Journal of Educational Statistics
Fall 1983, *Volume* 8, *Number* 3, *pp. 223–241*

INSTRUMENTAL VARIABLE METHODS FOR THE ESTIMATION OF TEST SCORE RELIABILITY

RUSSELL ECOB
Inner London Education Authority
AND
HARVEY GOLDSTEIN
University of London Institute of Education

Key words: *Instrumental Variables*, *Errors in Variables*, *Reliability*, *Longitudinal*, *Educational Attainment*

ABSTRACT. The method of Instrumental Variables is suggested as an alternative to traditional methods for estimating the reliability of mental test scores which avoids certain drawbacks of these methods. The consistency and efficiency of the instrumental variable method are examined empirically using data from the British National Child Development Study in an analysis of 16 year, 11 year and 7 year old scores on tests of mathematics and reading.

In the simple regression model (with the usual assumptions about the error term u)

$$y = \alpha + \beta x + u,$$

it is well known (Goldstein, 1979) that if the observed independent variable x contains errors of measurement and we wish to estimate the regression coefficient on the "true" value of x, then the ordinary least squares (OLS) estimator is inconsistent. The simplest and most common model relating the true value to the observed value of x is

$$x = T + e, \tag{1}$$

where T is the true value, e the random error of measurement, and $\mathrm{Cov}(T, e) = 0$. It is supposed, therefore, that we wish to estimate the parameters α, β in

$$\begin{aligned} y &= \alpha + \beta T + \varepsilon \\ &= \alpha + \beta x + (\varepsilon - \beta e). \end{aligned}$$

It is because x is correlated with $(\varepsilon - \beta e)$ that the OLS estimator, b, is an inconsistent estimator of β. A consistent estimator is given by b/R, where R is the reliability of x and defined as

$$R = \mathrm{Var}(T)/\mathrm{Var}(x),$$

where $\mathrm{Var}(x) = \mathrm{Var}(T) + \mathrm{Var}(e)$.

In many situations, the value of R is very close to 1, and any adjustment to the usual estimate can be safely ignored. In other applications, for example in mental testing, R may be considerably less than 1 so that an adjustment becomes necessary. In a linear model with several further independent variables, the estimators of the associated regression coefficients will also be inconsistent if OLS is used. Consistent estimates may be obtained by adjusting the observed covariance matrix of the independent variables so that the observed variances corresponding to variables containing measurement error have estimates of their measurement error variance subtracted prior to inversion of the matrix and calculation of the coefficients (Warren, White, & Fuller, 1974). To do this, it is important to have accurate and consistent estimates of the measurement error variances, or alternatively, reliabilities. In this paper we explore some new procedures for obtaining such estimates based on instrumental variable techniques.

Reliability Estimation

A large body of literature exists concerning procedures for obtaining good estimates of reliability for mental tests that consist of a fixed set of response items. Details can be found in Lord & Novick (1968, Chapter 4). The methods that have been proposed can be divided usefully into "internal" and "external" methods.

Internal Estimates of Reliability

In the internal methods, various relationships between the test item responses are used to provide reliability estimates, of which the best known is "coefficient alpha." Under general assumptions about the conditional independence of item responses, and constancy of measurement error variances, a useful lower bound on the reliability can be estimated, which under certain further assumptions leads to a consistent estimate of the reliability. We write, similarly to (1),

$$x_{ji} = T_{ji} + e_{ji},$$

where x_{ji} is the observed value of item j for individual i, τ_{ji} is the true value of item j for individual i, and e_{ji} is the measurement error. A conditional independence assumption is usually made; that is, for given τ_{ji}, $\tau_{j'i}$, $j \neq j'$

$$\mathrm{Cov}(e_{j'i}e_{ji}) = 0. \tag{2}$$

The conditions for a consistent estimator of the reliability are

$$\tau_{ji} = \alpha_j + \tau_i, \tag{3}$$

where α_j and τ_i can be interpreted as the difficulty of item j and the ability of individual i. This implies that the items are not only "unidimensional" but essentially equivalent apart from a location shift. If (2) does not hold, then

coefficient alpha is not necessarily a lower bound and may even overestimate the true reliability.

We see, therefore, that there are two serious drawbacks to 'internal" estimates of reliability. First, there is the difficulty of satisfying (3) to obtain a consistent estimator of the reliability, or equivalently, of the measurement error variance. While there is a considerable literature on this and the more general concept of unidimensionality, in practice it seems extremely difficult to ensure that item responses are indeed determined by a single quantity for each individual such as given by (3). For the types of educational tests we deal with in this paper, it seems even less likely that a unidimensional trait is operating. A more detailed discussion of this topic is given by Goldstein (1980). Secondly, assumption (2), often known as the "local independence" assumption, seems somewhat unreasonable a priori. It is difficult to imagine that if a given individual fails one item then the probabilities of success on later items are the same as when the individual succeeds on the earlier item. Nevertheless, there seems to have been little, if any, serious study of this problem and the consequent effect of nonzero correlations on reliability estimates. A further discussion of this point in this context of latent trait models is given by Goldstein (1980). Thus, there is as yet no really satisfactory method for obtaining a consistent estimate of reliability using "internal" methods, nor of even providing a lower bound. We suggest that estimates based on these methods should be treated with some caution.

External Estimates of Reliability

The most obvious method of estimating reliability or measurement error variance is to carry out repeat measurements. Thus, we have (dropping the suffix i), for two applications of a test,

$$X_1 = T + e_1;$$
$$X_2 = T + e_2;$$

and $\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(e_1 - e_2) = 2\{\sigma_e^2 - \mathrm{Cov}(e_1, e_2)\}$, where σ_e^2 denotes the constant measurement error variance. For many physical measurements it is reasonable to assume independence errors, that is, $\mathrm{Cov}(e_1, e_2) = 0$, so that we have $\sigma_e^2 = \frac{1}{2}\,\mathrm{Var}(X_1 - X_2)$. For mental tests, however, this usually will not be a reasonable assumption due to the presence of memory effects, learning, and so on. If more than one test relating to the same attribute is available, then by assuming suitable relationships between the true scores on the tests, it is possible to obtain reliability estimates. The usual assumption is that the tests are congeneric so that for a set of p tests,

$$X_j = a_j + b_j T + e_j, \qquad j = 1,\ldots,p.$$

The observed covariance matrix of the X_j contains $\frac{1}{2}p(p+1)$ elements and if we assume $\mathrm{Cov}_{j \neq j'}(e_j e_{j'}) = 0$ and $\mathrm{Cov}(T_j e_j) = 0$, the matrix is a function of the b_j and error variances σ^2_{ej}, which gives $2p$ parameters. Hence, for three or more tests, unique estimates, based for example on maximum likelihood, are available. Details of this approach are given in Jöreskog (1971). Although it is not quite as serious as in the simple test-retest case, this method also has the difficulty that the measurement errors of the tests may be correlated, for example, because of day to day fluctuations among examinees, and the like. This immediately raises the question of definition of true score, but we shall postpone a discussion of that until a later section.

In the next section we propose a generalization of congeneric tests to include any variable having nonzero correlation with the test whose reliability we wish to measure. Such an "instrumental variable" does not require any assumption about unidimensionality or independence; also, unlike the simple test-retest or the congeneric test models, the possibility of choosing any variable means that we can search for those that are likely to be uncorrelated with the measurement error e_1. The possibility of dropping both these restrictive assumptions is attractive and the remainder of the paper investigates this problem using an extensive longitudinal data set.

The Data

The data come from the National Child Development Study (NCDS) that 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 (1976).

Testing and scoring in the NCDS was carried out by the class teacher who also noted the child's achievements. Because the study was a national study of all children born in a particular week, most children selected were tested in a different situation and by a different tester.

Four possible situations giving rise to response variation are as follows:

1. the environment in which the test is administered;
2. the process of test administration;
3. the coding and scoring of the test (this includes the interpretation of the correctness of the response); and
4. day-to-day variation in individual test performance.

Because only one test of a given type was done by each child at each occasion, the sources of variation 1–4 above are confounded. It is important, however, to distinguish "day-to-day" variation from changes in true score over time. We can regard variation over time as contributing either to measurement error or to true score variation or to both. A reasonable estimate of the true score at a

particular moment would be obtained from a moving average of scores taken at successive time intervals before and after. The continuous change in true test score over, say, a week is therefore regarded as being supplemented by random error to produce the observed day-to-day variation. The various educational measurements in the NCDS were completed within a week for each child, so that any true score changes over not more than a 1-week period are effectively regarded as part of day-to-day variation. In addition to these sources of measurement error, there will typically remain an unexplained variation that can be conceptualized as the variation between the response to an item and its hypothetical replication.

Cronbach, Gleser, Nanda, & Rajartnan (1972) argue that test evaluation or "generalizability" studies, which also view a particular test as a sample from a universe of tests and which use experimental designs to estimate individually the above components of variation, should be carried out prior to test administration. We use here a "test-specific" interpretation of true score that treats true score as relevant only to the particular test. A justification for this is given by Goldstein (1979), although the methods used in this paper can be extended to a full "generalizability" approach. Goldstein (1979), using the same NCDS data, also drew attention to the use of instrumental variables in estimating the relation between mathematics and reading attainments when measured at different ages. He emphasized the potential usefulness of this method when imprecise prior knowledge about the reliability of the earlier attainment scores is available, and pointed out that little was known about the degree to which the instrumental variables used satisfied the conditions of consistency.

In this paper, the properties of a variety of instrumental variables are examined in the context of the regression of 16 years attainment on 11 and 7 years attainment for mathematics and reading test scores separately. Comparisons are made with the use of ordinary least squares and also with the use of the internal estimates of the reliability coefficients for the 11-year attainment given in Goldstein (1979).

Theory of Instrumental Variables Estimation

General Theory

Suppose X_{1i}, X_{2i} are the observed values of test score variables measured as deviations from their means at the first and second occasions and let them be the predictor and dependent variables, respectively, in a simple linear regression model. Furthermore, denote their true values as T_{1i} and T_{21} and the errors of observation or measurement errors for the ith subject as e_{1i} and e_{2i} ($i = 1,\ldots,n$). Then, as before:

$$X_{1i} = T_{1i} + e_{1i};$$
$$X_{2i} = T_{2i} + e_{2i}, \tag{4}$$

and a model relating the true values at each occasion is

$$T_{2i} = \beta T_{1i} + u_i. \tag{5}$$

If Z_i is the observed value of another variable, called the instrumental variable, then

$$b_{\text{IV}} = \sum_{i=1}^{n} Z_i X_{2i} \left(\sum_{i=1}^{n} Z_i X_{1i} \right)^{-1} \tag{6}$$

is called the instrumental variable estimator of the regression coefficient β (Johnston, 1972, p. 279).

From (4), (5) we have

$$b_{\text{IV}} = \left(\beta \sum Z_i T_{1i} + \sum Z_i u_i + \sum Z_i e_{2i}\right)\left(\sum Z_i X_{1i}\right)^{-1},$$

and $b_{\text{IV}} - \beta = (\Sigma Z_i u_i + \Sigma Z_i e_{2i} - \beta \Sigma Z_i e_{1i})(\Sigma Z_i X_{1i})^{-1}$. In terms of the sample correlations, r_{Zu}, r_{Ze_2} and r_{Ze_1} between the instrumental variable Z and the disturbance and errors of measurement of X_2, X_1, respectively, and the reliability R of X_1,

$$b_{\text{IV}} - \beta = \left(\frac{\sigma_u}{\sigma_{e_1}} r_{Zu} + \frac{\sigma_{e_2}}{\sigma_{e_1}} r_{Ze_2} - \beta r_{Ze_1} \right) \frac{(1 - R)^{1/2}}{r_{ZX_1}}, \tag{7}$$

where σ_{e_1}, σ_{e_2}, σ_u are, respectively, the standard deviations of the errors on the 1st occasion, 2nd occasion, and the disturbance term. As the sample size tends to infinity, the following consistency condition is obtained:

$$\sigma_u \rho_{Zu} + \sigma_{e_2} \rho_{Ze_2} - \beta \sigma_{e_1} \rho_{Ze_1} = 0.$$

Thus, a general condition for consistency is that the three correlations ρ_{Zu}, ρ_{Ze_2}, ρ_{Ze_1} are all zero. Equation (6) shows that if the predictor and dependent variables are interchanged, the instrumental variable estimator becomes its reciprocal.

The efficiency of the instrumental variable estimator is the square of the multiple correlation of the instrumental variable set with the predictor. This assumes that the estimator is consistent.

The Use of Many Instrumental Variables

When we have p instrumental variables $Z_j, j = 1, \ldots, p$,

$$b_{\text{IV}} = \left(\sum_i \sum_j c_j Z_{ij} X_{2i} \right) \Big/ \sum_i \sum_j c_j Z_{ij} X_{1i}. \tag{8}$$

The combination of Z_j that gives the most efficient estimate of b_{IV} can be found by choosing the constant c_j so that $\text{Corr}(\Sigma_i c_j Z_{ij}, X_{1i})$ is a maximum. This is the usual two-stage, least squares estimator (Johnston, 1972, p. 380). The c_j are then the sample regression coefficients, b_j, of X_1 on $Z_j, j = 1, \ldots, p$. Letting $\hat{X}_{1i} = \Sigma_j b_j Z_{ij}$, we obtain $b_{\text{IV}} = (\Sigma X_{1i} X_{2i}) / \Sigma X_{1i} X_{1i}$.

The Use of Dummy Variables as Instrumental Variables

The previous discussion has assumed the existence of instrumental variables that can be modelled as having simple linear relationships with the first occasion variables. Two other cases can be distinguished. First, where an interval scaled instrumental variable has a nonlinear relationship to the first occasion variable, and second, when the instrumental variable is categoric (e.g., measured on an ordinal or nominal scale).

In the first case the nonlinear relationship can be modelled, say by a polynomial function, or the instrumental variable can be grouped into categories. In the latter case, each category can be represented in the usual way by a dummy variable. This takes the value 1 for this category and 0 for every other category. Let X_{1r_k}, X_{1r_k} be two observations on the first occasion variable belonging to the same instrumental variable category, r. Using the dummy instrumental variables to estimate the first occasion variable gives the estimate $\bar{x}_{1r}$, which is the mean value of all observations in category r.

Substituting in (8) gives

$$b_{\mathrm{IV}} = \left(\sum_r p_r \bar{X}_{2r} \bar{X}_{1r} \right) \left(\sum_r p r \bar{X}_{1r}^2 \right)^{-2},$$

where $\bar{X}_{2r}$ is the mean of the X_{2r_k} in the category r, and p_r is the proportion in category r. This is essentially the "Method of grouping" as introduced by Wald (1940).

The literature on grouping methods (e.g., see Madansky, 1959; Neyman & Scott, 1951; Wald, 1940) using observed values of X_1 has tended to focus on conditions for consistent estimates and on the relative efficiency of different groupings rather than quantifying the inconsistency of various grouping methods. The results on the NCDS data, given below, go some way to remedying this situation for a particular data set.

Application of Instrumental Variable Methods to the NCDS Data

Selection of Variables

In all, 50 variables are considered as instrumental variables, being measured at ages 7, 11 and 16. These consist of test scores, teacher ratings and background variables. The test scores are of reading and mathematics at each age, and in addition, of general ability and copying designs scores at age 11. The teacher ratings are of reading and mathematics at all ages, and in addition, of oral ability and creativity at age 7, of oral ability and general knowledge at age 11, and of practical subjects at age 16. The "background" variables are social class and indices of behavior in the home at all three ages; the number of children in the household, birth order of the child, an index of accommodation facilities and childrens' heights at ages 11 and 16, overcrowding at 11; and

region, indices of school behavior, and a variety of feelings toward school at 16. The reader is referred to Davie, Butler and Goldstein (1972) and Fogelman (1976) for a more complete description of these variables.

The variables used as dependent variables in the regressions, that is, the test scores at age 11 and 16 of mathematics and reading, are transformed to have standard normal distributions in the same way as in Goldstein (1979) who showed that near linear relationships between observed scores resulted.

As the relative efficiency of an instrumental variable estimator is proportional to the square of the correlation with the predictor, variables are only retained for further analyses when this correlation is greater than 0.3. This eliminates all the "background" variables save social class, but only one of the teacher ratings (that of outstanding ability in any area at age 11) and none of the test scores, leaving 25 variables in all. All cases with missing values on any of these 25 variables are excluded, leaving 5,371 cases with test scores at each age.

In the appendix to Fogelman (1976), Goldstein shows that the attrition of subjects in the study does not affect to any marked extent the relationships found and that test scores for subjects having missing values on the background variables show no significant differences from other subjects.

In the results reported here, all instrumental variables are treated as sets of dummy variables for reasons given earlier in that section. For the test scores the dummy variable coding into five roughly equal size groups ensures that the relative loss in predictive efficiency from dummy variables compared to simple linear regression is always less than 6 percent.

In fact, using these instrumental variables as dummy variables or as interval scale variables gives very similar results, with the maximum difference in regression coefficients for any instrumental variable being 2 percent.

Forming Hypotheses of Error Structure in Prediction Relations

As in the section on External Estimates of Reliability we assume that the true score of a test comprises that component of test score that is unaffected by day-to-day variation by the particular testor or by the test situation, and is specific to the particular test used. We now examine the correlation of the variables to be considered as instrumental variables with measurement errors on the first occasion and second occasion tests and with the disturbance term. These variables are teacher ratings on a variety of attainments, test scores and social class. The teacher ratings, like the test scores, generally will contain measurement error, thus reflecting and being reflected by variations in the child's interest in subjects and day-to-day variations in the type of relationship to the teacher. Thus, a teacher who has very recently seen a child do a good piece of work or show a keen interest may tend to rate him higher than

otherwise. If the same contributory factors affect test score, then a teacher rating made at the same time as the test would be expected to have a positive correlation with the test score measurement error. Likewise, where a different attainement is rated at the same time as the test, similar correlations may exist, although presumably they would be smaller.

There are two variables that we hypothesize will have a zero correlation with test score measurement error. These are teachers' ratings taken at a different point in time and social class. We would expect none of the sources of measurement error to relate to teacher rating at a point in time 4 or 5 years away. Nor would we expect social class, which shows little variation for an individual over short time periods, to relate to any of the sources of measurement error.

Finally, we examine the relation of the disturbance terms to teacher ratings and social class. Either of these variables, particularly social class, may be correlated with the disturbances if they relate to the dependent variable once the predictor variable has been controlled for.

The hypotheses formulated above may be summarized as follows:

H1. Teacher ratings on a test, where the rating is at the same time as the test, will be positively correlated with test score measurement error.

H2. Teacher ratings on a different attainment from that tested, where the rating is at the same time as the test, will be positively correlated with test score error but to a lesser extent than for the same attainment.

H3. Teacher ratings when the child is at a different age from that of the test will be uncorrelated with test score measurement error whether or not the same attainment is tested.

H4. Teacher ratings are not correlated with disturbance terms from the regression of second-occasion score on first-occasion score.

H5. Social class is correlated with disturbance terms.

H6. Social class if not correlated with test score measurement error.

Generally, teacher ratings are held to correlate with test score measurement errors only when tested at the same time as the test and not to be correlated with equation disturbances. In contrast, social class is hypothesized as correlating with disturbances but not with test score measurement error when measured at the same time as the test.

Examining equation (7) we see that these six hypotheses give rise to the following predictions. The hypotheses giving rise to each prediction are given in brackets after the prediction.

P1. Comparing teacher ratings at 11 years, the lowest estimate of β will occur for the teacher rating of the same attainment (from H_1 to H_4, particularly H_2 affecting r_{Ze_1}), and at 16 years the highest value will occur for the rating of the same attainment (from H_1 to H_4, particularly H_2 affecting r_{Ze_2}).

P2. For a given attainment, teacher ratings will give higher estimates of β when measured at 16 than 11 years (from H_1 to H_4).

P3. For a given attainment, teacher ratings will give higher estimates of β when measured at 7 rather than 11 years (from H_1 to H_4).

P4. There is no difference in estimates between teachers ratings at 7 years (from H_3, H_4).

P5. Social class gives higher estimates of β than teacher ratings at 7 and 11 years (from H_1 to H_6).

P6. Social class will give similar estimates of β irrespective of the age at which measured (from H_5, H_6).

Note that these predictions are not unequivocal tests of the hypotheses. For instance, even if P6 holds, one could conceive of different correlations of social class with measurement error at different ages, these terms being counteracted by different correlations with equation disturbances. This, however, seems unlikely.

Results

Tables I and II give estimated regression coefficients for reading and mathematics, respectively, for 16-year attainment on 11-year attainment using a variety of teacher ratings and social class as instrumental variables at ages 7, 11, and 16. Using the ungrouped 11-year test score as instrumental variable gives the ordinary least squares estimate, and assuming the sample size is large enough to make use of the asymptotic properties, this enables reliability estimates to be calculated for each choice of instrumental variable by dividing the ordinary least squares estimate by the instrumental variable estimate. Each prediction will be examined in turn.

P1. This holds for mathematics using teacher ratings both at ages 11 and 16 and for reading for teacher ratings at 11 but not 16.

P2. This holds for comparable teacher ratings at ages 11 and 16 for both reading and mathematics test scores.

P3. This only holds for one out of the six possible comparisons, namely, teacher rating of "number" for reading attainment regression.

P4. This holds for both attainments.

P5. This holds in all cases.

P6. This holds at all ages for both attainments.

Note that predictions P4 and P6 specify no differences between regression coefficients, whereas the other predictions are of a difference in a specified direction. In fact, for P4 and P6 the differences between coefficients are small in relation to the standard errors, which have been estimated in the usual way, and are strictly applicable only if the estimates are consistent.

The predictions are all seen to hold generally with the exception of P3. This implies the rejection of H3 or H4 or both. Rejecting H3 implies that the

TABLE I

Estimated Regression Coefficients of Reading Test at 16 years on Reading Test at 11 years Adjusted for Measurement Error, Using a Number of Instrumental Variables Separately

Instrumental variable measured at:	7 years		11 years		16 years
Teacher rating of:					
Oral	0.955	Oral	0.972		
Reading	0.944	Use of Books	0.964	English	1.042
Number	0.990	Number	0.974	Mathematics	1.067
Creativity	0.990				
		General Knowledge	0.979	Practical Subjects	1.047
Social Class	1.057	Social Class	1.070	Social Class	1.064
Reading Test at 11 (interval scale) (OLS estimate)	0.797				
Reading Test at 11 (5 category groupings)	0.810				

Average standard errors of regression coefficients:			
	Using Teacher ratings at	7 years	0.017
		11	0.013
		16	0.015
	Social Class		0.027
	Reading Test at	11 years	0.008

TABLE II

Estimated Regression Coefficients of Mathematics Test at Age 16 *on Mathematics Test at Age* 11 *Adjusted for Measurement Error Using a Number of Instrumental Variables Separately*

Instrumental variable at:	7 years		11 years		16 years
Teacher rating of:					
Oral	0.883	Oral	0.089		
Reading	0.849	Use of Books	0.884	English	0.992
Number	0.854	Number	0.874	Mathematics	1.073
Creativity	0.911				
		General Knowledge	0.920	Practical Subjects	1.029
Social Class	0.994	Social Class	0.991	Social Class	1.025
Math Test at 11 (Interval Scale) (OLS estimate)	0.748				
Math Test at 11 (5 category grouping)	0.763				

Average standard errors of regression coefficients:	Using Teachers ratings at	7 years	0.018
		11	0.015
		16	0.016
	Social Class		0.030
	Mathematics Test at	11 years	0.010

differences in coefficients among the different teacher ratings at age 7 should be similar to those using the three corresponding teacher ratings at age 11. This is true with the one exception being the reversal in the relative magnitude of the teacher ratings of reading and number between ages 7 and 11 for mathematics.

The smaller coefficient estimates using 7-year rather than 11-year ratings could be explained by a correlation of 11-year rating with errors in the dependent variable, which counteracts the correlation with errors in the independent variable, the 7-year ratings having lower correlations with the dependent variable.

If H4 is the sole reason for the failure of P3, this suggests that given the 11-year test score, the partial correlation of teacher ratings with social class at 11, which if H5 holds is correlated with the disturbances, would be higher for 11-year teacher ratings than for 7-year teacher ratings. This is not the case.

It seems, then, that we should discard social class as a suitable instrumental variable due to its correlation with the equation disturbances, and teacher ratings at 16 years are positively correlated with test score error at 16 (form H1, H2) and probably also with equation disturbances. This leaves a choice between teacher ratings at 7 and 11 years. As H3 does not hold, we cannot be completely content with using the same-attainment, 7-year teacher ratings, and in addition, it is not known how highly correlated these are with the disturbances. It was suggested earlier, however, that we should expect the 11-year ratings to be more highly correlated with the disturbances. As H2 also holds, the wisest choice would seem to be the rating of a different attainment (out of mathematics or reading) at age 7. For the reading attainment this gives a reliability of 0.81 (using teacher rating of number at age 7) and for mathematics attainment gives a reliability of 0.89 using teacher rating of reading at age 7. In fact, the choice between 7 and 11 years for the instrumental variables makes little difference for mathematics attainment giving a reliability of 0.86 using reading rating at age 11, and for reading attainment the difference in reliability estimate is only 0.005.

The instrumental variable estimate chosen here is subject to possible inconsistencies from two causes operating in opposite directions. If H3 does not hold and the errors in the instrumental variable have positive correlations with the test score error, which are higher at age 11 that at age 16, then the expected value of the estimate is higher than the true value. On the other hand, if H4 does not hold and the errors in the instrumental variable have a positive correlation with the disturbance term, then the expected value is lower than the true value. If both H3 and H4 do not hold, then these biases operated in opposite directions and an evaluation of the merits of the estimate depend on further analysis of the relative strengths of the inconsistencies from the two causes.

The question naturally arises here as to whether the use of tests of the same attainment at different ages is necessary to obtain reasonable estimates of reliability coefficients by this method. Estimates of the reliability of 11-year reading test score obtained by regressing 16-year mathematics score on 11-year reading score using as instrumental variables teachers' ratings of reading and mathematics attainments separately at 11 years, gave reliability estimates of 0.76 and 0.61, respectively, compared with the value of 0.81 give above. This suggests that the disturbance terms in this regression are correlated with the mathematics teachers' ratings. For the regression of 16-year reading test on 11-year mathematics test using teachers' ratings of mathematics and reading separately as instrumental variables, reliability estimates of 0.85, 0.68, respectively, are obtained, compared with the value of 0.88 given above. Care should therefore be taken when estimating reliabilities by this method to use similar attainments as dependent and predictor variables.

Use of Grouped First-Occasion Variable as Instrumental Variable

Table III gives estimated regression coefficients for both reading and mathematics when the first-occasion variable is grouped into 2, 3, 5, or 7 groups of equal size.

The inconsistency of the grouping estimator (b_G) relative to the ungrouped (OLS) estimator (b_{OLS}) is given by

$$k = \frac{b_{IV} - b_G}{b_{IV} - b_{OLS}}, \tag{9}$$

TABLE III

Estimated Regression Coefficients and Standard Errors Using the Grouped Predictor as Instrumental Variable

	Reading	k	Mathematics	k
Ungrouped (OLS)	0.797 (0.0082)	1.00	0.748 (0.0097)	1.00
No. of groups (of equal size)				
7	0.808 (0.0085)	0.94	0.755 (0.0100)	0.93
5	0.810 (0.0086)	0.93	0.763 (0.0102)	0.86
3	0.818 (0.0092)	0.89	0.777 (0.0109)	0.73
2	0.827 (0.0103)	0.84	0.780 (0.0121)	0.70
Varying position of dichotomy with 2 groups				
Proportion in lower test group				
0.2	0.810 (0.012)	0.93	0.687 (0.014)	1.58
0.4	0.823 (0.010)	0.87	0.765 (0.012)	0.84
0.6	0.822 (0.010)	0.87	0.802 (0.012)	0.49
0.8	0.789 (0.012)	1.04	0.805 (0.014)	0.46

Note. Standard errors in brackets; k is defined in equation 9.

where b_{IV} is the instrumental variable estimator (using an appropriate teacher rating) and is assumed to be consistent. As suggested in the Use of Dummy Variables section, substantial inconsistencies are indicated in these data, which are greater with a larger number of groups. Thus, there is a trade-off between inconsistency and efficiency, the latter being greater as the number of groups is increased. For reading attainment the lowest estimate of reliability, arising from the division into two groups, is 0.97. This is higher than any of the values derived from the regression estimates by instrumental variables methods given in Table I. Furthermore, the estimated regression coefficient is seen to vary with the point of dichotomy. Whereas for reading the lowest estimate occurs for both extreme divisions, for mathematics the lowest estimate occurs when division is at the lower end of the scale and the highest estimate when division is at the higher end.

If the assumption is made that the correlations of the grouped first-occasion variable with the disturbances and error in the second-occasion variable are both zero, then using the reliability estimates from the previous section we can substitute in (7) to obtain the correlations with the error in the first occasion variable, r_{Ze_1}. These are given in Table IV, assuming the reliability estimates given in the Results section (0.81 and 0.89) are correct.

Thus, the correlations, while reasonably constant for reading, are systematically decreasing for mathematics. We have no good explanation for this but possible causes are nonhomogeneity of errors in the mathematics test or a nonzero correlation between true score and errors of measurement.

The Use of Grouped Test Score as an Instrumental Variable

Test scores of reading and mathematics at 7, 11, and 16 years, a score of General Ability at 11 years (with Verbal and Nonverbal components) and a Copying Design Test are considered as instrumental variables, and the results are given in Table V.

Since short-term fluctuations in attainment will be correlated to some extent over all attainments, we would expect the arguments and predictions (given previously) in relation to teacher ratings to apply to test scores. P1 is satisfied trivially in the light of the results on the use of a grouped predictor as

TABLE IV

Correlations of Dichotomized Instrumental Variable and First-occasion Measurement Error for Different Division Points

	Proportion below division point				
	0.2	0.4	0.5	0.6	0.8
Reading	0.280	0.302	0.329	0.305	0.324
Mathematics	0.390	0.226	0.233	0.128	0.120

TABLE V

Regression Estimates of 16*-year Attainment Test on* 11*-year Attainment Test in Reading and Mathematics Using Grouped Test Scores as Instrumental Variables*

	Reading	Mathematics
Instrumental variable		
At 7 years:		
Reading Test	0.920 (0.015)	0.822 (0.017)
Mathematics Test	1.004 (0.020)	0.850 (0.018)
At 11 years:		
Reading Test	0.810 (0.009)	0.866 (0.014)
Mathematics Test	0.995 (0.012)	0.763 (0.010)
General Ability Test: Verbal	0.942 (0.012)	0.826 (0.013)
: Nonverbal	0.982 (0.014)	0.901 (0.014)
: Overall	0.957 (0.012)	0.860 (0.013)
Copying Designs Test	0.989 (0.032)	0.933 (0.032)
At 16 years:		
Reading Test	1.197 (0.013)	0.949 (0.015)
Mathematics Test	1.053 (0.015)	1.315 (0.017)

Note. Standard errors in brackets.

instrumental variable. P2 is satisfied, but P3 is again contradicted in one of four cases by the behavior of the reading tests at 7 and 11 when used as instrumental variable for mathematics attainment.

Generally, the behavior of test scores is similar to the teacher ratings and the standard errors are similar, giving little indication for preference of one set of variables over the other.

Discussion

Our results suggest that differing correlations of instrumental variables with measurement errors account for the observed differences in regression and reliability estimates, although social class has a negligible correlation with measurement error but a nonnegligible correlation with the error of prediction.

The estimated correlation coefficient between true scores on reading and mathematics tests at 11 and 16 years is respectively 0.96 and 0.90. The estimated reliabilities using the selected instrumental variables (teacher ratings of an unrelated ability at age 7 as the predictor) are 0.81 and 0.89 for reading and mathematics, respectively. These compare with the values of 0.82 and 0.94 given in Goldstein (1979) by split-half item analysis on a subsample of 300 cases. While the values for reading are similar, the value obtained by item analysis for mathematics is somewhat higher than any obtained for the instrumental variables used here, although the difference is of the same order as the standard error of the separate estimates.

For reading attainment the estimated standard error of the reliability estimate obtained by item analysis is 0.030, while the instrumental variable method gives 0.012. A split-half estimate using all available data would have a standard error of about 0.007. For mathematics the relevant standard errors are 0.020, 0.014 and 0.005.

Using a grouping of the predictor variable itself as an instrumental variable gives estimates of the reliability that are higher than any obtained using other variables as instrumental variables, irrespective of the number of groups used. These estimators, we suggest, are not to be recommended.

While instrumental variable estimation has had a long history (early papers on theory and application in the economic field include Durbin, 1954; Madansky, 1959; Reiersol, 1945; Sargan, 1958; Wald 1940), it has not yet become generally accepted as an estimation method in the social and educational fields. Sargan (1958), in discussing the (unknown) correlation of instrumental variables with measurement error in an economic context states:

> It is not easy to justify the basic assumption concerning these errors, namely that they are independent of the instrumental variables. It seems likely that they will vary with a trend and with the trade cycle. Insofar as this is true, the method discussed here will lead to inconsistent estimates of the coefficients. Nothing can be done about this since presumably if anything were known about this type of error, better estimates of the variables could be produced. It must be hoped that the estimates of the variables are sufficiently accurate, so that systematic errors of this kind are small. (p. 396)

We have argued that comparisons of different instrumental variables, considered separately, can throw some light on the error structure in the data, and thus lead to better knowledge of the consistency of the estimates produced. Furthermore, it is also our view that this approach provides a flexible tool for an empirical study of the various assumptions needed to produce good estimates.

Finally, four issues seem particularly worthy of attention:

1. Obtaining estimates of the standard errors of the difference between different instrumental variable estimates (these will be lower than those obtained using the individual standard errors and assuming independence). This would enable a more careful analysis of the hypotheses of the paper.
2. Obtaining good estimates of the standard errors of the reliability estimates produced by instrumental variables methods.
3. Examination of the use of more than one instrumental variable in connection with a single predictor in terms of the efficiency and consistency of estimates.

4. The study of differing reliabilities and measurement error variances in different groups such as social classes, to incorporate these into linear model estimates.

Acknowledgements

We are grateful to the National Children's Bureau for permission to use the data and to Dougal Hutchinson for his comments on an early draft of the paper. The work was carried out largely on a grant from the National Institute of Education, Washington (NIE-G-77-0065).

References

Anderson, E. B. Comparing latent distributions. *Psychometrika*, 1980, *45*, 121–134.

Brown, R. L. Bivariate structural relation. *Biometrika*, 1957, *44*, 84–90.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajartnan. *The dependability of behavioural measurements: Theory of generalisability for scores and profiles*. New York: Wiley, 1972.

Davie, R., Butler, N. R., & Goldstein, H. *From birth to seven: A report of the National Child Development Study*. London: Longmans, 1972.

Durbin, J. Errors in variables. *Review of Institute of International Statistics*, 1954, *22*, 23–54.

Fogelman, K. *Britain's sixteen year olds*. London: National Children's Bureau, 1976.

Goldstein, H. Some models for analysing longitudinal data on educational attainment. *Journal of the Royal Statistical Society*, 1979, A, *142*, 402–442.

Goldstein, H. Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 1980, *33*, 234–246.

Johnston, J. *Econometric methods*. New York: McGraw-Hill, 1972.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. *Psychometrika*, 1971, *36*, 109–133.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 2). London: Griffin, 1977.

Layard, M. W. J. Robust large sample tests for homogeneity of variances. *Journal of the American Statistical Association*, 1973, *68*, 195–198.

Lord, F. M., & Novick, M. R. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley, 1968.

McDonald, R. The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 1981, *34*, 100–117.

Madansky, A. The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 1959, *54*, 173–205.

Nair, K. R., & Banerjee. A note of fitting straight lines if both variables are subject to error. *Sankya*, 1942, *6*, 331–333.

Neyman, J., & Scott, E. L. On certain methods of estimating the linear structural relation. *Ann. Math. Statist.*, 1951, *22*, 352–355.

Reiersol, O. Confluence analysis of means of instrumental sets of variables. *Arkiv. for Mathematik, Astronomi Och Fysik*, 1945, *32*.

Sargan, J. D. The estimation of economic relationships using instrumental variables. *Econometrica*, 1958, *26*, 393–415.

Wald, A. Fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.*, 1940, *11*, 284–300.

Warren, R. D., White, J. K., & Fuller, W. A. An errors-in-variables analysis of managerial role performance. *Journal of the American Statistical Association*, 1974, *69*, 886–893.

Authors

ECOB, RUSSELL, Research statistician in the Research and Statistics division of the Inner London Education Authority, England. Specializations: Longitudinal studies, educational measurement, effects of schooling.

GOLDSTEIN, HARVEY, Professor of statistical methods in the department of Mathematics, Statistics and Computing at the University of London Institute of Education. *Specializations*: Longitudinal studies, test score theory, educational measurement.