to assess both the extent to which methodological limitations may have distorted findings and the extent to which the findings are in keeping with or in opposition to other empirical research using different research strategies. In our book we attempted to follow both courses in deciding what conclusions could reasonably be drawn. Goldstein expresses scepticism and raises several different methodological objections which we regard as mistaken on the following grounds.

Firstly, he points to limitations in the procedures used to check whether the differences in pupil outcomes could be due, not to school influences, but rather to individual or family characteristics.

Of course, there are very real limitations in the strategies and tactics we followed (as we tried to point out in the book), but the argument on possible school influences does not rest on one statistical procedure but rather on the combination of (a) showing that school variation was not explicable in terms of the intake measures used and (b) showing that, to an important extent it was explicable in terms of our school measures.

Thus, we deliberately chose a longitudinal strategy (necessitating a nine year project) because we were aware of the crucial need to take into account the children's characteristics at the time they entered secondary school. None of the previous studies of schools had been able to provide any control for individual differences at the time of intake. Accordingly, the fact that we were able to control at all for intake variations was a major step forward. Of course, we were not able to have data on all potentially relevant intake characteristics. On the other hand, we did have data on the children's behaviour, nonverbal intelligence, reading, sex, socio-economic background and ethnicity. Previous studies have found these to be the most powerful predictive variables for the outcome with which we were concerned. It is most unlikely that the addition of, for example, family size (as suggested by Goldstein) would have substantially affected the findings. A comparison of various predictors of 16-year-olds' attainment in the NCDS cohort (Hutchison *et al.*, 1979) confirms this conclusion, showing that family size makes only a very limited additional contribution to prediction when test scores at 11 have already been taken into account. Nevertheless, we agree (as we clearly spelled out in the book) that the fact that school differences remained after controlling for intake differences does not mean that the differences were due to school influences. That inference requires several other steps, the most important of which is the determination of whether or not the school differences were systematically associated with the characteristics of the schools themselves.

Here, we are criticized by Goldstein on the grounds that the school process findings could just as easily have arisen by chance, and that similar results could be obtained with purely random data. In fact, that is not so. There are three essential complementary strands in the argument. First, the number of statistically significant associations between the school process variables and the outcome measures far exceed those expected on the basis of chance alone. Second, the statistically significant associations were not randomly distributed; rather, they followed a pattern. Physical and administrative features were largely unrelated to outcomes, whereas features concerning the social organization of school life did show consistent relationships with our measures of the children's progress. Thirdly, other studies, using rather different research designs, have produced closely comparable findings in both of these respects. (As well as the studies cited in the book there are other investigations to which reference might be made: see for example Goldman, 1961; Pablant and Baxter, 1975; Edmonds, 1979; Lezotte and Passalacqua, 1978; Brophy, 1979; Brookover *et al.*, 1979; Pederson *et al.*, 1978.) It is on that combination of steps that our argument on the probable importance of school process variables rests. However, as we make clear in the book, strong inferences on causation require the further step of the evaluation of planned change in schools (i.e. some form of "experiment" which constitutes the essential sequal to epidemiological studies). Such investigations have yet to be undertaken, but they are much needed.

Secondly, Goldstein criticizes our use of log-linear methods of contingency table analysis. Of course, all methods of multivariate analysis have their weaknesses, which vary from one technique to another. In order to remedy one limitation, all too frequently it is necessary to substitute another. Accordingly, the choice of method has to be made with proper attention to the particular characteristics of the sample and of the variables in each case. Not surprisingly, statisticians are not always in complete agreement on which method is most appropriate in any specific sample. Nevertheless, in our own study the main reason for using log-linear methods was that many of the independent and the dependent variables were of a categorical nature. Some were essentially qualitative (e.g. sex or delinquency). Others, such as exams or attendance, involved some implicit ordering but were not, perhaps, entirely satisfactory as ratio scales. The log-linear approach requires no assumptions about interval scales, linearity of correlation or homoscedasticity. The price paid is that there have to be rather arbitrary and fairly crude groupings. However, with the data to be analysed, this price is worth paying if the assumptions required in the more traditional correlational approaches cannot be met.

However, aware of the statistical differences of opinion on this issue we also used linear regression analysis which, like the log-linear modelling, showed a significant school process effect. We take the point about the relationship between the analyses in Tables 5.9 and 9.4 but an analysis along the lines

suggested by Goldstein does not alter the finding that school process is still significantly related to outcome.*

These issues are well known to researchers working in the field of study of possible school effects and it is regrettable that Goldstein dismisses our methods of analysis in disregard of both the issues and the fact that most statisticians working in this area now favour the approach we used.

Our use of G^2 statistics is also criticized as a means of comparing the relative importance of different variables—adding that these technical deficiencies "do not encourage the reader to place a great deal of confidence in the authors' results". His argument is based on the fact that the size of G^2 depends on sample size. Quite so, but of course in this instance we are using G^2 for comparative purposes only, within analyses in which the sample size remained constant.

Goldstein also comments on our use of significance tests, but within a single study in which the sample size remains constant, they do provide an indication of the relative importance of different variables at a given level of analysis. Thus, the fact that the school "process" variables gave rise to statistically significant associations with outcome, whereas variables such as size of school or sex composition did not, means that within this sample of schools, school process variables were likely to be more important with respect to the outcome than the other variables studied. It was for these purposes only that we used statistical probabilities. Obviously, non-significant findings do not mean that the variables were of no educational importance even within the sample of 12 schools, neither do they mean the variables could not be of even greater importance in other samples of schools. The point is spelled out in the book (see pages 105 and 161) but Goldstein chose to ignore it. The only secure way to determine whether a relationship is actually important is to find out whether the same pattern recurs in other investigations. If it does, then, and only then, can you be sure that it is meaningful, although of course it may still be the result of systematic bias. In that connection, our conclusions are strengthened by the closely similar findings from other empirical studies using rather different methods of research. While, undoubtedly, many questions remain to be tackled and further research in this important area is greatly needed, we still hold to our conclusions that "the results carry the strong implication that schools can do much to foster good behaviour and attainments".

References

BROOKOVER, W., BEADY, C., FLOOD, P., SCHWEITZER, J. and WISENBAKER, J. (1980) *Schools Can Make a Difference* (in press).

BROPHY, J . A. (1979) Advances in teacher effectiveness research. The Institute for Research on Teaching Occasional Paper, No. 18. Michigan State University.

EDMONDS, R. R. (1979) Some schools work and more can. *Social Policy* March/April.

GOLDMAN, N. (1961) A socio-psychological study of school vandalism. *Crime Delinquency* **7**, 221–230.

HUTCHISON, D., PROSSER, H. and WEDGE, P. (1979) The prediction of educational failure. *Educ. Stud.* **5**, 73–82.

LEZOTTE, L. W. and PASSALACQUA, J. (1978) Individual school buildings *do* account for differences in measured pupil performance. The Institute for Research on Teaching Occasional Paper, No. 6. Michigan State University.

PABLANT, P. and BAXTER, J. C. (1975) Environmental correlates of school vandalism. *J. Am. Inst. Planners* **241**, 270–279.

PEDERSON, E., FAUCHER, T. A. and EATON, W. E. (1978) A new perspective on the effects of first-grade teachers on children's subsequent adult status. *Harv. Educ. Rev.* **49**, 1–31.

MICHAEL RUTTER
BARBARA MAUGHAN
PETER MORTIMORE
JANET OUSTON
ALAN SMITH

*We should also point out that Table 9.4 contains an error in the reporting of the variance attributable to verbal reasoning. This was noticed too late to correct before publication, but has been corrected in later printings. It does not affect the variance attributable to school process.

Rejoinder

Professor Rutter and his colleagues have clarified a number of points in their reply to my critical review of *Fifteen Thousand Hours*. It seems worth emphasising that we all are agreed on the importance of longitudinal studies in this area and have no real difference of opinion on the need to exercise caution when making causal inferences from observational data. Also, as I said in my review, the basic idea of the study is a useful one and the study does present *prima facie* evidence for genuine school differences. Nevertheless, the reply from the authors contains some misunderstandings of the points I made, and I would like briefly to clear these up.

Firstly, there is the question of whether there were other intake factors which might have been able to account for school differences. Interestingly, Rutter *et al.* quote the paper by Hutchinson *et al.* (1979) to support their case for using V.R. group and social class as basic intake control variables. In fact, this paper shows that over and above test score at 11 years, teachers' ratings make a significant contribution and one which is larger than that of social class. Moreover, the contribution of family size and income taken together is about the same as that of social class, with additional but somewhat smaller contributions from crowding and housing amenities (Fig. 4). Thus Hutchinson *et al.* support rather than contradict my point.

Secondly, on the question of school ethos, the procedure used by the authors to derive the process score essentially defined this score on the basis of those variables which happened to turn out to have statistically significant correlations with the outcome variables. I was simply pointing out that this is a quite different procedure to that which seeks to provide a definition of school ethos based on substantive educational theory.

Thirdly, I did say in my review that the log-linear analyses were in many ways the most interesting ones in the book. The authors themselves point out that this type of analysis has limitations, but their response does nothing to dispel my doubts about the adequacy of using V.R. score in only three groups to adjust for intake differences. On the question of the G^2 statistic, my earlier comment was somewhat cryptic. I take the point about overall sample size, but nevertheless one of the difficulties with the use of this statistic is that its value depends on the way in which the total sample is distributed across the categories of occupation, V.R. score, etc. For a given set of actual differences between these categories in terms of the proportions of children with high examination scores, etc., the size of G^2 depends on this sample distribution, which in turn depends on how the original sample was selected.

Fourthly, my comments about statistical significance when using a sample of size 12 were not directed towards a study of the relative "importance" of different factors like school ethos and sex composition. I was merely pointing out that the observed differences for these factors are large and that non-significant results with such a small sample are extremely uninformative, since large real differences have only a small chance of attaining statistical significance. The authors partly defend themselves by referring to page 161 of their book, but in fact they claim there that "parental choice . . . was not a crucial determinant", and when this statement is set against the large observed differences in Table 8.10, it serves only to support my original point.

In short, the authors' reply to my criticism does little to alter my view that their conclusions are largely unsupported by their evidence. While I believe that more studies of school effectiveness would be very useful, such studies will need to pay attention to the shortcomings of *Fifteen Thousand Hours*.

HARVEY GOLDSTEIN