*J. R. Statist. Soc.* A (1979),
**142**, *Part* 4, *pp.* 407–442

# Some Models for Analysing Longitudinal Data on Educational Attainment

By HARVEY GOLDSTEIN

*Institute of Education, University of London*

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday, May 23rd, 1979, the President Sir CLAUS MOSER in the Chair]

SUMMARY

Longitudinal educational and social data on over 9000 children measured at the ages of 7, 11 and 16 years are analysed using linear models. The paper discusses and analyses the assumptions of these models, in particular the choice of scale units and transformations of educational attainment test data, the definition and interpretation of measures of "change" and methods for dealing with errors of measurement. Complex interrelationships across time between several variables are studied, and emphasis is given to the real life interpretation of the models used.

*Keywords*: LONGITUDINAL; ERRORS IN VARIABLES; PATH ANALYSIS; EDUCATIONAL ATTAINMENT; SIMULTANEOUS EQUATION ESTIMATION; COVARIANCE ANALYSIS

## 1. INTRODUCTION

IN an earlier paper, Fogelman and Goldstein (1976) used data from the British National Child Development Study (NCDS) to investigate factors associated with changes in the attainment of primary school children between 7 and 11 years. Their method of analysis was to regress 11 year attainment scores in reading and mathematics on corresponding 7 year attainment scores, and then to make comparisons between the equations fitted for different categories of various factors. Where these equations differed, this implied that the expected 11 year score for given 7 year score varied according to the category of the factors being studied.

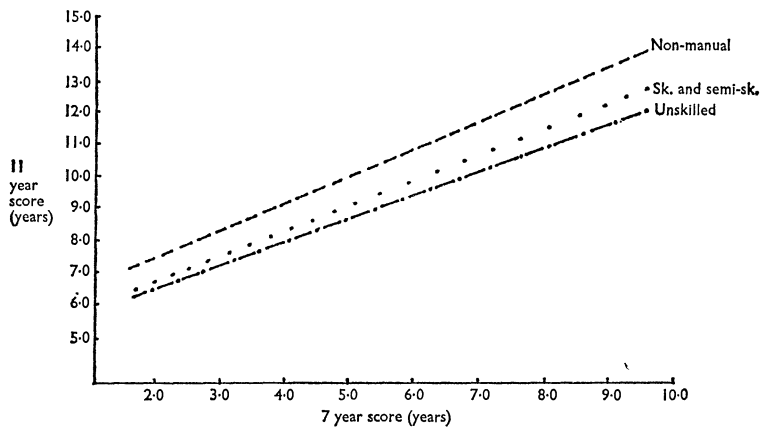


FIG. 1. Estimated regression of 11 year reading score on 7 year reading score, for three social class groups at 7 years.

Fig. 1, for example, shows the fitted regressions of 11 year on 7 year reading score (using a transformed score designed to produce linear relationships and measured in years of age) fitted to three social class groups. It is clear that over the range of 7 year scores found in

practice, for any given 7 year score, children with fathers in non-manual occupations at the age of 7 have, on average, higher 11 year scores than those with fathers in skilled or semi-skilled occupations who in turn have higher expected 11 year scores than children with fathers in unskilled occupations. The average differences, however, do not remain constant over the whole range of 7 year scores and this 'interaction' can be seen in the non-parallel regression lines. The set of regression models can be written, with the usual notation, as

$$x_{2ij} = \alpha_j + \beta_j x_{1ij} + \varepsilon_{ij}, \tag{1}$$

where $x_{2ij}$, $x_{1ij}$ are respectively the 11 and 7 year reading scores of the $i$th child in the $j$th social class. The estimated parameters are given in Table 1.

TABLE 1

*Estimated coefficients of regression of* 11 *year on* 7 *year reading score for three social class groups*

| | *Social class* | | |
|---|---|---|---|
| | *Non-manual* | *Skilled and semi-skilled manual* | *Unskilled manual* |
| $\hat{\alpha}$ | 5·68 | 5·28 | 4·90 |
| $\hat{\beta}$ | 0·86 | 0·77 | 0·75 |

Despite the apparent simplicity of this model a number of technical and interpretational problems arise which need careful analysis. In the following sections we examine each one in detail and further extend the model to incorporate an additional measurement occasion when the same children were measured at 16 years. We give first a brief description of the educational background to the measurements used and discuss the choice of scale units. We then discuss definitions and interpretations of measures of "change". Various models are introduced in increasing order of complexity, starting with simple relationships between the test scores at 7, 11 and 16 years, then introducing social class as an independent variable. The problems arising from the recognition of "measurement errors" in the test scores is then dealt with, followed by the introduction of social class changes between the three ages into the models. Finally, the fact of having more than one dependent variable at each age is dealt with explicitly. The assumptions of these models are discussed and tested using the NCDS data.

## 2. CHOICE OF MEASURING INSTRUMENTS

Measures of educational attainment in subjects such as reading and mathematics, made at widely separated ages, generally use different instruments. This occurs simply because different aspects of reading ability become manifest at different stages of development and this is reflected both in teaching and learning patterns. Thus in the NCDS at the age of 7 a word recognition test of reading was used, whereas at the ages of 11 and 16 a reading comprehension test was used. Even in those cases where the same test can be used at two occasions as at 11 and 16 years here, different portions of the test will be effective at each age, the easier items at the earlier age and the harder items at the later age. Thus the result is similar to using different measuring instruments. An immediate question is whether, in any useful sense, the instruments used can be said to measure "the same thing" at each occasion and whether it is meaningful therefore to compare scores across ages, for example by postulating underlying latent traits or factors which change in value over time, giving rise to longitudinal factor analytic type models (Jöreskog and Sörbom, 1976). While this problem is basically an educational rather than a statistical one,

it is nevertheless of some importance to establish that the statistical models have a useful interpretation. It appears that this problem has had little attention paid to it, and the following short discussion merely sketches out the nature of one solution.

With typical body measurements such as height and weight, the use of a constant measuring instrument applied in a standard fashion would normally be regarded as a sufficient condition to claim that the same characteristic is being measured at each occasion. Strictly speaking, however, this is a matter of definition rather than an empirical fact. Thus, for example, a child's height at puberty will be responding to a somewhat different set of underlying physiological mechanisms than his height as an adult, and in this sense we might choose to regard the same instrument as reflecting different underlying characteristics. Nevertheless, height has proved to be a useful measurement to make during the growth period, being sensitive to other important factors such as nutrition. The fact that it is convenient or useful to use the same instrument at each occasion, however, does not imply that this way of reflecting the effect of some other factor, such as nutrition, is universally superior to the use of different instruments on some occasions. With educational measurements we are in a logically similar situation to the measurement of height although, of course, there is less agreement over which instruments to use. We *could* devise a test instrument to cover a very wide age range (one method would be simply to combine together the separate test instruments applicable to each narrow age range), although as pointed out earlier, different parts of it would be effective at different ages. As is well recognized with educational measurements, it is their *appropriateness at particular ages* which is relevant. That is, they should reflect the attribute of interest as closely as possible. Thus a word recognition test was regarded as an appropriate instrument at 7 years, whereas it was appropriate to test reading comprehension at 11 and 16 years. With educational attainments the expectation of teachers and others as to what constitutes reading skills at any age will play an important part in determining what is to be measured, and this raises the further difficulty associated with varying expectations for different groups of children. Since this problem takes us into controversial areas of educational theory it is not appropriate to pursue it here. Instead we assume that the available measurements have been chosen to allow useful interpretations to be made for each test at the appropriate age. The interpretation of any function of such measurements will then be a combination of these separate interpretations.

## 3. Definitions of Change

Two broad approaches can be found to the definition and measurement of change. The first and simplest approach is to define a common scale and use the simple difference between the measurement at two occasions as a measure of change. Such an approach is commonly used with physical body measurements like height, and it can be extended by relating the measurement to a function of time or age as is done in growth curve fitting. This approach is also sometimes used with mental measures, the measurements at each occasion being standardized to have the same population distribution. For example, IQ is commonly reported in such standardized units. In this case the arithmetic difference between two standardized measurements measures relative change. While there may be circumstances where such a use is justified, an example using height will illustrate the general difficulty of choosing appropriate scale units. According to Tanner *et al.* (1966) the mean height difference between boys and girls at the age of 4·0 years is 1·2 cm. At the age of 8·0 years the difference is the same at 1·2 cm. The population standard deviation, however, increases by a third from 4·34 to 5·77 cm. If, therefore, we were to standardize height to give the same population standard deviation at each age we would find that the mean difference at age 8·0 had decreased to 0·21 standard deviations from 0·28 standard deviations at age 4·0. Thus the interpretation of any changes in group differences may depend crucially on the scale chosen and it is by no means clear always which is the appropriate scale to use. Furthermore, various assumptions about the form of the distribution of the measurements can also lead to differing conclusions.

The second approach to the measurement of change avoids the more troublesome of the above problems. Instead of trying to define a simple measure of change between two occasions we can turn our attention to the relationship between the measurements on the two occasions. (This can only be done of course where longitudinal data are available.) We can now allow the data themselves to determine the precise form of the relationship and we can compare the relationships for different groups, etc. This is the approach adopted by Fogelman and Goldstein (1976).

Let us denote the first occasion measurement by $x_1$ and the second by $x_2$. There are three broad approaches to the study of the relationship between $x_1$ and $x_2$. We could regard the two measurements symmetrically and consider, for example, a functional relationship between them (Kendall and Stuart, 1967, Chapter 29). Alternatively we could study the regression of $x_2$ on $x_1$ or $x_1$ on $x_2$. There is one obvious choice, namely the regression of $x_2$ on $x_1$. This choice incorporates the important information about the direction of time, namely that occasion 1 precedes occasion 2 and hence that our interest lies in the distribution of the measurement at occasion 2 given knowledge of the value of the measurement at occasion 1. Certainly if we are interested in using our model to contribute evidence relevant to an attempt at understanding causal mechanisms which operate through time, then this seems the obvious approach to use. While there may be circumstances when other kinds of interpretations are required and one of the other models may be appropriate, they do not appear to be so in the present case and we shall not consider them further.

In the simplest case the regression will be linear and we write, using the same notation as before,

$$x_{2i} = \alpha + \beta x_{1i} + \varepsilon_i.$$

It is worth noting that by a suitable change of scale of $x_{1i}$ setting $x_{1i} = \beta x_{1i}$, we obtain

$$x_{2i} - x_{1i} = \alpha + \varepsilon_i.$$

The $x_{1i}$ here are not treated as random variables so that despite its apparent similarity, this is not equivalent to the use of simple differences, where both variables are treated as random. We also note that any inferences based on the regression model are essentially unaltered by a linear transformation of either variable, a desirable property which is not possessed when simple differences are used to measure change. Non-linear transformations, however, do create problems and we shall deal with these in later sections.

In the following analyses we shall make use of and extend the regression model to more than two occasions, suggesting methods for dealing with non-linearity, measurement error and identification problems with simultaneous equations.

## 4. DATA SOURCES

The present data come from the National Child Development Study which followed up a cohort of 17 000 children born in one week of March 1958, at the ages of 7, 11 and 16. The children belonged to the first year-group for whom the minimum school leaving age was 16 years. A description of the social and educational data (among others) collected at these ages is given in Fogelman (1977), who also discusses the representativeness of the sample in terms of response rate, showing an overall response of 91, 91 and 87 per cent at the three ages, with certain identifiable biases resulting in some under-representation of children from "disadvantaged" groups, but with little effect on test score distributions. At the age of 7, a word-recognition test and an arithmetic test were given. At 11, a reading comprehension and a general mathematics test were given, and at 16 the same reading test as at 11 and a different general mathematics test were given. A full description of the reading and mathematics tests used at these ages is given in Fogelman *et al.* (1978). Apart from these test scores, the analyses in the present paper use social class group which is defined in terms of the Registrar General's (1960) classification.

## 5. MODELS

The first and simplest model to be analysed is represented in Fig. 2 as a path diagram.

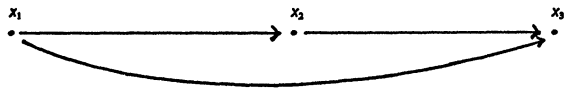


FIG. 2. Path diagram showing the direction of relationships between measurements at three occasions.

We write this as a pair of regression functions

$$x_2 = f(x_1), \quad x_3 = g(x_1, x_2), \tag{2}$$

where $x_1, x_2, x_3$ are the test scores in reading or mathematics at 7, 11 and 16 years.

In the interest of parsimony we would prefer (2) to be linear and additive and so we first consider the detailed relationships among the test scores, and where necessary introduce transformations to satisfy these requirements. We would also like to have the residual terms obeying the usual assumptions of normality, homoscedasticity and independence so that standard regression theory can be applied.

Fig. 3 shows plots of mean 11 year reading score on 7 year reading score and mean 16 year score on 7 and 11 year reading scores. Fig. 4 shows similar plots for mathematics. While it would be possible to make empirical transformations for each plot separately to approximate

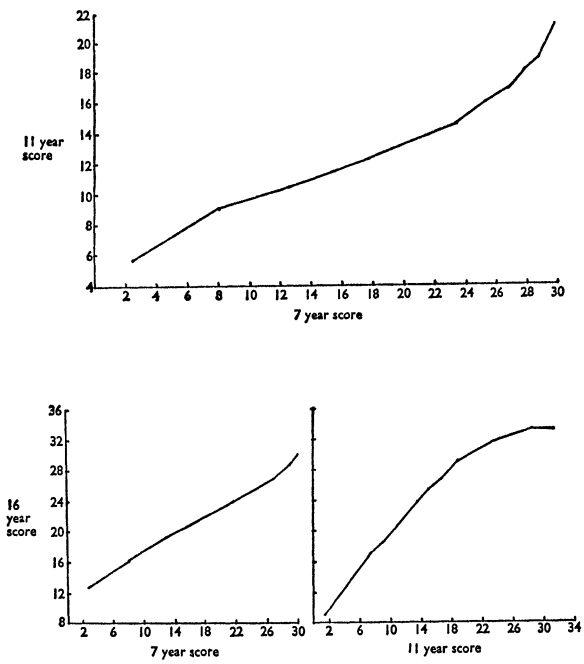


FIG. 3. Mean reading scores at 11 and 16 years for reading scores at 7 and 11 years.

the linearity, normally or homoscedasticity assumptions, we would ideally like to satisfy them simultaneously. Fogelman and Goldstein (1976) carried out an empirical transformation of the 7 year reading score to produce linearity of the 11 on 7 year plot. A modification of this transformation together with empirical transformations to give standard $N(0, 1)$ distributions for the 11 and 16 year reading scores give near-linear relationships and also tend to stabilize the error variances, the ratio of maximum to minimum variance never exceeding 2·0. The

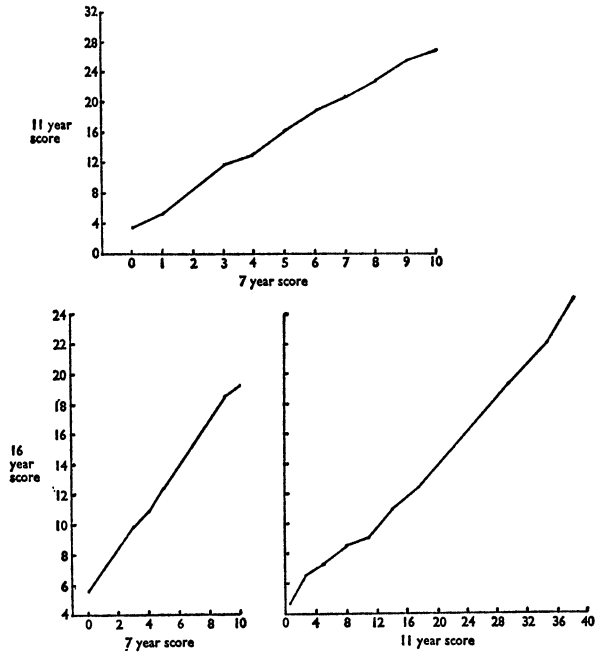


FIG. 4. Mean mathematics scores at 11 and 16 years for mathematics scores at 7 and 11 years.

modification to the 7 year reading score transformation also reduces the size of the social class by 7 year score interaction mentioned in Section 1, a point to which we return in Section 9.

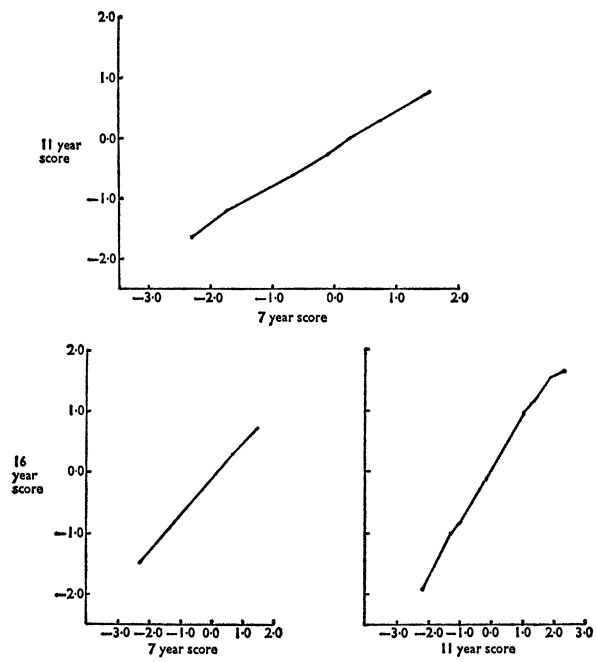


FIG. 5. Mean reading scores at 11 and 16 years for reading scores at 7 and 11 years, after transformation.

Fig. 5 shows the transformed reading score plots which are now more nearly linear. As shown by Fig. 4, the mathematics relationships are approximately linear and remain so after empirical transformations of the 11 and 16 year scores to give standard $N(0, 1)$ distributions with the ratio of maximum to minimum variance likewise never exceeding 2·0.

We therefore rewrite equations (2) as linear regression equations in the usual way as follows (omitting the individual subscript)

$$x_2 = \alpha_1 + \beta_1 x_1 + \varepsilon_1, \tag{3a}$$

$$x_3 = \alpha_2 + \beta_2 x_1 + \gamma_2 x_2 + \varepsilon_2. \tag{3b}$$

Thus we have a pair of linear equations. In order that ordinary least squares applied to each equation gives efficient and consistent estimates, we need to satisfy the usual conditions that the correlations between the random error terms $\varepsilon_1$, $\varepsilon_2$ and also between these and the independent variables are zero. Failure to satisfy the former condition will lead to inefficient estimates but failure to satisfy the latter will lead to inconsistent estimates. Both these violations of the usual assumptions can arise from misspecification of the model by omitting a relevant variable. In the present case we have transformed the variables where necessary to give linear relationships, and standardized residuals from fitting the above two equations show no dependencies when plotted against 11 and 16 year fitted values or against each other.

Table 2 shows the fitted coefficients for equations (3a) and (3b) together with standard errors. From Table 2B we see that there are small multiplicative interaction terms (i.e. $x_1 x_2$)

TABLE 2

*Coefficient estimates for equations (3a) and (3b)*

A. *Dependent variable* 11 *year score*

| | *Reading* | | *Mathematics* | |
|---|---|---|---|---|
| | *Fitted value* | *S.E.* | *Fitted value* | *S.E.* |
| $\hat{\alpha}_1$ | 0·015 | | 0·065 | |
| $\hat{\beta}_1$ | 0·664 | 0·009 | 0·561 | 0·008 |
| Residual variance | 0·61 | | 0·64 | |

Sample size = 9200

B. *Dependent variable* 16 *year score*

(Figures in brackets are for model including multiplicative interaction)

| | *Reading* | | *Mathematics* | |
|---|---|---|---|---|
| | *Fitted value* | *S.E.* | *Fitted value* | *S.E.* |
| $\hat{\alpha}_2$ | 0·159 (0·184) | | −0·004 (−0·043) | |
| $\hat{\beta}_2$ | 0·136 (0·131) | 0·010 | 0·076 (0·072) | 0·009 |
| $\hat{\gamma}_2$ | 0·734 (0·735) | 0·009 | 0·758 (0·757) | 0·009 |
| Interaction | (−0·037) | 0·007 | (0·075) | 0·007 |
| Residual variance | 0·46 (0·45) | | 0·49 (0·48) | |

Sample size = 9100

In subsequent tables the number of children in the analysis (to the nearest 100) varies, since where data were not available on any variable then the complete case was omitted. The test scores of the omitted cases have been compared with those included and in no instance do significant differences emerge.

for mathematics and reading. Since they do not markedly affect the predicted values and their interpretation is unclear, we shall retain the simpler additive model.

## 6. MEASUREMENT ERROR

Nearly all measurements made on human beings contain a component of error which, if large enough, can substantially modify our inferences, as we shall show. By "measurement error" we mean that if an individual is independently and repeatedly measured in a short space of time, different observed values would be obtained. In practice, we normally cannot do this without the relation between the individual and the measuring instrument altering and inducing dependencies between measurements. This is particularly true for mental measurements where learning effects take place. Furthermore, the phrase "in a short space of time" is vague and difficult to generalize. For example, is a single day an appropriate unit of time, or do we expect a trend in the "true" value even within a day? If such a trend occurred and we chose a day as our unit, then the "error" will include changes in "true" value as well as other random sources. Where underlying trends can be estimated it may be possible to obtain estimates of the "instantaneous" true error, but for the variables in this paper the trend over a short period of time is so small compared to the size of the likely measurement error, that there is little to be gained in such an attempt. For example, the change in 11 year reading score is about 0·01 units per week, whereas the measurement error has a standard deviation of about 0·3 units.

A variety of methods is available for the estimation of measurement errors. Some of these use the information given by the individual items of a single test and others use "parallel" test forms. Each makes particular assumptions and has drawbacks. A detailed discussion is given by Lord and Novick (1968), although this does not deal explicitly with the above "time span" problem. In the present case we use estimates based on "split half" analysis of the NCDS data since suitable external estimates are unavailable. Such "split half" estimates tend to underestimate the true reliability based on notionally immediately repeated testing. On the other hand, they do not take into account other sources of variation such as the conditions of testing, etc. which would themselves tend to reduce the reliability. Since we have no information concerning the magnitude of these opposing tendencies we are unable to pursue this point further. It has also been suggested that since any single test is only one possible selection of items from a very large number, the variation between all possible tests should be taken into account. Although this may have application in some contexts it seems more reasonable in the present case to regard any departure of the test mean from a hypothetical population mean as a systematic factor, the same for each individual. Thus the conditional analysis would be unaffected. A further problem arises, however, because the data were collected during several months at each occasion. The estimate of measurement error thus will include a component of the time trend which is of the same order of magnitude as the standard deviation of measurement error over about a 9 month period. Of the apparent measurement error, about 5 per cent for reading and 15 per cent for mathematics is due to this trend, and the sample based estimates have been adjusted on the basis of these values.

The simplest model for measurement error is as follows,

$$X_i = x_i + u_i, \tag{4}$$

where $X_i$ is the observed measurement on the $i$th individual, $x_i$ is the "true" value and $u_i$ the measurement error. We assume that the $u_i$ are identically and independently distributed with zero mean and also independent of the $x_i$. The $u_i$ are also sometimes assumed to have a normal distribution and general procedures for solving the likelihood equations when (4) is added to (3) are discussed by, for example, Jöreskog and Sörbom (1976). The method adopted in the present paper uses a least squares technique which assumes no particular distributional form.

If we imagine an individual to be independently remeasured within the basic time unit, the correlation between the successive measurements is, from (4),

$$R = \text{var}(x_i)/\text{var}(X_i)$$

where $\text{var}(X_i) = \text{var}(x_i) + \text{var}(u_i)$.

$R$ is known as the reliability of $X_i$. Hence if we know the value of $R$ and $\text{var}(X_i)$, or have estimates of them, we can obtain $\text{var}(x_i)$ and $\text{var}(u_i)$.

It follows that the usual estimate of the correlation between a variable containing measurement error (an "unreliable" measurement) and a variable with no measurement error is a biased and inconsistent estimator of the correlation between the true values. The same is true for a simple regression coefficient where the measurement error is in the independent variable. Since it is the variance term which is inflated here, the true regression is underestimated in absolute value. If a consistent reliability estimate is available, then a correction to the observed regression coefficient can be made by dividing it by the reliability to yield a consistent estimate of the coefficient of the true score. This procedure can be extended to several independent variables where corrections are made to the covariance matrix of independent variables. Where the measurement errors are independent of each other this involves subtracting measurement error variance estimates from the observed variances. We assume this to be the case in the present paper. The coefficient estimates obtained in this way are referred to as "moment" estimates. Before going on to study the properties of such estimates one theoretical difficulty needs to be mentioned.

It was shown by Lindley (1947) that if an exact linear regression relationship exists between the true values of two variables, then there continues to be an exact linear regression relationship between the observed values if and only if the cumulant generating function of the errors is proportional to that of the independent variable. Where both distributions are normal then this condition is satisfied and an analogous condition holds for the case of several independent variables. We do not know, in the present data, whether the conditions to maintain linearity hold, but since, in practice, the linear relations are, at best, useful approximations or simply convenient graduations of the data, it seems reasonable to continue to assume linear relationships. Cochran (1975) suggests that where Lindley's condition does not hold, the linear component of the relationship still dominates for moderately high values of the reliability.

Before we go on to describe the further analysis of the data, it is instructive to examine how inferences may be changed drastically by correcting for measurement error. Fig. 6 is an

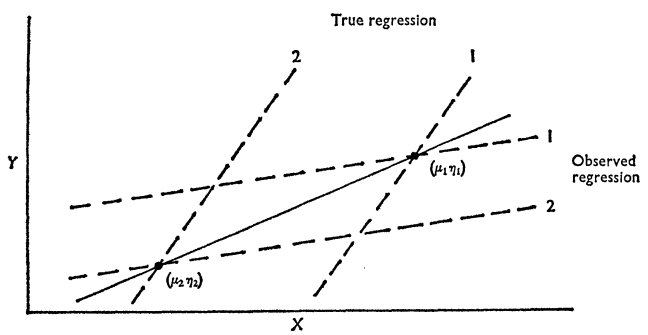


FIG. 6. Hypothetical illustration of an effect of correcting for measurement error.

hypothetical example in which two groups 1 and 2 are compared by fitting linear regressions of $Y$ on $X$ where $Y$ is the second and $X$ the first occasion measurement (sampling variation is ignored).

The lines labelled "observed regression" are the (assumed) parallel lines for the observed scores, passing through the means of the observed score distributions. Since the measurement errors have zero means, the observed score distributions have the same means as the true score distributions. The lines labelled "true regression" are the observed lines whose slopes have been divided by the reliability. We see that the relative position of the lines has changed, that for population 2 now lying above that for population 1. Hence, an inference based on true scores is that for given $X$ values the average $Y$ value is greater in population 2 than in population 1, whereas based on observed scores the opposite inference would be made.

Although we have tacitly assumed up to now that measurement errors are "nuisance" variables to be eliminated from our inferences, there are circumstances when it is appropriate to base estimates on observed scores. If we are interested solely in the prediction of a test score from another test score, and if the second test score can only ever be observed with error, then the appropriate prediction equation is that based on these observed scores. This may be of particular relevance when studying the workings of an educational system where the allocation of children to different curricula or types of school is made on the basis of unreliable measurements. In such cases we may equally well be interested in differences between, for example, types of school for given observed scores, as well as in differences for given "true" scores. We shall return to this point when studying some of the results and take it up further in the discussion.

## 7. MOMENT ESTIMATORS

Although the proposed method of moment estimation yields consistent estimates, for a finite sample we may not be able to rely on this asymptotic property. While investigation of bias in the multiple independent variable case is difficult, results are available for simple linear regression which will at least indicate the order of magnitude of expected biases. It has been shown by Richardson and Wu (1970) that for simple linear regression the expected value of the estimated observed slope is

$$E(b) = R\beta\left\{1+\frac{2R(1-R)}{n}+O\left(\frac{1}{n^2}\right)\right\},$$

where $R$ is the reliability and $n$ is the sample size. Thus with sample sizes and values of $R$ in the present data, the expected relative bias of $b/R$ is no larger than about 1 in 10 000.

The properties of the moment estimators have been studied by Warren *et al.* (1974) and Fuller and Hidiroglou (1978) and the relevant results from the former paper will now be given. The latter paper presents analogous results when reliability estimates, rather than measurement error variance estimates, are available from independent data. The use of independent estimates of reliability seems to be of less practical utility for educational tests, however, than measurement error variance estimates. A common difficulty is that reliability estimates often will not be estimated on a sample from the same population to which they are subsequently applied. Although there seems to be little evidence, it may be reasonable to assume that measurement error variances do not change markedly from one population or group to another, whereas the observed between-individual variances often will. In such circumstances, where an external estimate is available, the measurement error variance rather than the reliability should be used. The model is

$$Y = x\beta + e, \quad X = x + u, \tag{5}$$

where

$Y$ is the $(n \times 1)$ vector of observed dependent variables;

$X$ is the $(n \times k)$ design matrix of observed independent variables;

$\beta$ is the $(k \times 1)$ vector of true unknown parameters;

$e$ is a $(n \times 1)$ residual vector;

$u$ is the $(n \times k)$ matrix of measurement errors;

$x$ is the $(n \times k)$ design matrix of true independent variables.

We assume that for the $i$th individual ($i = 1, \ldots, n$) the covariance matrix

$$\operatorname{cov}(e_i, u_{1i}, \ldots, u_{ki}) = \operatorname{diag}(\sigma_e^2, \sigma_{u_1}^2, \ldots, \sigma_{u_k}^2),$$

where $\sigma_{u_j}^2$ is the measurement error variance for the $j$th independent variable. Thus the measurement errors are uncorrelated with each other and with the residual error. We also assume that they are uncorrelated with $x$, and that the estimate $\hat{\sigma}_{u_j}^2$ of $\sigma_u^2$ is uncorrelated with $x$ and $e$.

In the present case the $\sigma_{u_j}^2$ associated with the independent variable social class will be zero, since we assume no error attaches to this measurement. Define

$$\hat{H} = n^{-1} X^{\mathrm{T}} X - D \Lambda D,$$

where

$$D = \operatorname{diag}(S_{X_1}, \ldots, S_{X_k}),$$

$$S_{X_i}^2 = n^{-1} \sum_{i=1}^{n} X_{ij}^2$$

and $X_{ij}$ is the value of the $j$th independent variable on the $i$th individual,

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_k),$$

where

$$\lambda_j = \sigma_{u_j}^2 / \sigma_{X_j}^2 = 1 - R_j.$$

Thus $\hat{H}$ is obtained by subtracting the measurement error variances from the appropriate diagonal terms of the observed covariance matrix. This "corrected" covariance matrix is then used in the subsequent calculations. A complication which may arise is when $\hat{H}$ ceases to be positive definite because of sampling fluctuations. In this case it can be adjusted by adding a suitable small multiple of the matrix $D \Lambda D$ (Warren *et al.* 1974). This problem does not occur in the present data. A consistent estimator of $\beta$ is given by

$$\hat{\beta} = n^{-1} \hat{H}^{-1} X^{\mathrm{T}} Y$$

with a consistent estimator of the covariance matrix

$$\operatorname{cov}(\hat{\beta}) = n^{-1} \hat{H}^{-1} S_v^2 + n^{-1} \hat{H}^{-1} (D \Lambda D S_v^2 + D \Lambda D \hat{\beta} \hat{\beta}^{\mathrm{T}} D \Lambda D + 2 \hat{R}) \hat{H}^{-1},$$

where $S_v^2$ is the residual variance $(n-k)^{-1} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}} \hat{\beta})^2$, and $R = n d^{\mathrm{T}} u$, where

$$u^{\mathrm{T}} = (\hat{\sigma}_{u_1}^4, \ldots, \hat{\sigma}_{u_k}^4),$$

$$d^{\mathrm{T}} = (d_1^{-1} \hat{\beta}_1^2, \ldots, d_k^{-1} \hat{\beta}_k^2)$$

and $d_j$ is the degrees of freedom used in estimating $\sigma_{u_j}^2$.

We can use the estimate $\operatorname{cov}(\hat{\beta})$ to construct significance tests and confidence intervals for $\beta$ in the usual way. In the present data, the 16 and 11 year measurement error estimates are based on a sample size of about 300. The estimates for the 7 year tests are based on an instrumental variable approach described in the next section. Using the estimated variances of the regression coefficients given by this method, we can estimate the equivalent sample size, namely that which would produce the same variances by use of a split-half method, by substituting in the above formulae. This gives approximate equivalent sample sizes of 3000, although some caution is needed since these estimates do not completely satisfy all the independence assumptions given above.

Where the $\lambda_i$ are small and $d_i$ large compared to $n$, this estimate of the covariance matrix should be close to the usual estimate derived from the observed matrix $X$.

As we pointed out earlier, other methods of estimation, based on maximum likelihood, have been proposed for the estimation of linear models with measurement error. Apart from having to accept assumptions of normality, one of the more unsatisfactory features of these is that they require the parameters to have either completely unknown values, or to be subject to known linear constraints, and do not allow for the possibility of imprecise knowledge such as is available for the present measurement error variance estimates. As we shall see below, the precision of the estimates of the measurement-error variances makes an important contribution to the variance of the other parameter estimates. Thus if we assume that the values of these measurement error variances are known without error we may gain a quite false impression of precision. Of course, if a "split half" procedure were carried out on the total sample, the halves could be treated as "parallel" tests and the method of Jöreskog and Sörbom (1976) will then make the due allowance for the sampling variability of the measurement error variances.

We now apply the above results to the data previously analysed according to the models given by (3a) and (3b). The estimated reliability values for the test scores are given in Table 3.

TABLE 3

*Estimated reliability coefficients for test scores*

| *Age* | *Reading* | *Mathematics* |
|---|---|---|
| 7 years | 0·79 | 0·65 |
| 11 years | 0·82 | 0·94 |
| 16 years | 0·86 | 0·85 |

The 7 year reading test of word recognition was originally used by Southgate (1962) and has a quoted reliability of 0·94. However, as a result of the non-linear transformation used, we would expect this value to be reduced, and we have used instead an "internal" estimate, the rationale for which is described in the next section. The 7 year mathematics test consisted of problem arithmetic items. No independent reliability estimate is available, and so again an "internal" estimate is used. At 11 years, the reliability estimates were obtained from an item analysis of test scores which had been carried out on a sub-sample of the NCDS subjects using the split half technique mentioned earlier. Thus there is a reasonable amount of uncertainty associated with each of these estimates, and in the following analysis we see how this inflates the estimates of the standard errors. We first show how the parameter estimates vary with the reliabilities.

Fig. 7 shows how the estimate of the coefficient of the 7 year score in equation (3a) changes with the reliability and Figs 8a and 8b show how the coefficients in equation (3b) alter with changing reliabilities. For reading it can be seen that even small changes in the reliability values given in Table 3 can produce for equation (3b) relatively large changes in $\hat{\beta}_2$ and $\hat{\gamma}_2$. For mathematics, small changes in the given reliabilities produce relatively smaller changes in the parameter estimates—due in the main to the high value of the 11 year reliability. As would be expected, a high 7 year reliability does not imply the same stability. It is also worth noting that for high 11 year reliabilities the parameter values decrease with decreasing 7 year reliability value, whereas for low 11 year reliability they increase with decreasing 7 year reliability.

Table 4 gives adjusted estimates of the coefficients together with their estimated standard errors. Comparing these estimates with those in Table 2 it will be seen that the coefficients in equation (3a) are adjusted upwards. The standard error for reading is increased by about 30 per cent, about 70 per cent of which is due to the error in estimating $\sigma_u^2$, as a result of the

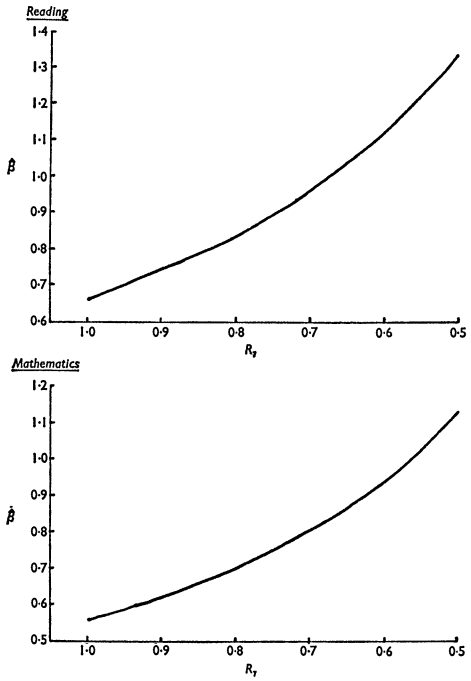


FIG. 7. Regression coefficient estimates adjusted for different reliability values of 7 year score (equation (3a))

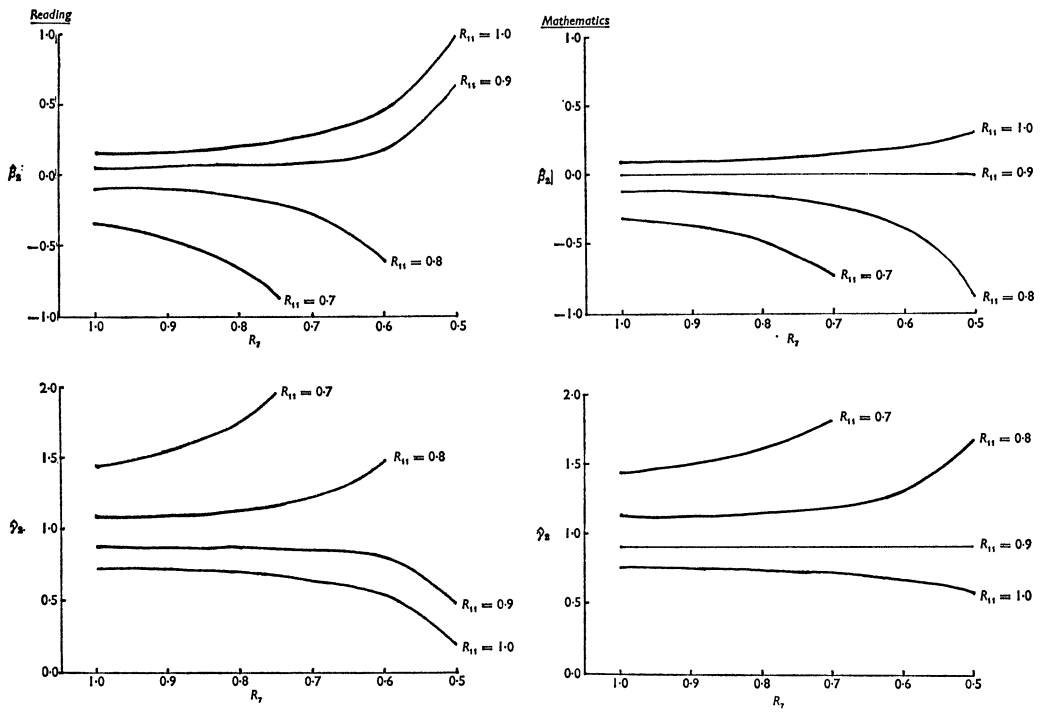


FIG. 8. Regression coefficient estimates adjusted for different reliability values, for 7 year ($R_7$) and 11 year ($R_{11}$) scores (equation (3b)). (a) Reading. (b) Mathematics.

TABLE 4

*Coefficient estimates in equations (3a) and (3b) after adjusting for measurement error*

A. *Dependent variable* 11 *year score*

| | *Reading* | | *Mathematics* | |
|---|---|---|---|---|
| | *Fitted value* | *S.E.* | *Fitted value* | *S.E.* |
| $\hat{\alpha}_1$ | 0·015 | | 0·065 | |
| $\hat{\beta}_1$ | 0·841 | 0·012 | 0·864 | 0·017 |
| Residual variance | 0·49 | | 0·47 | |

B. *Dependent variable* 16 *year score*

| | *Reading* | | *Mathematics* | |
|---|---|---|---|---|
| | *Fitted value* | *S.E.* | *Fitted value* | *S.E.* |
| $\hat{\alpha}_2$ | 0·154 | | −0·005 | |
| $\hat{\beta}_2$ | −0·147 | 0·051 | 0·074 | 0·021 |
| $\hat{\gamma}_2$ | 1·113 | 0·059 | 0·808 | 0·017 |
| Residual variance | 0·30 | | 0·45 | |

Sample size = 9200

relatively small number of associated degrees of freedom. For mathematics, due to the low reliability, the standard error is increased by about 100 per cent, of which about 60 per cent is due to the error in estimating $\sigma_u^2$.

For equation (3b) for reading there is a large decrease in the coefficient of the 7 year score and a large increase in that of the 11 year score. This is also clear from Fig. 8a as pointed out above. For mathematics, because of the high reliability of the 11 year score, the coefficients are little changed. The standard errors for reading increase substantially however, being 5 to 6 times as large as in Table 2. Almost 90 per cent of this is due to the relatively large sampling error of the measurement error variance estimates, especially the 11-year-old one. For mathematics, due to the high reliability of the 11 year score, the increase is more modest the standard errors being at most about $2\frac{1}{2}$ times those in Table 2 and only about 20 per cent being due to the sampling variability of the measurement error variance estimates. These results underline the importance of obtaining measurement error variance or reliability estimates which are of the same order of accuracy as the observed variances, especially when the reliabilities tend to be low. With large samples such as the present one, the sample size will be much larger than that for the quoted measurement error or reliability estimates for the tests used. In these circumstances it may be worth the effort to obtain new estimates from the total sample. It is also of interest to note that the sign of $\hat{\beta}_2$ changes for reading, giving a negative coefficient after adjustment.

## 8. INSTRUMENTAL VARIABLE ESTIMATORS

If we incorporate (4) in (3a) and (3b) we can write the latter in terms of true values as

$$x_2 = \alpha_1 + \beta_1 x_1 + \varepsilon_1, \tag{6a}$$

$$x_3 = \alpha_2 + \beta_2 x_1 + \gamma_2 x_2 + \varepsilon_2 \tag{6b}$$

with observed values

$$X_i = x_i + u_i. \tag{6c}$$

Despite an apparent similarity to the model for a functional relationship (Anderson, 1976), this model is concerned with the relationship between the expected values of the dependent variable for the values of the independent variables rather than the symmetrical relationship implied by the functional equation model.

Rewriting the above equation in terms of observed variables we have

$$X_2 = \alpha_1 + \beta_1 X_1 + (\varepsilon_1 + u_2 - \beta_1 u_1), \tag{7a}$$

$$X_3 = \alpha_2 + \beta_2 X_1 + \gamma_2 X_2 + (\varepsilon_2 + u_1 - \beta_2 u_1 - \gamma_2 u_2). \tag{7b}$$

Thus $X_1$ and $X_2$ are correlated with the error terms which involve $u_1$ and $u_2$, which implies that ordinary least squares gives inconsistent parameter estimates. We also note that errors in the dependent variables do not give rise to inconsistent estimates, merely increasing the residual variance. In the previous section we incorporated information about reliability estimates into these equations to provide consistent estimators. Another approach is to use an independent estimate of $X_1$ in equation (7a) and of $X_1$ and $X_2$ in (7b) which are uncorrelated with the error terms. Write

$$X_1 = \sum_j \delta_{1j} z_j + v_1, \quad X_2 = \sum_j \delta_{2j} z_j + v_2 \tag{8}$$

Then if the $z_j$, known as instrumental variables, the $v_i$, $u_i$ and $\varepsilon_i$ are mutually uncorrelated, we can obtain consistent estimates of the parameters. If normality is assumed we can use maximum likelihood, otherwise we may use two-stage least squares, both of which give similar estimates for the present data. The efficiency of this procedure depends on the predictive efficiency of (8). For a single independent variable the relative efficiency of two stage least squares is equal to the square of the correlation between the independent variable and the set of instrumental variables (Johnston, 1972).

The procedure for two-stage least squares is to estimate the unknown parameters in equation (8) and then use the predicted values of $X_1$ and $X_2$ from (8) in (6a) and (6b) instead of the observed values. The consistency property essentially depends on the $v_i$ errors being uncorrelated with the errors of measurement $u_i$. When the instrumental variables are measured at the same time as the test score this assumption may be open to question. On the other hand, if we use instrumental variables measured at a quite different occasion we will not obtain as much precision.

A five point scale of teacher's ratings of number work is available at 7 years and we use the mean 7 year mathematics score at each scale point to estimate $\beta_1$ in (6a). This is the value given in Table 4 and its efficiency is 0·37. It corresponds to a reliability of 0·65 after adjusting for age trends, and it is this value which is given in Table 3. The reliability estimate for reading was obtained similarly using teachers ratings at the age of 7. Similar ratings are available at 11 and 16 years. We have not, however, used these latter in Table 3 but rather the item analysis values mentioned above. Although the former have smaller standard errors, they may not be providing unbiased estimates. In fact for the 11 year scores the instrumental variable estimates are quite close to those in Table 3, being 0·80 and 0·95 respectively for reading and mathematics, but considerably smaller at 16 years being 0·72 and 0·77. For the 7 year scores we have to use instrumental estimates since no others are available, but it should be remembered that these may not be very good.

## 9. Attainment Changes and Social Class

We have referred already to the finding of Fogelman and Goldstein (1976), that the expected 11 year reading and mathematics scores for given 7 year scores differed between social classes. We now extend this result, first to include 16-year-old attainments, and then to study the effect of changes in social class.

Equations (3a) and (3b) become

$$x_2 = \alpha_1 + \beta_1 x_1 + w_{1j} + \varepsilon_1, \tag{9a}$$

$$x_3 = \alpha_2 + \beta_2 x_1 + \gamma_2 x_2 + w_{2j} + \varepsilon_2, \tag{9b}$$

where the $w_{1j}, w_{2j}$ are additive constants for three social classes at 7 and 11 years. The social class groupings, based on the occupation of the child's father, are as follows (Registrar General, 1960):

(a) non-manual workers (Social classes I, II, III Non-manual);

(b) skilled and semi-skilled manual workers (Social classes III, IV);

(c) unskilled manual workers (Social class V).

As has already been mentioned the interaction between social class and 7 year reading score is reduced by the modified 7 year reading score transformation and Table 5 shows the separate fitted regression lines for each social class, not corrected for reliability. The slopes are in fact very similar and not significantly different.

TABLE 5

*Separate regression lines for each social class. Eleven year on 7 year reading score*

| *Social class* | $\hat{\alpha}$ | $\hat{\beta}$ |
|---|---|---|
| Non-manual | 0·278 | 0·641 |
| Skilled and semi-skilled manual | −0·074 | 0·599 |
| Unskilled manual | −0·266 | 0·592 |

Sample size = 8500

Test for equality of slopes $\chi^2$ (2 d.f.) = 2.6

N.B. For significance tests

| | |
|---|---|
| *** | $P < 0{\cdot}001$ |
| ** | $0{\cdot}001 < P < 0{\cdot}01$ |
| * | $0{\cdot}01 < P < 0{\cdot}05$ |
| Otherwise | $0{\cdot}05 < P$ |

The fact that the monotone scale transformation can change the relative slope differences underlines the arbitrary nature of these attainment test scales. In the present case, the estimated lines have become more nearly parallel as a result of "stretching" the higher 7 year scores where relatively more of the non-manual children are found. A similar result could be obtained by a non-linear but monotone transformation of 11 year scores which tended to "shrink" the high 11 year scores relative to the low ones. Thus, interpretations of particular interactions are problematical and we can adopt one of two procedures to avoid the difficulty. The first, and perhaps the simplest, is simply to constrain the fitted lines to be parallel and this is the usual assumption made in the analysis of covariance. The second procedure is to fit non-parallel lines and then average the differences between them over a suitable distribution of 7 year scores, for example that for all the children in the population. A discussion of such averaging procedures is given by Cochran and Rubin (1973). The different methods will give similar results where the degree of non-parallelism is not very great, but where say the

lines intersect in the range of scores of interest, considerable care will be needed (Aitkin, 1973).

A further complication is introduced if the measurement error variances of the 7 year scores differ between the social classes, since "unadjusted" parallel lines then become non-parallel after adjustment.

If a common adjusted regression slope is fitted for each social class, then if $\sigma^2_{ui}$ is the measurement error variance in social class $i$ containing $n_i$ subjects, the adjustment to the variance of the test score is

$$-\sum_i n_i \sigma^2_{ui} / \sum_i n_i.$$

This is simply a weighted average of the separate measurement error variances. Thus, where these variances do not differ greatly between the social classes, a correction based on the overall reliability will approximate to the correct one. We shall assume in the remainder of the paper that all the adjusted regression slopes are parallel, with the adjustments based on the overall reliability values given in Table 3.

TABLE 6

A. *Eleven year scores related to 7 year scores and 7 year social class, adjusted for measurement error. (Figures in brackets are unadjusted for measurement error)*

| | *Reading* | | | | *Mathematics* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Fitted value* | *S.E.* | *d.f.* | $\chi^2$ | *Fitted value* | *S.E.* | *d.f.* | $\chi^2$ |
| Overall | −0·001 (−0·013) | | | | 0·021 (0·013) | | | |
| 7 year score | 0·789 (0·609) | 0·013 | 1 | 3683·6*** | 0·820 (0·508) | 0·018 | 1 | 2075·3*** |
| Social class at 7 | | | | | | | | |
| Non-manual | 0·217 (0·305) | | | | 0·309 (0·408) | | | |
| Skilled and semi-skilled | −0·051 (−0·060) | | 2 | 227·9*** | −0·054 (−0·061) | | 2 | 465·2*** |
| Unskilled | −0·166 (−0·245) | | | | −0·255 (−0·347) | | | |
| Residual variance | 0·48 | | | | 0·44 | | | |

B. *Sixteen year score related to 7 year scores and 11 year social class, adjusted for measurement error. (Figures in brackets are unadjusted for measurement error)*

| | *Reading* | | | | *Mathematics* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Fitted value* | *S.E.* | *d.f.* | $\chi^2$ | *Fitted value* | *S.E.* | *d.f.* | $\chi^2$ |
| Overall | 0·115 (0·112) | | | | −0·008 (−0·009) | | | |
| 7 year score | −0·117 (0·130) | 0·047 | 1 | 6·2* | 0·084 (0·079) | 0·037 | 1 | 5·2* |
| 11 year score | 1·068 (0·706) | 0·057 | 1 | 351·1*** | 0·767 (0·718) | 0·027 | 1 | 807·0*** |
| Social class at 11 | | | | | | | | |
| Non-manual | 0·027 (0·127) | | | | 0·115 (0·144) | | | |
| Skilled and semi-skilled | 0·046 (0·023) | | 2 | 16·0*** | −0·033 (−0·038) | | 2 | 52·9*** |
| Unskilled | −0·073 (−0·150) | | | | −0·082 (−0·106) | | | |
| Residual variance | 0·30 | | | | 0·45 | | | |

Returning to equation (9a), we note that although Fogelman and Goldstein (1976) did not present their results adjusted for measurement error, they did investigate the effect of using reliability values down to 0·8. They concluded that the adjustments to the social class fitted values were relatively small and did not lead to markedly different conclusions. For 7 year mathematics, however, the reliability is below these values and Table 6A can therefore be regarded as an updated version of the corresponding table in Fogelman and Goldstein.

In fact, for both reading and mathematics in Table 6A, the social class differences are reduced by no more than 35 per cent after adjusting for measurement error. Thus the results tend to support this assertion of Fogelman and Goldstein. The differences for mathematics, however, are about 50 per cent greater than those for reading, whereas using an age equivalent scale Fogelman and Goldstein have found them only 20 per cent greater. At the age of 7 the average difference between the non-manual and skilled or semi-skilled manual is about 0·50 units and between the latter and unskilled manual about 0·43 units for reading, and correspondingly 0·39 and 0·25 for mathematics. We see therefore that for reading the additional differences between these social groups (0·27 and 0·12 respectively) are rather less than the pre-existing ones, and for mathematics (0·35 and 0·21 respectively) are approximately equal to the pre-existing ones.

The analysis of the 16 year scores shows that for reading there is now a negligible difference between the non-manual and skilled or semi-skilled group, after allowing for 7 and 11 year score, although there is a difference of 0·12 units between the latter and the unskilled group. We would infer therefore that the difference between the non-manual and skilled or semi-skilled group does not increase further between 11 and 16 years, whereas the unskilled group continue to fall behind and by the same amount as between 7 and 11 years. We note, however, that the unadjusted constants for social class show a fairly large difference (0·10 units) between the non-manual and skilled or semi-skilled groups. This is the difference we would use if we wished to make predictions about expected differences in 16 year scores between children from these two social groups on the basis of known 11 year scores. This statement would change if an essentially equivalent but more reliable test were used. Hence the estimates adjusted for reliability provide a baseline for inferences, although we must recognize that perfectly reliable tests do not occur in practice. We shall return to this point in the discussion.

It is worth pointing out that when average differences between the social classes are expressed in terms of proportions of children with attainments below or above particular extreme values, then very large differences become apparent. This is relevant where decisions are made about, say, remedial or special education. For example, the average rate of change of reading score with age at 11 years is about 0·5 standard deviation units per year (Fogelman and Goldstein, 1976). Thus the 3rd percentile of the distribution of 11 year score corresponds to approximately 4 years retardation of reading, although we should be careful not to take such a linear extrapolation too far. Nevertheless, this corresponds to the 1st percentile for the non-manual children and about the 4th percentile for the children of unskilled manual workers. We can also see from Table 6 that for those children with reading attainments at about the 3rd percentile at age 7, about 20 per cent of those who are in the non-manual group are below the 3rd percentile at age 11 compared to about 40 per cent of the children of unskilled manual workers.

### 9.1. *Changes in Attainment and Social Mobility*

The analyses so far have been only partly longitudinal in that the independent variable, social class, has entered only at one point in time. As Table 7 shows, however, there is considerable social class mobility between the three ages with only 75 per cent remaining in the same social class group at all ages. It is also of interest that there is a general "upward" mobility with an overall 6 per cent of children moving into the non-manual group and 1 per cent moving out of the unskilled group. We now construct a full longitudinal model which takes account of these changes.

TABLE 7

*Social class changes between 7, 11 and 16 years of age*
*(Figures in brackets are percentages on sample total)*

| *Social class at 16* | *Social class at 11* | Social class at 7: *Non-manual* | *Skilled or semi-skilled manual* | *Unskilled manual* | *Total* |
|---|---|---|---|---|---|
| Non-manual | Non-manual | 2043 (25·3) | 321 (4·0) | 8 (0·1) | 2372 (29·4) |
| | Skilled and semi-skilled manual | 108 (1·3) | 385 (4·8) | 10 (0·1) | 503 (6·2) |
| | Unskilled manual | 2 (0·0) | 7 (0·1) | 8 (0·1) | 17 (0·2) |
| | Total | 2153 (26·7) | 713 (8·8) | 26 (0·3) | 2912 (36·0) |
| Skilled and semi-skilled manual | Non-manual | 141 (1·7) | 186 (2·3) | 5 (0·1) | 332 (4·1) |
| | Skilled and semi-skilled manual | 165 (2·0) | 3841 (47·5) | 194 (2·4) | 4200 (52·0) |
| | Unskilled manual | 11 (0·1) | 119 (1·5) | 82 (1·0) | 212 (2·6) |
| | Total | 317 (3·9) | 4146 (51·3) | 281 (3·5) | 4744 (58·7) |
| Unskilled manual | Non-manual | 5 (0·1) | 7 (0·1) | 3 (0·0) | 15 (0·2) |
| | Skilled and semi-skilled manual | 7 (0·1) | 125 (1·5) | 49 (0·6) | 181 (2·2) |
| | Unskilled manual | 2 (0·0) | 84 (1·0) | 150 (1·9) | 236 (2·9) |
| | Total | 14 (0·2) | 216 (2·7) | 212 (2·6) | 432 (5·3) |
| Total | Non-manual | 2189 (27·1) | 514 (6·4) | 16 (0·2) | 2719 (33·7) |
| | Skilled and semi-skilled manual | 280 (3·5) | 4351 (53·9) | 253 (3·1) | 4884 (60·5) |
| | Unskilled manual | 15 (0·2) | 210 (2·6) | 240 (3·0) | 465 (5·8) |
| | Total | 2484 (30·8) | 5075 (62·8) | 509 (6·3) | 8068 (100·0) |

The following path diagram shows one such model. The arrows denote which independent variables are influencing the dependent variables. Thus the 11 year score can be related to social class at both 7 and 11 in addition to 7 year test score. (To avoid too much complication we have omitted the distant effect of 7 year social class on 16 year score.) The equations for

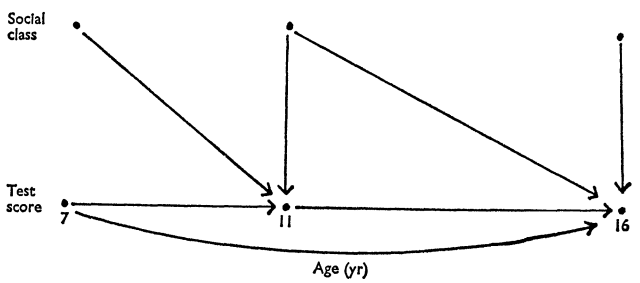


FIG. 9. Path diagram showing the direction of assumed causal relationships between social class and test score measurements at 7, 11 and 16 years.

this model can readily be written down by adding appropriate terms to equations (9a) and (9b). This is not, however, the most useful paremeterization of the problem. Social class at 7 years is highly correlated with social class at 11 and the latter is highly correlated with social class at 16. Thus we are in danger of some instability occurring in the parameter estimates. Also, it is

unlikely that a purely additive model will suffice to explain the social class effects. Hence we have parameterized social class in terms of social class at 11, change between 7 and 11 years and change between 11 and 16 years, these change variables being easily interpretable components of the first order interactions. (Tests for higher order interactions between the social class factors give non significant results). Table 8 presents the results of the analysis.

TABLE 8

A. *Eleven year scores related to 7 year scores, social class and social class changes, adjusting for measurement error*

| | | *Reading* | | | | *Mathematics* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Fitted constant* | *S.E.* | *d.f.* | $\chi^2$ | *Fitted constant* | *S.E.* | *d.f.* | $\chi^2$ |
| Overall | | 0·036 | | | | −0·065 | | | |
| 7 year score | | 0·781 | 0·013 | 1 | 3637·0*** | 0·814 | 0·018 | 1 | 2045·0*** |
| Social class at 7 | Non-manual (a) | 0·231 | | | | 0·314 | | | |
| | Skilled and semi-skilled (b) | −0·067 | | 2 | 248·6*** | −0·079 | | 2 | 458·4*** |
| | Unskilled (c) | −0·164 | | | | −0·235 | | | |
| Change in social class between 7 and 11 | c → a+b | −0·027 | 0·050 | | | −0·081 | 0·048 | | |
| | b → a | 0·169 | 0·030 | | | 0·154 | 0·029 | | 69·1*** |
| | a+b → c | −0·167 | 0·046 | 4 | 66·8*** | −0·136 | 0·044 | | |
| | a → b | −0·187 | 0·043 | | | −0·193 | 0·041 | | |
| | No change | 0·000 | | | | 0·000 | | | |
| Residual variance | | 0·48 | | | | 0·44 | | | |

B. *Sixteen year scores related to 11 and 7 year scores, social class at 11 and social class changes, adjusting for measurement error.*

| | | *Reading* | | | | *Mathematics* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Fitted constant* | *S.E.* | *d.f.* | $\chi^2$ | *Fitted constant* | *S.E.* | *d.f.* | $\chi^2$ |
| Overall | | 0·112 | | | | −0·005 | | | |
| 7 year score | | −0·120 | 0·046 | 1 | 6·8** | 0·082 | 0·037 | 1 | 4·9* |
| 11 year score | | 1·066 | 0·057 | 1 | 349·8*** | 0·760 | 0·026 | 1 | 854·4*** |
| Social class | Non-manual (a) | 0·039 | | | | 0·155 | | | |
| | Skilled and semi-skilled (b) | 0·054 | | 2 | 11·3** | −0·041 | | 2 | 69·3*** |
| | Unskilled (c) | −0·093 | | | | −0·114 | | | |
| Change in social class between 7 and 11 | c → a+b | −0·098 | 0·043 | | | −0·064 | 0·053 | | |
| | b → a | −0·060 | 0·033 | | | −0·148 | 0·041 | | |
| | a+b → c | 0·059 | 0·058 | 4 | 10·1* | 0·028 | 0·072 | 4 | 14·4** |
| | a → b | 0·007 | 0·041 | | | −0·026 | 0·050 | | |
| | No change | 0·000 | | | | 0·000 | | | |
| Change in social class between 11 and 16 | c → a+b | −0·015 | 0·060 | | | 0·029 | 0·074 | | |
| | b → a | 0·033 | 0·034 | 4 | 7·2 | 0·115 | 0·042 | | |
| | a+b → c | −0·087 | 0·056 | | | −0·091 | 0·069 | 4 | 10·2* |
| | a → b | 0·079 | 0·044 | | | −0·028 | 0·054 | | |
| | No change | 0·000 | | | | 0·000 | | | |
| Residual variance | | 0·29 | | | | 0·45 | | | |

Sample size = 5200

By comparison with Table 6 the addition of changes in social class between 7 and 11 years to the analysis of 11 year scores has little effect on the constants fitted for social class at 7. The effect of moving out of the unskilled group is similar to no change for both mathematics and reading. For reading, the difference between those moving from the skilled or semi-skilled group into the non-manual group and those staying in the skilled or semi-skilled group is 0·17 units which is about half the difference between these groups at 7. A similar result holds for those moving in the opposite direction, and also for those moving into the unskilled group. Similar results are found for mathematics. It should be noted in this and other tables which include social class changes, that the constants fitted to the change categories have very low intercorrelations—mostly about 0·05, and always less than 0·20.

For the analysis of 16 year scores, the changes in social class have relatively small effects, as can be seen by studying the estimated standard errors. For both reading and mathematics, a reduction in score seems to be associated with upward mobility between 7 and 11, but this is simply a reflection of the fact that 11 year score is also fitted, so that compared to those with no change, namely both non-manual at 7 and at 11, the upwardly mobile are those who were in the non-manual group at eleven but not at 7. For mathematics, those moving from the semi-skilled or skilled group at 11 into the non-manual group at 16 have an estimated increase of about 0·12 units over those who remain in the former groups.

## 10 Relationships between Test Scores

Until now we have analysed reading and mathematics attainment separately. In general, however, school attainments influence each other, for example the ability to read and comprehend the written description of a mathematical problem is an important component of the ability to solve it. Thus the 11 and 16 year mathematics scores may depend partly on reading attainments. It seems less likely that reading attainment depends in the same way upon mathematics attainment, but we shall include this possibility in the models we now develop.

Using $x$ to denote reading score and $y$ mathematics score equation (3a) becomes

$$x_2 = \alpha_{11} + \beta_{11} x_1 + \beta_{21} y_1 + \gamma_{21} y_2 + \varepsilon_1 \tag{11a}$$

$$y_2 = \alpha_{12} + \beta_{12} x_1 + \beta_{22} y_1 + \gamma_{12} x_2 + \varepsilon_2 \tag{11b}$$

with a similar extension to (3b). If we substitute for $y_2$ and $x_2$ in (11a) and (11b) we obtain

$$\left.\begin{aligned} x_2 &= (1-\gamma_{21}\gamma_{12})^{-1}\{(\alpha_{11}+\alpha_{21}\alpha_{12}) + (\beta_{11}+\gamma_{21}\beta_{12})x_1 + (\beta_{21}+\gamma_{21}\beta_{22})y_1 + (\gamma_{21}\varepsilon_2+\varepsilon_1)\}, \\ y_2 &= (1-\gamma_{21}\gamma_{12})^{-1}\{(\alpha_{12}+\alpha_{21}\alpha_{12}) + (\beta_{12}+\gamma_{12}\beta_{11})x_1 + (\beta_{22}+\gamma_{12}\beta_{21})y_1 + (\gamma_{12}\varepsilon_1+\varepsilon_2)\} \end{aligned}\right\} \tag{12}$$

Thus $x_2$ is not independent of the error term $\varepsilon_2$ unless $\gamma_{21}$ is zero, and nor is $y_2$ independent of $\varepsilon_1$ unless $\gamma_{12}$ is zero. Thus, if we use ordinary least squares with equations (11a) and (11b) we shall obtain inconsistent estimates. If we have instrumental variables then we can use these as before to obtain consistent estimates. Provided they satisfy the appropriate assumptions they will also adjust for measurement error in $x_2$ and $y_2$. The "reduced form" parameters of equation (12) can be consistently estimated, however. If we use ordinary bivariate least squares we obtain consistent and efficient estimates of the six composite coefficients together with the three elements of the covariance matrix of residuals. The original number of parameters, however, is ten so that these are not individually identified. In order to obtain identifiability we need to introduce at least one linear constraint. If we set $\gamma_{21} = 0$, for example, then we shall be able to identify the parameters, and also obtain consistent estimates using ordinary least squares with equations (11a) and (11b).

A similar situation exists for the extension of equation (3b) where we can either use instrumental variables, or introduce a linear constraint.

Table 9 shows the fitted coefficients using the teachers' ratings of reading and mathematics at 11 and 16 years as instrumental variables, together with the measurement error variance

TABLE 9

*Relationships between reading and mathematics test scores*
*Two-stage least squares estimates for equations (11a) and (11b) using teachers' ratings as instrumental variables and adjusted for measurement error*

A. *Dependent variable* 11 *year score*

| | *Reading* | | *Mathematics* | |
|---|---|---|---|---|
| | *Fitted constant* | *S.E.* | *Fitted constant* | *S.E.* |
| Overall constant | 0·007 | | 0·036 | |
| 7 year reading | 0·600 | 0·023 | 0·261 | 0·024 |
| 7 year mathematics | 0·071 | 0·024 | 0·487 | 0·026 |
| 11 year reading | — | — | 0·395 | 0·016 |
| 11 year mathematics | 0·331 | 0·015 | — | — |
| Residual variance | 0·45 | | 0·35 | |

B. *Dependent variable* 16 *year score*

| | *Reading* | | *Mathematics* | |
|---|---|---|---|---|
| | *Fitted constant* | *S.E.* | *Fitted constant* | *S.E.* |
| Overall | 0·150 | | −0·047 | |
| 7 year reading | −0·133 | 0·043 | — | |
| 7 year mathematics | — | — | 0·053 | 0·021 |
| 11 year reading | 1·167 | 0·112 | 0·097 | 0·027 |
| 11 year mathematics | −0·176 | 0·059 | 0·571 | 0·028 |
| 16 year reading | — | — | 0·292 | 0·017 |
| 16 year mathematics | 0·155 | 0·017 | — | — |
| Residual variance | 0·30 | | 0·42 | |

Sample size = 8200

estimates for the remaining variables. For the 11 year reading score, the 7 year mathematics score contributes rather little and the coefficient of the 11 year mathematics score is only just over half that of the 7 year reading coefficient. For 11 year mathematics, however, both the 7 and 11 year reading score coefficients are substantial, with the latter not much less than the 7 year mathematics score coefficient. For the 16 year reading score, the dominant coefficient is that for 11 year reading, and for 16 year mathematics the 16 year reading coefficient is just over half the 11 year mathematics coefficient.

We saw that setting $\gamma_{21} = 0$ gave identifiable coefficients in (11a) and (11b), and if we additionally set $\beta_{21} = 0$ we have a model in which reading score is allowed to influence mathematics but not vice versa. Table 10 shows the results of analysing this model and the corresponding one for 16 year scores. For the 11 year reading score there is a 40 per cent increase in the 7 year reading coefficient when the mathematics coefficients are set to zero. For the 11 year mathematics score there are considerable changes in the coefficients resulting from the substitution of the measured 11 year reading score for the predicted 11 year reading score, with the 7 year reading coefficient being very small. As mentioned in a previous section, there is some doubt about whether the chosen instrumental variable, the teachers' rating of the child's use of books, satisfies all the necessary assumptions.

For the 16 year reading score the coefficients are little changed by the omission of mathematics scores and for 16 year mathematics the 11 and 16 year reading coefficients show fairly

TABLE 10

*Relationships between reading and mathematics test scores, adjusted for measurement error*

A. *Dependent variable* 11 *year score*

| | *Reading* | | *Mathematics* | |
|---|---|---|---|---|
| *Source* | *Fitted constant* | *S.E.* | *Fitted constant* | *S.E.* |
| Overall constant | 0·015 | | 0·035 | |
| 7 year reading | 0·841 | 0·012 | −0·033 | 0·033 |
| 7 year mathematics | — | — | 0·376 | 0·022 |
| 11 year reading | — | — | 0·700 | 0·040 |
| Residual variance | 0·49 | | 0·25 | |

B. *Dependent variable* 16 *year score*

| | *Reading* | | *Mathematics* | |
|---|---|---|---|---|
| | *Fitted constant* | *S.E.* | *Fitted constant* | *S.E.* |
| Overall constant | 0·154 | | −0·063 | |
| 7 year reading | −0·147 | 0·051 | — | — |
| 7 year mathematics | — | — | 0·031 | 0·021 |
| 11 year reading | 1·113 | 0·059 | −0·199 | 0·069 |
| 11 year mathematics | — | — | 0·681 | 0·034 |
| 16 year reading | — | — | 0·393 | 0·044 |
| Residual variance | 0·30 | | 0·42 | |

Sample size = 8200

large changes in value. Thus, with the reservation about the chosen instrumental variables, these results suggest that there is an appreciable reciprocal relationship between mathematics and reading at 11 years but that there is little effect of mathematics scores on 16 year reading score. The latter result is as expected, and an explanation of the former might be sought in the closer integration of subject teaching found in primary schools, with the mathematics test score possibly partly being a surrogate for other unmeasured variables.

We may introduce further variables such as social class into these models, but our interpretation of the fitted constants will change. By including the effect of reading attainment in the model for mathematics we are no longer studying simply the difference between social classes and mobility patterns, conditioned on earlier mathematics attainment. Rather we are studying the way in which any such differences are altered by taking account of reading attainment, especially that measured at the same age, which is itself dependent on social class.

Table 11 shows the results of incorporating the social class variables into the models analysed in Table 10 for mathematics and it will be seen, by comparison with Table 8, how this reduces the difference between the social class fitted constants. The extent to which these differences are reduced could be interpreted as the extent to which social class acts on mathematics attainment through its effect on reading attainment. For example, the negligible difference between the skilled or semi-skilled and unskilled social classes at 11 when compared with a difference of 0·07 units in Table 8 together with the residual differences between the change categories, might indicate that the additional change in mathematics attainment between these social class groups between 11 and 16 years could be almost entirely accounted for by the effect of social class on reading attainment.

TABLE 11

*Mathematics attainment related to earlier attainment in mathematics, reading and social class, adjusted for measurement error*

| | | 11 *year mathematics* | | | | 16 *year mathematics* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Fitted constant* | *S.E.* | *d.f.* | $\chi^2$ | *Fitted constant* | *S.E.* | *d.f.* | $\chi^2$ |
| Overall | | −0·014 | | | | −0·011 | | | |
| 7 year reading | | −0·023 | 0·032 | 1 | 0·5 | | | | |
| 7 year mathematics | | 0·375 | 0·022 | 1 | 290·5*** | 0·038 | 0·027 | 1 | 2·0 |
| 11 year reading | | 0·670 | 0·040 | 1 | 280·6*** | −0·266 | 0·091 | 1 | 8·5** |
| 11 year mathematics | | — | | | | 0·690 | 0·044 | 1 | 245·9*** |
| 16 year reading | | — | | | | 0·420 | 0·060 | 1 | 49·0*** |
| 7 year social class | Non-manual (a) | 0·070 | | | | — | | | |
| | Skilled and semi-skilled (b) | −0·024 | | 2 | 28·1*** | — | | | |
| | Unskilled (c) | −0·046 | | | | — | | | |
| 11 year social class | Non-manual (a) | — | | | | 0·111 | | | |
| | Skilled and semi-skilled (b) | — | | | | −0·054 | | 2 | 47·5*** |
| | Unskilled (c) | — | | | | −0·057 | | | |
| Change in social class between 7 and 11 | c → a+b | −0·051 | 0·036 | | | −0·018 | 0·052 | | |
| | b → c | 0·020 | 0·023 | | | −0·121 | 0·040 | | |
| | a+b → c | 0·008 | 0·034 | 4 | 2·9 | 0·012 | 0·070 | 4 | 9·8* |
| | a → b | −0·008 | 0·032 | | | −0·038 | 0·049 | | |
| | No change | 0·000 | | | | 0·000 | | | |
| Change in social class between 11 and 16 | c → a+b | — | | | | 0·028 | 0·072 | | |
| | b → a | — | | | | 0·082 | 0·040 | | |
| | a+b → c | — | | | | −0·034 | 0·066 | 4 | 5·8 |
| | a → b | — | | | | −0·057 | 0·052 | | |
| | No change | — | | | | 0·000 | | | |
| | Residual variance | 0·25 | | | | 0·42 | | | |

Sample size = 5200

## 11. CONCLUSIONS

Several problems have arisen during the course of analysing the models in this paper. Some of these are to do with the substantive subject matter and although interesting, they are not the central concern of this paper which is methodological. Nevertheless, many of the findings presented here are new and it is hoped that further more detailed analyses along these lines will be made. Three methodological problems have concerned us.

First, the definition of "change" adopted here solves certain problems of scale definition which arise when simple difference scores are used. It also raises the problem of interactions or non-parallelism of regression lines, to which there seems to be no completely satisfactory solution. In some circumstances, preliminary transformations of the variables to produce linearity, homoscedasticity or normality may also have the effect of eliminating such interactions, and this has been shown in the present data.

A second problem is the need to provide accurate estimates of measurement error variances or reliability coefficients. From the point of view of efficiency these should be estimated on as large a sample as possible. If this is not done then there will be little purpose in increasing the study sample size beyond a certain point where their sampling errors begin to dominate the standard errors of the estimated coefficients. We need to recognize also that the various

methods of estimating measurement error or reliability have unsatisfactory aspects and, where possible, results using different methods should be compared. Related to this problem is that of finding suitable instrumental variable estimators, either for estimating measurement errors or for providing consistent estimators for systems of nonrecursive simultaneous equations such as (11a) and (11b). The data analysed here suggest that the most obvious of such estimators do not necessarily satisfy the necessary assumptions, at least for reading attainment. It is planned to pursue elsewhere a more detailed study of alternative instrumental variable estimators.

The third major problem is in the interpretation to be given to measurement error adjusted estimates. It is clear that if we accept the notion of a "true" attainment score and if we wish to make statements about these scores then the adjusted estimates should be used. In practice it may be possible to approximate to these true scores by carrying out sufficiently detailed testing of individuals, although we need to remember that different tests will generally be reflecting somewhat different underlying "true" attainments. On the other hand, statements concerned with prediction on the basis of an observed score should be based on models which incorporate a measurement error component appropriate to the prediction variable used. There are certainly some situations where educational decisions are made on the basis of observed scores which include measurement errors.

Transfer arrangements from primary to secondary school are often still based, at least partly, on test scores and it might be appropriate to base comparisons of, say, different schools on scores which include measurement errors. In this case we would start with the estimated model from which measurement errors had been removed and adjust it by incorporating the error variances appropriate to the tests actually used. These might be different from those used in the present or a similar study, especially where several tests or assessments are used jointly. If we found that our inferences were considerably changed when adjustment was made for unreliability, this might direct our attention to particular types of interpretation which took account of the way in which information, derived from different assessment instruments with differing reliabilities, was actually used to make educational decisions.

## ACKNOWLEDGEMENTS

My grateful thanks are due to the following who made constructive criticisms of an early draft; Murray Aitkin, Carlyle Maw, Lee J. Cronbach, Ken Fogelman, Michael Healy, Dougal Hutchison and Ian Plewis. The analysis was partly supported by the National Institute of Education, Washington, U.S.A. (contract no. 400–76–0041).

## REFERENCES

AITKIN, M. (1973). Fixed width confidence intervals in linear regression with applications to the Johnson–Neyman technique. *Brit. J. Math. and Statist. Psychol.*, **26**, 261–269.

ANDERSON, T. W. (1976). Estimation of linear functional relationships: approximate distributions and connections with simultaneous equations in econometrics. *J. Roy. Statist. Soc.* B, **38**, 1–36.

COCHRAN, W. G. (1975). Some effects of errors of measurement on linear regression. *Proc. 6th Berkeley Symp.*, Berkeley: Univ. of California Press.

COCHRAN, W. G. and RUBIN, D. R. (1973). Controlling bias in observational studies: a review. *Sankhyā*, **35**, 417–446.

FOGELMAN, K. R. and GOLDSTEIN, H. (1976). Social factors associated with changes in educational attainment between 7 and 11 years of age. *Educational Studies*, **2**, 95–109.

FOGELMAN, K. R., GOLDSTEIN, H., ESSEN, J. and GHODSIAN, M. (1978). Patterns of Attainment. *Educational Studies*, 4, 121–130.

FOGELMAN, K. R. (1976). *Britain's Sixteen-Year-Olds.* London: National Children's Bureau.

FULLER, W. A. and HIDIROGLOU, M. A. (1978). Regression estimation after correcting for attenuation. *J. Amer. Statist. Ass.*, **73**, 99–104.

JOHNSTON, J. (1973). *Econometric Methods.* New York: McGraw-Hill.

JÖRESKOG, K. G. and SÖRBOM, D. (1976). Statistical models and methods for analysis of longitudinal data. In *Latent Variables in Socio-Economic Models* (D. J. AIGNER and A. S. GOLDBERGER, eds). Amsterdam: North Holland.

KENDALL, M. G. and STUART, A. (1967). *The Advanced Theory of Statistics*, Vol. 2. London: Griffin.

LINDLEY, D. V. (1947). Regression lines and the linear functional relationship. *J. R. Statist. Soc.* B, **9**, 218–244.

LORD, F. M. and NOVICK, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison–Wesley.

REGISTRAR-GENERAL (1960). *Classification of Occupations.* London: HMSO.

RICHARDSON, D. M. and WU, D. (1970). Least squares and grouping estimators in the errors in variables model. *J. Amer. Statist. Ass.*, **85**, 724–728.

SOUTHGATE, V. (1962). *Southgate Group Reading Tests: Manual of Instruction.* London: University of London Press.

TANNER, J. M., WHITEHOUSE, R. H. and TAKAISKI, M. (1966). Standards from birth to maturity for height, weight, height velocity and weight velocity: British children, 1965. II. *Arch. Dis. Child*, **41**, 813–635.

WARREN, R. D., WHITE, J. K. and FULLER, W. A. (1974). An errors-in-variables analysis of managerial role performance. *J. Amer. Statist. Ass.*, **69**, 886–893.

## DISCUSSION OF PROFESSOR GOLDSTEIN'S PAPER

Professor MURRAY AITKIN (University of Lancaster): I am pleased to be able to move the vote of thanks to Harvey Goldstein for this important paper. There have been few studies of the effects of measurement error on the practical conclusions to be drawn from complex educational studies, and this careful and detailed examination of the consequences of measurement error on the interpretation of social class differences is therefore very welcome.

Measurement error is an ever-present reality in educational measurement, and of course the quoting of reliabilities of standard tests is commonplace. It is widely accepted that tests should be highly reliable, but the actual *use* of known or estimated reliabilities in the analysis of experimental or observational data is still uncommon.

The effect of unreliability in the 7 and 11 year reading scores is particularly noticeable in the equation for the 16 year reading score. The regression equation based on the observed score is, from Table 2B,

$$\hat{\mu}_{x_{16}} = 0{\cdot}16 + 0{\cdot}14X_7 + 0{\cdot}73X_{11},$$

giving a conditional mean $\hat{\mu}$ of 0·75 for $X_7 = -1$ and $X_{11} = +1$. The estimated true score regression equation corrected for unreliability is, from Table 4B,

$$\hat{\mu}_{x_{16}} = 0{\cdot}15 — 0{\cdot}15x_7 + 1{\cdot}11x_{11}.$$

For *true* scores of $x_7 = -1$ and $x_{11} = +1$, we have $\hat{\mu} = 1{\cdot}41$, nearly 0·7 of an (observed score) standard deviation above the first estimate. Of course, we do not know the true scores. However, if we estimate them from the corresponding observed scores, using the reliability estimates in Table 3, we have $\hat{x}_7 = \hat{\rho}_7 X_7 = -0{\cdot}79$, and $\hat{x}_{11} = \hat{\rho}_{11}x_{11} = 0{\cdot}82$, and for these values, $\hat{\mu} = 1{\cdot}18$. Unreliability here makes a very substantial difference to the conclusions. Comparable differences occur in the more complex models.

The last half of the last paragraph of Section 7 deserves particular emphasis. If reliability or measurement error estimates are based on a sample (e.g. from a pilot study) much smaller than that used for the main investigation, the standard errors of estimated parameters will be greatly inflated after adjustment for unreliability, particularly if the reliabilities are low. The consequences for the design of such studies are important: every effort should be made to obtain internal reliability or measurement error variance estimates from the complete sample, for example by using parallel form or split-half methods.

A further analytical advantage results from such a procedure: straightforward ML methods (Jöreskog and Sörbom, 1976) are then available for parameter estimation (if we are willing to assume normality of the appropriate test scores after suitable transformation), in both the simple models of Section 8 and the complex models of Sections 9 and 10. Where these are available and valid, they will give greater precision than the consistent estimates described in the paper.

In a paper of this complexity, there are bound to be some unresolved issues. I would like clarification on three points:

(1) In paragraph 2 of Section 3, the simple difference score measure of "change" is rejected in favour of "the relationship between the measurements on the two occasions". The measure

adopted—based essentially on ANCOVA—suffers from its own difficulties of interpretation. For example, in Section 9, penultimate paragraph, there is a negligible difference in reading intercepts between the non-manual group and the skilled group, at age 16, "after allowing for 7 and 11 year score". The conclusion is drawn that "the difference between the non-manual and skilled or semi-skilled group *does not increase further between* 11 *and* 16 *years* . . .". Does this mean that the mean difference between classes at 16 is essentially the same as that at 11? For this is not the case. This same problem of interpretation is seen most clearly in the two-group experiment where the groups already differ systematically in mean on the covariate. If the two groups now receive different treatments, how are we to assess the effect of treatment? The ANCOVA tests whether the response mean difference is $\beta$ times the covariate mean difference. If response and covariate are measured on similar scales, and $\beta<1$, then a "no significant difference" result of the test simply means that the response mean difference has regressed towards zero by comparison with the covariate mean difference. But would this have been expected, if the treatments had been identical? It seems to me that the difficulties of interpretation of non-randomized studies remain severe.

(2) Given the care with which models are treated in the paper, I am unhappy about the disappearance of the interaction terms in Table 2B. It is said that their interpretation is unclear, which is not in itself sufficient reason for ignoring them. It may be true that they do not markedly affect the predicted values—though it is worth noting that the interaction for mathematics is not negligible, being larger than the 7 year intercept—but what is the effect of the correction for unreliability on the interaction, and how does it affect the social class differences examined subsequently?

(3) Finally, a minor point. I am puzzled by the statement at the end of Section 3 that if $\beta = 1$, the regression of $x_2$ on $x_1$ is not equivalent to the use of simple differences. The use of weight gain in diet experiments is equivalent to ANCOVA if the regression of final weight on initial weight has slope equal to 1. What is the difference here?

In conclusion, this paper has shown very clearly the importance of measurement error and the effect it can have on substantive conclusions drawn from educational studies. I have much pleasure in moving the vote of thanks.

Dr A. M. ADELSTEIN (Member of Council): It is a great pleasure to be able to second this vote of thanks. In doing so, I would like to make a few remarks, not so much about the technical aspects of the paper but rather about some of its implications.

To me, the analysis of this kind of data of educational attainment is enormously important. The political and social implications of this kind of analysis stretch across the world. Wherever one looks, be it in the capitalist or the communist world, there are these class differences in education. People interpret these differences to suit their own preconceived political views, and usually to maintain a particular kind of *status quo*. It thus becomes extremely important that at least the basic and fundamental facts about these measurements are properly and thoroughly analysed. I would say that this paper goes a long way in starting along that road. Although it does not directly lead us to the political and social conclusions which should be drawn from this, at least it does not mislead us in the social group sense, nor even in the individual sense. I am constantly amazed by the glib way, the sort of superficial way, in which attainment scores are used—as Professor Goldstein said—to determine the fate of individuals, in terms of their education or their therapy, or just in terms of their status as compared to their peers. In that sense, this paper seems an extremely important paper. The other aspect in which I am interested is this question of longitudinal studies. This particular study is world-famous. It is a very famous study, not only because of the educational statistics which Professor Goldstein has presented to us tonight, but for much else that was done in the study. Some of its fame, however, rests on fairly precarious grounds because critics say that studies of this kind are very expensive, they take a long time—so that by the time the results come out most people have lost interest in what the studies were all about—and it is difficult to get people to stay with the studies. This study was started in 1958, so we are 20 years on with the results.

This is one of a series of studies. There are three well-known ones, which people not familiar with them may not know: 1946, 1958 and 1970. It is quite likely that another will be carried out. A good deal of argument is going on among those who will pay for any such study. It is a sign of the times that study like this would cost today some millions of pounds, whereas in 1958 it cost very much less—it was really a very cheaply-run study. However, the decision whether another

study will take place depends to a large extent on how well the studies are analysed, and how much comes out of them. Of course, the critics are waiting to ask what has been done with all this work. Therefore, I am particularly pleased to be able to be here today to learn that Professor Goldstein has made this analysis. He has done two or three other analyses of longitudinal data from these studies of which I am aware. There are other aspects of these studies which do not depend on the fact that they are longitudinal studies, but Professor Goldstein has busied himself particularly with the longitudinal aspects of them. Because of the work which he and some of his colleagues have put into this, we are possibly closer to being able to make a decision that it is worth doing another study of this kind.

I would like to second the vote of thanks.

The vote of thanks was carried by acclamation.

Mr DOUGAL A. HUTCHISON (National Children's Bureau): First, I would like to say that I am very impressed by this careful and painstaking investigation of some of the problems which arise with longitudinal data. I have already commented on an earlier draft of the paper, but one or two comments remain that I want to make.

The social sciences have taken over techniques containing simplifying assumptions. These assumptions, on the one hand, have become so built in to the discipline that it is extremely difficult even to notice them but, on the other hand, they are not necessarily reasonable. Perhaps the two most striking of these are the assumption of the absence of measurement error in independent variables and the ability to conduct randomized experiments, thus allowing the researcher to ignore specification error, as far as omitting important contributory factors in analysis is concerned.

Until now measurement error, as far as social statistics is concerned, has been rather like the weather. Everybody talks about it, but nobody does anything about it. This careful exposition, therefore, is quite a pioneer effort.

I particularly liked the clear exposition not only of statistical questions, but also his investigation of the problems that had arisen, the steps he took to solve them and his estimates of how successful he was and also the real-life relevance of the statistical procedures.

There are some questions. It is probable that the standard assumptions of error theory are violated. For example, the 11 year test score for reading was the simple sum of the number of correct answers on a multiple choice test. Thus, it may be that the error score is negatively correlated with the true score.

I also wonder how reasonable it is to assume that social class is measured without error (see Osborne and Morris, 1979).

It is possible, too, that in disposing of the possible source of bias due to measurement error other biases have been increased. As I commented earlier, experimental techniques allow us to ignore the effect of mis-specifications, since it is assumed that by randomization nuisance variables will have no systematic effect. However, this certainly does not hold in this case—there is no question of being randomly assigned to social class. Cronbach and co-workers have shown that the bias that can arise in non-randomized analyses of covariance depends on the intercorrelations between the ideal covariate, the actual variate used and the linear function which best discriminates between the groups. This may well be an aspect to be considered in any further analysis of these data by Professor Goldstein, or in the comparison of types of school discussed in the paper.

My remarks have referred to a number of questions which are not dealt with in this paper. This is not intended as criticism but rather as a suggestion of where to go next.

Mr I. PLEWIS (Institute of Education, University of London): It is very pleasing to listen to a paper which faces up to the problems of analysing social and educational data and which sees those problems in the context of substantive issues. A major aim of longitudinal studies is to measure change and I agree that the seemingly obvious ways of doing this—simple individual differences and mean differences—are not usually satisfactory. Nevertheless, if there is prior knowledge which suggests that both true and error variances are constant over time then using mean differences to define group change does have the advantage of simplicity and, because true and observed means are equal, it does eliminate the problems caused by measurement error. Of course, such knowledge is rarely available and such conditions are unlikely to be found for measures concerned with aspects of children's development. Thus, Professor Goldstein has chosen to use regression models to measure change and, by considering the effects of measurement error in great detail, he has

shown that this problem is not quite as intractable as some of the critics of regression models in the social science literature have suggested. However, it is perhaps worth pointing out that although measurement errors in the dependent variables are unlikely to cause problems with these data, longitudinal data collected, for example, by the same interviewers on more than one occasion can result in correlated errors which will lead to inconsistent estimators—such errors are difficult to estimate. Also, I too wonder how reasonable it is to assume that social class is measured without error particularly when the information is not collected by personal interview?

An alternative approach to the identification of (11a) and (11b) would be to re-specify them as follows:

$$x_2 = \alpha_{11}+\beta_{11}x_1+\beta_{21}y_1+\gamma_{21}y_2+\delta_{21}x_2'+\varepsilon_1, \tag{1a}$$

$$y_2 = \alpha_{12}+\beta_{12}x_1+\beta_{22}y_1+\gamma_{12}x_2+\delta_{12}y_2'+\varepsilon_2. \tag{1b}$$

Here, $x_2'$ and $y_2'$ are the teachers' ratings of reading and mathematics; (1a) and (1b) are both exactly identified and could be motivated by the idea that the ratings are in fact measuring teacher expectations which might affect children's test scores. Unfortunately, this more complex model leads to further problems of measurement error.

Mr J. R. Ecob (Institute of Education, University of London): Professor Goldstein mentions in the conclusion the problem of finding suitable instrumental variable estimators which are uncorrelated with the errors of measurement but highly predictive of observed values. I am not quite clear what he means when he says that "the reliability estimate for reading was obtained similarly using teachers' ratings at the age of 7 years". How many ratings and which ones? Any number of variables can be used and I wonder if he may have excluded some useful ones.' For example, the following are relevant to reading test score: at 7 years: teacher ratings both of oral ability and of "reading", and an assessment of reading standard (book no. on reading scheme). At 11 years teacher ratings of the use of books and oral ability. In addition any other variable categorical, ordinal/or interval scaled and measured at any point in time which is related to the criterion could be used although there may be a trade off between increasing the number of variables and satisfying the assumptions involved.

I would support Professor Goldstein's decision to deal with observed variables rather than latent traits in regard to these data although a latent trait analysis could serve as an interesting comparison. The Jöreskog-type latent trait models have the advantage that they can include parameters for the correlation of measurement error over time (Jöreskog and Sörbom, 1976b, give an example of fitting their model where estimates of measurement errors of tests of reading and mathematics are found to be quite highly correlated over time, especially for the Scholastic Aptitude Tests, over periods 2 years apart). However, these models seem unable to cope with indicators of traits where these indicators are themselves correlated with background variables included in the model conditional on the trait value.

Recent results of mine suggest this is true in these data, when teacher rating and test score are taken as indicators of reading ability. Reading test score at age 7 was subtracted from teachers' rating of reading at the same age which was scaled to have the same distribution as the test score (this in a sense provides a measure of the change in ordering of subjects by both criteria). This difference was substantially explained by six factors. The order listed is that of the extent of explanation: Social class; behaviour in the home (an index of parental rating of aspects of behaviour in the last 3 months); Adequacy of accommodation (an index based on presence of bathroom, indoor lavatory, hot water supply, type of accommodation); Number of children in family younger than the respondent; Number of children in family older than the respondent; School behaviour (an index of teachers' ratings of aspects of classroom behaviour).

There may also be an argument for including teacher ratings of arithmetic or reading ability as a predictor. Indeed Hutchison, Prosser and Wedge (1979) found that at 7 years old this was more predictive of a 16 year test score than the corresponding test score for both reading and arithmetic. It is not clear whether this is because it measures a more appropriate aspect of ability at the age (in reading, "comprehension" as well as "word recognition"), whether it substitutes to a certain extent for the continuing effect of Social Class (shown in this paper) whether the reliability of measurement is higher, whether the distribution of scores is more even (the reading test at age 7 has a marked ceiling effect) or whether there are other reasons.

Dr F. H. HANSFORD-MILLER (Inner London Education Authority): I think perhaps the most valuable result in the field of education from Professor Goldstein's research is his finding of a close statistical relationship between attainment in mathematics with that in reading. This is very much in line with other recent educational research. Choat (1978), for example, states categorically that "language must be regarded as an integral part of mathematics", whilst Bruner (1964, 1966) believes that language provides both a means for transforming experience in the child and a vehicle for higher thought, including mathematics. Vygotsky (1962) contends that thought development in a child is actually determined by language.

Although Professor Goldstein makes the point that his central concern is methodological, I do not think he can escape so easily from some criticism of educational content. Indeed his educational findings have already brought forth comments this evening, and among the key words of the paper are "educational attainment", so obviously they will be widely discussed. I would like to ask, therefore, how far is his division of children according to their father's so-called "social class" still justified? In any event his "unskilled manual" class appears to be only some 5—6 per cent of the total. It may make an interesting statistical exercise, but is it not unfair and rather degrading to a child to categorize him in this way? Might not also the categorization itself affect the resulting data? I know much current educational thought seeks acquisition and documentation of social knowledge of this kind, but another school of thought sees teaching as a meeting of minds, to be kept, if possible, free from bias by what it would see as extraneous matters.

Table 7 and part of the associated text is concerned with social class changes between 7, 11 and 16 years of age—but how do such changes happen in practice? Consider a child moving out of the unskilled manual social class; does the child's father cease to be a dustman, say, and instead service the dust carts before they go out? In any case, the mother probably stays in the same social class, and how about Women's Lib. and the Equal Opportunities Commission? The mother, surely, has as big, if not a bigger, role in the upbringing of a child as the father and yet she is not mentioned as a causal element in categorizing social class. And how does the teacher get to know of any change in social class? Does the child put up his hand and say, "Please, Sir, I've moved up out of my social class?"!

Seriously, I believe such divisions are non-productive in research today, and on a par with the question of whether a house had an outside toilet, which similarly dominated research in housing policy for many decades. Modern elections, with their almost universal national trend, show that we are all today very much subject to the same influences. Instead of social class, I believe that in education we should be looking for more relevant criteria. What I should like to see is more research into the basic educational question of how children learn.

Consider some recent research on the learning processes in mathematics. Choat (1978) believes that perceptual space—that is, geometry—is the beginning of mathematics for the young child. He writes:

> "From birth a child is surrounded by space. Therefore, it is only natural that he wishes to explore this space to discover how he fits into it ... He begins to discover the properties of shapes by touching them ... He becomes conscious that objects have an identity and permanence ... Mathematics for young children is their development of the awareness of the spatial relationships".

This may be overstating the case for geometry but I feel geometry and other branches of mathematics should certainly have come into Professor Goldstein's research, instead of just arithmetic, for which he admits in Section 7 that "The 7 year mathematics test consisted of problem arithmetic items". Surely, that is a narrow and obsolete criterion for a test in mathematics today. The methodological standard of this paper is much needed for current educational research but the more difficult task, I believe, is the obtaining of more relevant and appropriate basic data of children's learning processes and achievements so as to ensure more meaningful and useful results.

Mr J. R. B. KING: It is important that the work which goes on in education should be brought to the attention of statisticians: for this reason, this is an important paper. As a general remark, I am horrified to see that in the past measurement error has not been included in work on which administrative decisions are based. Perhaps, on the other hand, I ought to be grateful for that because otherwise I might not have gone to the school to which I went.

Turning to the mathematical and methodological parts of the paper, the paper leans fairly heavily on well-established econometric techniques. It is, therefore, fair for us to ask the same questions of this work that we do of econometric work. Questions to do with the derivation of the models and the kind of explanation provided by them.

In econometrics we start by looking at a particular phenomenon, then we build a model which tries to explain how that phenomenon occurs. Having built the model and tested it with the data, we look at the results and ask whether they are sensible—a very practical test—whether the coefficients are of the right kind of magnitude; whether they have the right sign or whether, to use the econometrician's language, they have a "perverse" sign— a sign opposite to that expected.

May I give an example of this by referring to the results in Table 4. Table 4B gives the formula $1{\cdot}11X_{11}-0{\cdot}15X_7$, for predicting 16-year-old reading ability score in terms of the 11-year-old reading score and the 7-year-old reading score. Why should we take the 11-year-old score and subtract from it some multiple of the 7-year-old score? I can understand that equation because I can rewrite it slightly as (approximately) $0{\cdot}96X_{11}+0{\cdot}15(X_{11}-X_7)$; in other words, it is the 11-year-old score plus a multiple of the change between 7 and 11. That seems to be sensible.

I would like to ask Professor Goldstein why, if that is the kind of underlying process, do we not start off with an equation of that sort, and estimate the coefficient on $X_{11}$ and the coefficient on the difference variable $(X_{11}-X_7)$?

Incidentally, if we try to rewrite the similar equation for predicting mathematical score, the end-result is a rather peculiar equation, perhaps with a perverse sign—or else it has some rather peculiar implications for the way we learn or dislearn mathematics.

My second question is, again, the usual one asked of econometricians. Professor Goldstein makes the sweeping assumptions that the models are additive; and the heroic assumptions, usually made for identifiability, that coefficients like $\gamma_{21}$ or $\beta_{21}$ are zero. Are these assumptions really justifiable, particularly when one considers the explanations which should be in the models?

Mr BRYAN RODGERS (National Survey of Health and Development): I am pleased that Professor Goldstein has raised the question of the use of different types of tests administered at different ages. I wish to present some data from the National Survey of Health and Development (the 1946 birth cohort which was mentioned in Section 8 of the main paper) which illustrate the effects of using different types of tests at different ages, and the degree to which the use of different tests can influence the interpretation of the data.

The National Survey of Health and Development used a word pronunciation test at 8 and 11 years old—Professor Goldstein calls this a word recognition test. The obtained data have been used to carry out a similar regression analysis to the type discussed today, but I stress that no correction was made for the unreliability of the test.

If we carry out the analysis using this 8- and 11-year old data by regressing the scores at 11 on those at 8 for two different groups (I will use only *two* groups, the children of non-manual fathers and those of manual fathers) we find that for a given 8-year-old score, the children of non-manual fathers are 0·22 of a standard deviation ahead of those children of manual fathers, by the age of 11.

The problem then arises as to whether the same results would be obtained if a different test were used at 11. We can go some way to answering this question using the data which the National Survey has collected. A number of other tests were given to the survey children when they were 8 years old, one of which was a vocabulary test and used the same items as the pronunciation test—that is, the children were asked to supply the meaning of the words which they had previously attempted to read.

If we regress the vocabulary scores on the pronunciation scores for the same age (8 years old) this produces what I consider a rather remarkable result, that for a given pronunciation score non-manual children perform 0·32 of a standard deviation ahead of the manual children on the vocabulary test. This means that we can apparently "achieve" a larger divergence of the two social groups by administering a different test on the following day than can be "achieved" by administering the same test three years later.

Similar results, but of lesser magnitude, are obtained if we look at the other tests administered at 8 years old. If we take manual and non-manual children of equivalent word pronunciation scores, we find that the non-manual children are performing ahead of the manual children by 0·13 of a standard deviation on a sentence completion test, and by 0·18 of a standard deviation on a picture intelligence test.

This raises a problem: how are we to interpret an analysis where we equate on a single test given at age 8—or age 7 in the case of the National Child Development Study? I have asked what would happen if we equated the manual and non-manual children, not just on the word pronunciation test but on all four tests so that we could say, for a given score on the word pronunciation test *and* sentence completion test *and* picture intelligence test *and* vocabulary test, what differences do we then find between the two groups at age 11? We find if we do that, that the initial observed difference of 0·22 of a standard deviation for word pronunciation is reduced to 0·16 of a standard deviation.

What I am worried about is that if we can reduce what was called initially a change between Time 1 and Time 2 by attributing some of that change to what are, in fact, observed differences existing at Time 1, how much further can we carry on this process by introducing an increasing number of independent variables, or covariates, from Time 1? It appears to me that we cannot answer this question, using the type of data we have. However, if we could build into the model an increased number of pre-existing differences between the social groups at Time 1, the observed change—or so called change—between Time 1 and Time 2 may be substantially reduced.

The following contributions were received in writing after the meeting.

Mr JOHN BIBBY (The Open University): In common with other discussants I should like to congratulate Professor Goldstein on his careful investigation of the impact of measurement error on regression estimates. However, I confess to feeling somewhat perplexed when taking account of such errors changes a plausible positive estimate (0·136 in Table 2B) into a "perverse" negative estimate (−0·147 in Table 4B). Does this mean that my prior assessment of plausibility is wrong, or simply that it is not compatible with the existence of measurement error (at least not with *this model* of measurement error)?

I should also like to raise some issues in connection with the social class variable, and in particular the assumption (Section 7) that no error attaches to the measurement of this variable. To assess the reliability of this assumption one would need detailed information concerning the measuring instrument used, which I do not have at my disposal. Perhaps Professor Goldstein could elaborate on the wordings of the relevant questions, the reliability of the respondents to whom the questions were addressed, the coding procedures employed and the class-perceptions of the actors associated with various stages of the measurement process, especially in so far as these considerations might throw some light on the likely structure of measurement error in the final data or perhaps this is an unfair question.

One reason why I suspect that considerable measurement error might exist in the social class variable is the behaviour of the percentage of the sample who were allocated to the "Unskilled manual" category. This varies from 6·3 per cent at age 7, through 5·7 per cent at age 11, to 5·3 per cent at age 16. These figures may be compared with the 1971 census figure of 8·6 per cent of economically active males in Registrar General's category V. Even allowing for changes over time and for the considerable age-variation exhibited by the Census data, it seems that Professor Goldsteins' sample may be under-representing the Unskilled manual class by as much as 40 per cent. I should be interested to hear his comments on this suggestion.

The question also arises of what is the effect of measurement error in a categorical variable such as social class on any regression estimates. As in Professor Goldstein's equation (6C) we have an equation of the form

$$X_i = x_i + u_i.$$

However, if the $x$s represent class measurements then the $u$s will usually *not* have expectation zero. If there are just two classes, coded 0 and 1, then it may be reasonable to assume that

$$E(u_i|x_i = 0) = p_{10} \quad \text{and} \quad E(u_i|x_i = 1) = -p_{01},$$

where $p_{10}$ is the probability that $X_i = 1$ when $x_i = 0$, and $p_{01}$ is defined similarly.

If $Y_i = \alpha + \beta x_i + \varepsilon_i$ then with the usual assumptions the probability limit of the least squares estimator $\hat{\beta}$ is

$$\frac{\beta \text{var}(x) + \beta \text{covar}(x,u)}{\text{var}(x) + 2\ \text{covar}(x,u) + \text{var}(u)}$$

$$= \frac{\beta\pi_1\pi_0 - \beta\pi_1\pi_0(p_{10}+p_{01})}{\pi_1\pi_0 - 2\pi_1\pi_0(p_{10}+p_{01}) + \pi_0 p_{10} + \pi_1 p_{01} - (\pi_0 p_{10} - \pi_1 p_{01})^2}$$

where $\pi_0$ and $\pi_1$ are the probabilities of the events $x_i = 0$ and $x_i = 1$ respectively. In the special case where $p_{10} = p_{01} = p$ and ignoring quadratic terms in $p$, this simplifies to

$$\frac{\beta\pi_1\pi_0(1-2p)}{\pi_1\pi_0(1-4p)+p}$$

For instance, we might characterize the distinction between manual and non-manual workers in Professor Goldstein's sample by taking $\pi_0 = 0{\cdot}4$ and $\pi_1 = 0{\cdot}6$. The above expression then simplifies to

$$\beta(1-2p).$$

In other words, with these assumptions the estimator $\hat{\beta}$ is inconsistent and asymptotically downwardly biased. The asymptotic shrinkage is represented by the factor $(1-2p)$. For instance, if 5 per cent of respondents were misclassified (not an unlikely figure, I suspect) then $\hat{\beta}$ would tend to be shrunk by around 10 per cent.

Clearly many questions have been left unanswered by the above, but I would like to ask Professor Goldstein whether he has investigated these or related issues.

My final point concerns the impact of measurement error upon social mobility tables such as Table 7. In general, even cells which are "really" empty in these tables will tend to have entries in them if there is measurement error. This may be what is happening in Table 7. Of the 27 cells in this table *less than 1 per cent of the sample lies in the 12 cells indicating mobility over more than one class-boundary*. This may be contrasted with the fact that 75 per cent of the sample lies in the three cells which exhibit zero stability. Professor Goldstein's summary of this table as exhibiting "considerable social class mobility" thus strikes me as rather bizarre—unless the definition of 'considerable" is by comparison with some even more extreme reference point. However, my main point is that *even the 1 per cent referred to above may be no more than a reflection of measurement error in the allocation of social class categories.*

A short back-of-the-envelope calculation suggests that even if there were absolutely no "real" mobility, the proportion of the sample exhibiting mobility over more than one class boundary would be approximately $3p(\pi_1+\pi_3)$, where $p$ is the misclassification error and $\pi_i$ is the proportion of the population in category $i$. In Professor Goldstein's example we may take

$$\pi_1+\pi_3 = 0{\cdot}4.$$

This suggests that measurement error of only $p = 0{\cdot}0077$ is sufficient to generate the observed proportion of mobility over more than one class boundary. I would be interested to hear Professor Goldstein's comments on this point, and anything else he might like to say on the problems of measurement error in categorical variables.

Mr CLIFFORD SPIEGELMAN (Statistical Engineering Laboratory, N.B.S., Washington, D.C.): It should be noted that the assumption of normality for $u_i$ in equation (4) plays a crucial role. It was shown by Reiersöl (1950) that consistent estimates of slope may exist without resorting to reliability estimates if $u_i$ is not normal. In a survey paper by Moran (1971) several such estimates are referenced. A new estimate of this sort is given in Spiegelman (1979).

If the $u_i$ are normal, changes due to social class changes may also be analysed without resorting to reliability estimates. For pedagogical purposes suppose there are two social classes $a$ and $b$, and all students we are analysing start in $a$.

$$\text{Let} \quad z = \begin{cases} 1 & \text{if } a \to b. \\ 0 & \text{if } a \to a \end{cases}$$

Let $x_1$ be the 7 year score satisfying equation (4) and the 11 year score $x_2 = \theta_0 z + p_k(u) + \varepsilon_2$, where $\theta_0$ is the fixed additive effect of class change and $p_x(u)$ is a polynomil of degree $k$ in $u$ with unknown coefficients. Then if $[Ez|x]$ (the average proportion changing $a \to b$ for each $x$ value) is not too highly multicolinear with 1, $x, \ldots, x^k$, and $E[z|x]$ were known, standard regression tactics could be applied to estimate $\theta_0$.

However, $E[z|x]$ is not known. Still $E[z|x]$ is an identifiable function. Universally consistent procedures for estimating $E[z|x]$ may be found in Stone (1977). These estimates of $E[z|x]$ may be

used in place of the true function in the regression. The exact conditions for and the asymptotic distribution of the resulting estimates (which are asymptotically unbiased) may be found in Spiegelman (1976).

The AUTHOR replied later, in writing, as follows.

I would like to thank all the discussants for their useful comments. In his vote of thanks, Professor Aitkin stressed the importance of obtaining good measurement error variance estimates and, if possible, using ML methods. As well as this, I would emphasize the importance of investigations into the validity both of the usual formulae for obtaining reliability or measurement error variances and of the normality assumptions in the ML approach, which are also commented upon by Mr Spiegelman. Mr Russell Ecob is presently carrying out analyses along these lines using the NCDS data. Professor Aitkin is also right to draw attention to problems of interpreting covariance adjusted estimates in non-randomised studies, a point also raised by Mr Hutchison. I think it is important, however, to draw a distinction between two kinds of non-randomised studies. One is where different treatments are applied to groups who differ in their initial characteristics but where one would like to make the kinds of inference normally associated with randomised studies. These are often known as quasi-experimental designs. The other type of study is a purely observational study, such as the NCDS, where the emphasis lies in a description of change. In the former kind of study, the purpose of covariance adjustment is to attempt to equate the initial characteristics of the groups and this leads to familiar difficulties, some of which Professor Aitkin has mentioned. In the present paper, I have been concerned with developing a measure of change based on future expectations of attainment, from a given attainment level, so that the purpose of the covariance analysis is different. Thus, when I talk of differences between groups not changing over time, I am referring to the expected test score at the later occasion of children from the different groups who have the same attainment at the earlier occasion, and I do not see that there is any ambiguity in this. On the question of interactions, also raised by Mr King, I agree that it is not entirely satisfactory to ignore them. Nevertheless, the inclusion of the $7 \times 11$ year score interaction has a very small effect on the other estimates, including those for social class, and it seems justifiable in the interests of parsimony to omit it. In response to Professor Aitkin's final point, I should, perhaps, have made it clearer that I was referring to the difference between a conditional (ANCOVA) model and an unconditional model where the simple difference between initial and final score is used as the dependent variable. I was pointing out that a change of scale to give $\beta = 1$ only made the final equation in Section 3 *appear* to refer to an analysis of simple differences, whereas the initial score still retained the status of a variable with fixed values.

I am extremely grateful to Dr Adelstein for raising some wider issues. I very much agree that educational measurements are often used uncritically without their limitations being made clear. A number of discussants mentioned this. On the one hand, there is a proliferation of educational tests and, on the other hand, there is a body of statistical theory about test construction and use, but there seems to be little empirical work with large data sets which is able to evaluate the test material against the statistical theory. As Dr Adelstein says, the present paper is only a start. I also think he is correct to question the cost of large-scale longitudinal studies. My own view is that there was a good justification for the three studies he mentions, but that future large-scale child development studies should be designed differently. There may no longer be need to follow up a large cohort from birth, and the results of the existing studies could be used to guide us in choosing an efficient design for future studies, which would presumably involve follow-up of several groups of children with overlapping age ranges.

Mr Plewis and Mr Ecob discuss the use of teachers' ratings of attainment. Mr Plewis's extension of equations (11a) and (11b) is an interesting one. While it is true that the parameters are now identifiable, this is achieved by assuming that certain pathways between ratings and test scores are absent, and I wonder how reasonable such an assumption might be?

In reply to Mr Ecob, the teachers' rating of reading at 7 years was used to estimate the reliability. The question of raising the precision of instrumental variable estimates of reliability at the increasing risk of violating the model assumptions is an interesting one, which I hope he will pursue.

Dr Hansford-Miller raises the important question, only briefly touched upon in the paper, of the relevance and importance of the variables used. I can only reiterate that I do consider this an important question, but that my paper was explicitly methodological and not directly

concerned with such issues. There are a number of papers based on NCDS data, however, which do address such issues and I would refer him to Fogelman and Goldstein (1976) and to Davie, Butler and Goldstein (1972).

In reply to Dr King, I have deliberately avoided writing equations involving simple difference scores such as $x_{11}-x_7$ for the reasons explained in Section 3. Mr King and Mr Bibby raise a query about the estimates in Table 4b. I would interpret the coefficient for 7-year-old reading in terms of the greater progress made by a low scoring child between 7 and 11 years, compared to a high scoring child where both children have the same 11 year attainment. The interesting educational question is why this should be so, and why a different result occurs for mathematics. It is not clear to me that either of these two results is obviously unreasonable. Likewise, the differing results for equations (11a) and (11b) where different assumptions are made require further study and my intention was not to make "heroic assumptions" about parameter values, but to study the effects of different assumptions.

Mr Rodgers raises the question of how to interpret, at age 8, differences in vocabulary scores between children from non-manual and manual social class groups who are equated for pronunciation test scores. While that is undoubtedly an interesting problem, it seems not to help us very much in answering the question why the same children differ on the pronunciation test given 3 years later. Mr Rodgers is worried also by the increasing ability to explain social class (or other) differences using more covariates. I would take the view rather, that one of the aims of a study is to find variables which do account for such differences, at least in part, and I comment on this towards the end of Section 10.

Mr Bibby, Mr Plewis and Mr Hutchison raise questions about errors of measurement of social class. In reply to Mr Bibby, these data were obtained by interviewing the child's mother and asking her for details of her husband's occupation. The subsequent coding was carried out by social class coders at the Office of Population Censuses and Surveys. As far as I know, there has been no study of likely sources of measurement error in social class coding for the NCDS data and I agree that it would be useful if this were done. Comparisons of social class distributions with census data are difficult since the appropriate statistic is the percentage of 16-year-old children whose fathers are in the different classes rather than the percentage of economically active males in the different classes. The figure for 13-year-olds for Great Britain from the 1971 census (when the NCDS children were 13) is 7·4 per cent (unpublished special tabulation) which is still rather higher than the (interpolated) NCDS figure of 5·6 per cent and I have no convincing explanation for this. The simple model for measurement error in categorical variables suggested by Mr Bibby can be incorporated within the approach of the paper in a straightforward way. The difficulty lies in estimating the various misclassification probabilities and I would be interested to know of work which bears on that. One of the reasons for choosing the particular three social class groupings was to minimize coding and reporting errors, so that a figure of 5 per cent misclassifications seems to be rather on the high side, and as far as Mr Bibby's penultimate paragraph is concerned, I still think that Table 7 indicates a high social class mobility. In Mr Bibby's final paragraph, the misclassification probability refers to misclassifying a child from a non-manual background as one from an unskilled manual background and vice versa. These seem highly unlikely events and the value of $p$ quoted by Mr Bibby is likely to be much in excess of the true value.

## References in the Discussion


Bruner, J. S. (1964). The course of cognitive growth. *Amer. Psychol.*, **19**, 1–15.

——(1966). On cognitive growth. In *Studies in Cognitive Growth* (J. S. Bruner, R. R. Oliver and P. M. Greenfield, eds). New York: Wiley.

Choat, E. (1978). *Children's Acquisition of Mathematics.* Windsor, England: NFER Publishing Co.

Cronbach, L. J., Rogosa, D. R., Floden, R. E. and Price, G. G. (1977). Analysis of covariance in non-randomized experiments: parameters affecting bias. Mimeograph. Stanford Evaluation Consortium, School of Education, Stanford University, California 94305,U.S.A.

Davie, R., Butler, N. R. and Goldstein, H. (1972). *From Birth to Seven.* London: Longman.

Hutchison, D., Prosser, H. and Wedge, P. (1979). The prediction of educational failure. *Educ. Studies*, **5**, 73–82.

Joreskog, K. G. and Sorbom, D. (1976b). Statistical models and methods for test–retest situations. In *Advances in Psychological and Educational Measurement* (D. N. M. De Gruijter, L. J. Van Kempand Th. Van Kemp, eds). London: Wiley.

Moran, P. A. P. (1971). Estimating structural and functional relationships. *J. Mulitivar. Anal.*, **1**, 232–255.

OSBORNE, A. F. and MORRIS, A. C. (1979). The rationale for a composite index of social class and its evaluation. *Brit. J. Sociol.*, **30**, 39–59.
REIERSÖL, O. (1950). Identifiability of a linear relationship between two variables which are subject to error. *Econometrica*, **18**, 375–389.
STONE, C. (1977). Consistent nonparametric regression (with Discussion).
595–645.
SPIEGELMAN, C. (1976).
Ph.D. Thesis, Northwestern University, Illinois.
——(1979). On estimating the slope of a straight line when both variables are subject to error. *Ann. Statist.*, **7**, 210–206.
VYGOTSKY, L. S. (1962). *Thought and Language.* Cambridge, Mass.: M.I.T. Press.

As a result of the ballot held during the meeting, the following were elected Fellows of the Society

BARKER, Michael J.
BARKER, Thomas J.
BAWDEN, Derek
BLUMENFELD, Dennis E.
BURKE, Mrs Christine A.
CHADHA, Harbajan K.
COPE, David R.
DALENIUS, Professor Tore E.
DE SARBO, Wayne S.
ELBOURNE, Diana R.
ERICSSON, Neil R.
FAGBONGBE, Julius O.
GEDALLA, Brian
GLASBEY, Christopher A.
GOWERS, James I.
GUPTA, Hari D.
HARRIS, Raymond I.
HASTIE, Trevor J.
HIGGINS, Bernard R.
HUXHAM, Samuel H.
KHAN, Masood A. K.
KIMBER, Alan C.
MACDONALD DAVIES, Mrs Isobel M.
MORRISSEY, Susan M.
MUSTAFA, Rawda A. A.
PHIPPS, Diarmid F.
PIERCE, Jennifer M.
RAJAGOPAL, Pinayur
RAMYAR, Hooshang
RAWLES, Richard E.
ROSS, George D.
SAND, Peter W.
SAUNDERS, Ian W.
SOUZA, Reinaldo C.
THOMPSON, Valerie
THOMSON, David J.
TRUE, John
TRUELSEN, Peter E.
WANRODY, Gerard L.
WEBSTER, Susan
WEHRLY, Thomas E.
WELDON, Kenneth L.
WIGODSKY, Peter A.
YEE, Kan-Fat
YOUNUS, Muhammed