MODEL-BASED RANKING OF SCHOOLS

CSE Technical Report 297

Ita G.G. Kreft
UCLA Center for Research on Evaluation,
Standards and Student Testing

Jan de Leeuw
Center for Technology Assessment

Cent
Department of psychology and
mathematics

Abstract

This report presents an empirical study of the problem of ranking schools in terms of their quality. The unadjusted ranks are compared with rankings obtained by various adjustment schemes based on ANCOVA and random effects models. The advantage of random coefficient models is highlighted by the reanalysis of a 1959 data set that contains information about 1290 children in 37 schools in the city of Groningen, The Netherlands. The authors demonstrate that any simple ranking for the schools would be misleading because of important interactions of student background with school.

Introduction[1]

The empirical example analyzed in this report is from the field of school effectiveness research. The analysis model is a two-level linear model. These two choices are not accidental. In educational research there is a clear link between substantive and methodological issues, and it is also clear that the validity of statistical inferences is enhanced when appropriate analytical models are used. Thus this paper is about model building in educational research, and it is based on the knowledge we have about this field and about the way the data are collected. More specifically, we are interested in ways to rank schools in terms of their effectiveness to train students. The definition of *effectiveness* is quite controversial (Kreft, 1987), but we shall adopt a simple operational one for the purposes of this paper. We shall see that the rankings do not depend only on the measured properties of the schools, but also on the model chosen to represent the differences among the schools.

In this report we emphasize that a school effectiveness researcher has no other way to proceed than to base the choice of her analysis model on the knowledge she has about her data. In principle there are hundreds of possible ways to order schools in terms of their success. Moreover, all these different ways of ordering are not necessarily closely related with each other. On the contrary, as we will show, some may be negatively correlated, some may not be correlated at all. It is obvious that with so many choices a researcher is able to satisfy whomever she wants, just by choosing the one model that will be most pleasing to the audience from the many models that exist. Although we are aware that school effectiveness research is more often than not policy-oriented research (compare Kreft, 1987), pleasing criteria should not guide the choice of the analysis model and thus the way schools are ordered. What we need are more objective criteria—criteria that enable us to choose the best analysis method, best in the sense of statistical and substantive adequacy or appropriateness.

In this report we give some criteria for making this choice in the field of school effectiveness research. The arguments are based on statistical as well as theoretical reasoning, in close interplay with each other. We argue that these two types of considerations should be in agreement. As the starting point, we introduce some of the well known traditional linear models, such as analysis of variance and covariance and multiple regression techniques. These methods are often applied in situations where they cannot take account of certain salient features of the data. School effectiveness in general, and the data set analysed in this paper in particular, are cases in point. Later in the report we introduce models that are designed for the school effectiveness situation in educational research. School effectiveness research is complicated, and this leads to a more complicated statistical model (see Aitkin & Longford, 1986; De Leeuw & Kreft, 1986; Raudenbush & Bryk, 1986). We will examine how these more complicated models deal with the practical problem of ranking schools. Policy-oriented researchers often point out that more sophisticated statistical techniques are unnecessary because they do not give essentially different results. We shall see that rankings of schools in terms of output can be very misleading indeed, and that correcting the rankings for input characteristics must be done carefully because otherwise a lot of spurious variation is introduced. The researcher diligently has to search for an optimal place on the continuum between bad-fitting parsimoneous models, which promise spurious precision, and good-fitting models with many parameters, which capitalize on unstable effects in the data.

Description of the Data

For our example we use the Dutch GALO data (described by Peschar, 1983). These data contain information about primary school leavers in 37 schools in the city of Groningen in 1959. This is the same data we used in a previous article (De Leeuw &

---

[1]We would like to thank Nick Longford (ETS) for many valuable comments.

Kreft, 1986). In a sense this paper is an extension of the earlier article, within which we compared different estimation procedures. (It is also interesting to compare our analysis with the similar work of Aitkin & Longford, 1986.) We also compare different methods of analysis in order to show that different models lead to different solutions. To do this we use variables that are measured at the pupil level.

The dependent variable is the advice, or recommendation, given by the head teacher about the most appropriate school for a particular student after primary education. (In the Netherlands, a head teacher's recommendation has great influence on a student's choice of secondary school.) Advice has seven categories which represent the seven options for secondary education in the Netherlands. These categories are scored 1 to 7, with 1 representing the least prestigous type of secondary school and 7 representing the most prestigious type of school. The resulting variable ADV is used as a numerical variable. This is a choice which is difficult to defend: Clearly another method of scoring (such as 1, 4, 9, ... , 49) would lead to different results. Given the nature of the linear statistical techniques that are available, we have to make some kind of choice. Previous experience with these data indicates that our equal interval integer scores behave rather nicely, in the sense that they lead to fairly linear regressions.

As predictors for our operationalization of school success we use three variables: students' sex (SEX), students' intelligence (IQ), and fathers' occupation (SES). Intelligence was measured using the Groninger Intelligence Test (see Peschar, 1975). Father's occupation comprises six categories, each treated as a numerical variable: (1) unskilled, blue collar laborers; (2) schooled laborers; (3) lower-level white collar workers; (4) owners of small businesses; (5) higher level white collar workers; and (6) professionals.

Another predictor is the distinction between the 37 schools as 37 groups. The number of students varies per school, from a low of 11 to a high of 66 (see Table 1a).

The First Ordering of Schools by Way of Means

To rank order the schools we start in a very simple way. We take the outcome characteristic, in our case ADV, and we compute the averages for each school. These averages then produce the ranking of the schools. Actually we have done a bit more. Table 1a gives the rank numbers of the 37 schools on all four variables. Column 1 of Table 1a gives the school number, Column 2 the school size, and the remaining four columns give the rank orders in terms of average SEX, IQ, ADV, and SES. Knowing that IQ and SES have a fairly direct influence on the advice of the head teacher, it is not surprising that the ordering by mean ADV in Table 1a is closely related to an ordering by mean IQ or mean SES. This can be seen in Table 1b, which gives Spearman rank correlations for the variables in Table 1a, including school size (SIZE).

Observe that Table 1b shows a significant positive correlation between SIZE and the three variables IQ, SES, and ADV. This is clearly because of schools such as Schools 25, 27, 29, and 37. These are small schools with low SES averages. They are, presumably, inner city schools with many children of unskilled workers. Larger schools typically have higher SES averages.

The data can be used to illustrate another interesting phenomenon. At an individual level the Spearman rank correlation between ADV and IQ is 0.743, while at school level the correlation between mean ADV and mean IQ is 0.881. This blowing up of the correlation coefficient when calculated over aggregated variables is a well-documented phenomenon, known as the Robinson effect. (This effect was named after Robinson, who was one of the first to describe and explain it (Robinson, 1950). Another example is the correlation between SES and ADV, which is 0.305 at the individual level and is blown up to 0.714 at the school level. Of course, the ranking of

Table 1a
Ranking of Schools by Variables

School	SIZE	SEX	IQ	ADV	SES
1	29	6	12	6	9
2	33	1	34	34	31
3	31	11	27	30	19
4	66	30	32	33	33
5	39	14	2	4	8
6	45	7	8	24	32
7	39	5	21	25	28
8	31	24	13	18	6
9	53	33	36	29	25
10	31	21	6	9	13
11	30	8	4	7	18
12	36	19	15	20	27
13	52	15	22	23	22
14	29	10	25	28	24
15	33	36	18	16	7
16	65	28	29	37	26
17	57	12	30	31	35
18	31	32	16	11	4
19	26	35	5	12	11
20	27	20	14	14	14
21	25	26	17	5	20
22	27	4	23	19	21
23	26	16	20	21	30
24	36	2	3	2	10
25	11	37	1	1	1
26	27	25	24	17	16
27	15	27	35	26	3
28	27	29	10	8	15
29	20	34	26	22	5
30	32	3	7	13	17
31	49	31	31	32	36
32	57	22	37	35	37
33	37	23	28	27	34
34	39	17	33	36	29
35	35	13	19	15	12
36	28	18	11	10	23
37	16	9	9	3	2

Table 1b
Rank Correlations between Rank Orders

	SIZE	SEX	IQ	ADV	SES
SIZE	1.000	-.104	.383	.574	.636
SEX	-.104	1.000	.154	.043	-.205
IQ	.383	.154	1.000	.881	.553
ADV	.574	.043	.881	1.000	.714
SES	.636	-.205	.553	.714	1.000

mean ADV reflects to a great extent the ranking of the means for IQ and SES (see Table 1b).

It is clear from our results so far that the ordering of schools from highest to lowest mean ADV is influenced by the student characteristics of the school population. If our goal is to see which schools are more successful than others, irrespective of the population, we have to correct the mean ADV for the influence of IQ and SES. The conclusion that in order for schools to have better results they need to attract high SES and high IQ students is a trivial one. If we want to know whether schools have an effect on students next to and above their background characteristics, we have to control for these differences in individual backgrounds.

Statistical Control for SES, IQ, and SEX

Our use of averages in the previous section can be formalized in terms of using linear models. This makes it possible to talk about assumptions, and it also points out various natural alternatives. In our models index j is used for schools, and index i is used for students, who are nested in schools. Boldface notation is used throughout to indicate random variables.

The first model we use is

$$\mathbf{y}_{ij} = \alpha_j + \boldsymbol{\epsilon}_{ij}, \qquad [1]$$

where the disturbances are normal, independent, centered, and homoscedastic (this last assumption means that they are assumed to have a common variance σ^2 for each individual). This is the one-way analysis of variance (ANOVA) model, in which there is a single α_j parameter for every school. From the linear model point of view, Model 1 is the null model in which coefficients for all predictor variables are set equal to zero, except for the intercept. Estimation of the parameters produces the ordering of schools by way of uncorrected school means (i.e., column ADV of Table 1a).

A method to test whether schools are associated with achievement levels is to *partial out* the influence of student and school characteristics (attributes). The simplest approach is to perform an analysis of covariance (ANCOVA), where schools are the groups and SES, SEX, and IQ are the covariates. Model 2 is the analysis of covariance model, with the dependent variable ADV,

$$\mathbf{y}_{ij} = \alpha_j + \beta_1 \mathrm{SEX}_{ij} + \beta_2 \mathrm{IQ}_{ij} + \beta_3 \mathrm{SES}_{ij} + \boldsymbol{\epsilon}_{ij}. \qquad [2]$$

This is a substantial improvement over Model 1 as far as the unexplained part of the variation is concerned; it is decreased from 0.815 to 0.346. The estimated αs can be used again to order the schools. We can use the residual variances, which are 2.07 and 0.88, respectively, to test the difference between Models 1 and 2 9i.e., we can test the hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ within Model 2). Normally we would use an F-test for this purpose. Because F-tests cannot be applied in the random coefficient models, we will use the likelihood ratio chi square. The chi square is 1290 *(ln 2.07 - ln 0.88) = 1103.44. With only three degrees of freedom, this clearly is highly significant.

If we compare the ANOVA and ANCOVA columns in Table 2a we see a different order. The rank correlation between them is r = 0.667, indicating only a moderate agreement between the two orderings. Our conclusion so far is that controlling for background characteristics does make a difference. Using only the aggregated means is misleading. In the following tables, high levels are coded with upper case I and S, low levels with lower case i and s.

Table 2a
Ranking of Schools by Fixed Models

School	SIZE	ANOVA	ANCOVA	is	iS	Is	IS
1	29	6	5	28	24	3	3
2	33	34	32	8	16	10	23
3	31	30	35	35	36	33	35
4	66	33	29	23	23	25	25
5	39	4	17	27	15	18	5
6	45	24	36	34	35	32	32
7	39	25	30	25	27	22	22
8	31	18	31	36	29	37	36
9	53	29	9	5	6	9	10
10	31	9	16	11	13	26	29
11	30	7	12	29	32	15	20
12	36	20	23	31	26	31	28
13	52	23	13	18	25	8	12
14	29	28	33	20	28	21	31
15	33	16	22	17	30	7	26
16	65	37	37	33	34	36	37
17	57	31	24	1	5	2	4
18	31	11	19	10	8	17	8
19	26	12	28	32	31	24	27
20	27	14	20	24	7	34	11
21	25	5	1	19	17	1	1
22	27	19	10	2	3	19	14
23	26	21	18	22	9	30	13
24	36	2	4	21	18	11	6
25	11	1	2	9	33	5	34
26	27	17	7	12	22	12	16
27	15	26	8	26	11	20	7
28	27	8	11	13	20	16	21
29	20	22	6	37	37	14	19
30	32	13	27	15	19	4	9
31	49	32	26	16	12	28	30
32	57	35	25	6	2	29	15
33	37	27	21	3	4	23	18
34	39	36	34	14	10	35	33
35	35	15	15	7	14	13	24
36	28	10	14	30	21	27	17
37	16	3	3	4	1	6	2

Table 2b
Rank Correlations between Rank Orders

	ANOVA	ANCOVA	is	iS	Is	IS
ANOVA	1.000	.669	-.036	-.062	.434	.383
ANCOVA	.669	1.000	.257	.226	.624	.621
is	-.036	.257	1.000	.734	.450	.315
iS	-.062	.226	.734	1.000	.099	.563
Is	.434	.624	.450	.099	1.000	.598
IS	.383	.621	.315	.563	.598	1.000

However, using ANCOVA as the way to avoid a bias in the direction of the school population characteristics has its own problems. It is based on at least one critical assumption, which is *a priori* unlikely to be true. This is the *homogeneity of slope* (i.e., the assumption that all regression lines are parallel in schools or that there is no interaction effect between school and student characteristics). If *heterogeneity of slopes* is more likely we can still control for student characteristics by fitting the same model as in Model 2, but now for each school separately; the result is Model 3.

Thus, Model 3 allows each school to have its own estimates for the regression coefficients,

$$y_{ij} = \alpha_j + \beta_{1j}SEX_{ij} + \beta_{2j}IQ_{ij} + \beta_{3j}SES_{ij} + \varepsilon_{ij}. \qquad [3]$$

If we fit this model the residual variance is 0.78, which means that the proportion of unexplained variance drops to 0.307. This corresponds to a likelihood ratio statistic (for testing the homogeneity of "within-school" slopes) of 1290*(ln 0.88 - ln 0.78) = 155.61, with 36 * 3 = 108 degrees of freedom. This transforms to a z-value of 3.24, which is quite small for a sample as large as this one, though significant. Although there is some evidence of heterogeneous within-school slopes, it is not very strong.

Let us ignore this statistical information for the moment, and act as if the within-school slopes are different. If we allow slopes to differ per school, the ordering of the schools becomes less simple. Some schools may be successful for high-IQ students or for females, while the same schools are not successful for low-IQ students or for males. Since we have 2 (SEX) x 7 (SES) x 85 (IQ from 60 to 144) = 1190 different conceivable students, the number of possible comparisons is large. To illustrate this we picked, rather arbitrarily, four different types of students by crossing high IQ/low IQ with high SES/low SES. This produces four different orderings. Observe that we give model-based rankings here. We do not compute average ADV for all groups on all schools, but we compute ADV predicted by the linear model for these combinations of the independent variables. The former would be more precise but very inefficient, because there will be very few Is or IS girls in any one school.

The first group (is) are girls with an IQ of 90 who have blue-collar workers as fathers (SES Category 2). The second group of girls (iS) have the same IQ, but their fathers own a small business or work as businessmen (SES Category 4). The third (Is) and fourth (IS) orderings are based on girls with the same two SES backgrounds, but their IQ is now considerably higher—it is equal to 110. The columns in Table 2a give, next to the orderings from Models 1 and 2, the orderings by using Model 3 in the last four columns as follows:

is predicted advice = $\alpha_j + \beta_{1j}$ (SEX=2) + β_{2j}(IQ=90) + β_{3j}(SES=2)

iS predicted advice = $\alpha_j + \beta_{1j}$ (SEX=2) + β_{2j}(IQ=90) + β_{3j}(SES=4)

Is predicted advice = $\alpha_j + \beta_{1j}$ (SEX=2) + β_{2j}(IQ=110) + β_{3j}(SES=2)

IS predicted advice = $\alpha_j + \beta_{1j}$ (SEX=2) + β_{2j}(IQ=110) + β_{3j}(SES=4)

Replacing the αs and βs with the different estimated values per school produces 37 outcomes for the prediction of ADV and thus for the rank order for all schools. The results are indeed different for different types of students, as we can see when comparing the last four columns in Table 2a. Each row in Table 2a contains the rank numbers for one single school over the six different methods of ordering. Table 2b has the Spearman rank correlations between the six rank orders.

Some examples of the different rankings that a school receives when we compare different types students in those schools are illustrated by Schools 1, 21 and 23 (see Table 2a). School 1 scores high for the IQ-110 students (columns Is and IS) by occupying a third place, but does poorly (28th and 24th place) for IQ-90 girls (columns

is and IS). The same is true for School 21, which scores high for IQ-110 students but does much worse for 90-IQ students. It drops from being the best school in the last two columns to being an average school (number 19 and 17, respectively) in the i columns. School 23 also jumps around: It does rather poorly on most scales but is up to 9th and 13th place for higher SES girls (see columns iS and Is). This shows an interaction effect between student characteristics and the school.

The correlations between the orderings, with different students in Model 3 as the ranking criteria, are moderate to low. The highest correlation is only 0.734. This is the one between is and iS, the low-IQ girls only differing in the occupation of their fathers. The association between ANOVA and ANCOVA is also somewhat higher than the others, r = 0.669. The correlation between orderings for low-SES girls that differ only in IQ (orderings is and Is) is r = 0.450. The largest discrepancy is between girls with different IQs and fathers with different occupations (orderings iS and Is); this correlation is as low as r = 0.099. The correlations between ANCOVA and the last four orderings is moderately high between the two groups with high-IQ girls (r = 0.624 and r = 0.621), and low for the low-IQ girls (r = 0.257 and r = 0.226). It is not surprising that ANOVA, the uncorrected means, has in general the lowest correlation with the other five.

Especially different are the conclusions based on low-IQ students when compared with the other orderings. Partly this is related to the size of the school. For low-IQ students the average predicted ADV is negatively correlated with the size of the school, while for high-IQ students there is a fairly strong positive correlation. This seems to indicate that small schools give relatively high advice to low-IQ students, while large schools give relatively low advice (that is, low-IQ students in small schools are given greater encouragement to go to prestigious secondary schools than are low-IQ students in large schools). The orderings given by ANOVA and ANCOVA agree more with the high-IQ orderings Is and IS. This is unfortunate, since often research in school effectiveness is interested in the success of schools with underprivileged students. In our case, and probably in a great deal of school effectiveness research, it is clear that neglecting this potential interaction effect, as the ANOVA and ANCOVA orderings do, can lead to different and biased conclusions about which schools are more successful.

The orderings based on Model 3, taken together, will generally produce a more complete picture. In comparison, the uncorrected means ANOVA seems to be the most biased and least informative way to order schools (of course, its fit is also bad when compared with the ANCOVA model). Here we must consider the indirectness of the methods used, estimate the school effect, and then rank it. On the one hand, the school effects may be heavily biased, but the ranking may still turn out without bias. On the other hand, it is important to note that in some cases estimated rankings are obtained where no ranking is really appropriate because of the extremely poor fit of the model.

However, we still encounter problems in using Model 3. This is already clear from the fact that the differences between the slopes are not very significant (compare the likelihood ratio test earlier), while ordering the schools by the various types of individuals (as we have above) produces wildly different results. If, in our ordering of schools, we estimate different models for different schools and then compare outcomes, we take the coefficients at face value. In doing so, we ignore the fact that some estimates are more efficient than others (small versus large standard errors) and that some may even be biased as a result of small non-random groups and/or outliers. In our case the number of students per school differs markedly, which causes some schools to have more reliable estimates than others. School 25, for instance, only has 11 students. It is ranked lowest on IQ, ADV, and SES. If we use ANCOVA to correct for background it moves up one place, but if we let the school determine its own regression coefficients, strange things happen. For high SES students this school turns out to be one of the best there is, but high SES students actually do not attend this school. Taking such things into account prompts the search for a better way to analyze the data.

Choosing a Better Model

We start our search for a better model with an examination of the assumptions behind the traditional linear models. One of the assumptions of the fixed linear model is simple random sampling. This assumption often is violated in educational research, and the violation has an impact on the analysis. This is shown in the equation of all fixed models where it is stated that the individual errors ε_{ij} are uncorrelated, and have a mean of zero and a constant variance σ^2. However, in our case, as well as in educational research in general, we know that students are sampled from within a well-defined population: a particular school. In fact, usually it is not students that are sampled, but schools, and students are nested within them. This gives us a good reason to assume that individual (student) error terms of students in the same group are correlated. The error term contains, in addition to random measurement error, various influences that are not measured, that is, the influence of the variables not in the model (Kreft 1987). Since students in the same class share many hours of common experiences each day, it is somewhat unrealistic to assume (as is done in the fixed models) that unmeasured influences are unsystematic. In a more appropriate model the random terms associated with the students in the same school should be correlated, or, equivalently, should share a common component.

Another problematic aspect is the use of fixed effects models such as AN(C)OVA. ANOVA and ANCOVA are analysis methods designed for the analysis of a fixed number of experiments. But schools are better thought of as a random sample from the population of possible schools, and not as a fixed number of treatments. What we need here is a random effects model. When we compare the traditional random effects model and with the fixed effects model, we find that the main difference is that we are no longer dealing with means or point estimators, but with variance components. Rather than estimating effects directly by taking differences of treatment means from the grand mean, the variance due to treatment is estimated. In random coefficient models it is assumed that the model deviations within the same schools are correlated, and that schools are a random sample from the population of schools. The last assumption allows us to make inferences to other schools not in the sample, while the first assumption provides more reliable estimates. The estimates are no longer based solely on individuals independent of each other, but upon individuals in relation to each other when in the same group. The model as a whole is more reliable, since the coefficients are weighted in relation to their reliability, the size of the group, and the correlations between the individuals within the group. This also makes the chance for Type I errors in the random model smaller than in the fixed models (see De Leeuw & Kreft, 1986; Raudenbush & Bryk, 1988).

Since in the analyses of (co)variance, fixed or random, all analyses have slopes that are parallel between groups, by assuming no interaction between individual student and school characteristics we have to adjust the traditional random effects model to incorporate the possibility of different slopes per school. As shown before when comparing the four orderings of schools for high- and low-IQ girls and for girls with blue collar workers and businessmen as fathers, the actual slopes for IQ and SES are very different for different schools. Therefore, allowing for the possibility that schools have different slopes is a necessary first step.

The random coefficient model is a special variance components model. Again there are several sources of variance in the dependent variable that are decomposed in a pupil variance component and a school variance component. This way the total variance is split into sampling variance and school-level variance. The researcher will eventually try to tie this last variance to a school characteristic. We do not do that here since we merely try to order schools by their outcomes. Because the error structure in this model is much more complicated as a result of the weighting procedure used to estimate the different sources of variance, estimation of the residual variance is less straightforward than in the fixed model. The usual Least Squares (LS) procedures are replaced by Maximum Likelihood (ML) methods, closely related to Bayesian and empirical Bayes methods for linear models (see for details: Aitkin & Longford, 1986; De

Leeuw & Kreft, 1986; Jennrich & Schluchter, 1986; and Longford, 1988a). This results in more efficient and reliable estimates, an outcome due to the application of a shrinkage factor to schools that are far away from the grand mean. The improvement over LS estimates is especially large when samples are small, because a Bayes shrinkage to the grand mean offsets the instability in coefficients that is a result of the presence of small groups (see Raudenbush, 1988). Each LS estimate βj is weighted in proportion to its precision. This improvement is greatest when much heterogeneity among micro parameters exists and least when sample sizes are large. If groups are large, LS estimators are more or less equal to ML estimators. In summary, we can state that very small schools can be left in the analysis because the estimation method makes the outlier problem and chance factors less disturbing. This is not the same as saying that the presence of small schools in the data set is an optimal condition. Small schools will be more subject to shrinkage to the mean (or shrinkage to a macro level variable, if this is in the model) than large schools.

Most statistical software packages provide techniques to analyze random treatment designs (see for instance the SAS module VARCOMP for random analysis of variance model), but these are often not useful in educational data analysis. The reasons are the limitations that are caused by the usual assumptions of equal slopes and equal error variances (or equal n) between schools and uncorrelated error terms within schools. The new 5V module of BMDP can handle the data structures we have in mind (Schluchter, 1988), but for really large data sets the input handling is not very efficient. For our analyses we have used a Macintosh version of the VARCL program of Longford (1988b), which has been designed specifically to handle these random coefficient models.

The estimates produced by the random coefficient models are more reliable and also more efficient. The standard error of the estimates are smaller than the errors around the estimates based on the other models. Standard errors are not only related to sample size and sampling variations, but also to the mean of the group and the deviation of the group parameters compared to the overall mean. Therefore, the number of parameters to be estimated is much smaller than in the separate models for separate schools method. In the latter method the schools are considered independent of each other, and separate and independent parameters are estimated for each school. In our example of 37 schools, using Model 3, this leads to 37 * 4 = 148 parameters (one intercept and three slopes for each school). With the error variance σ^2 this produces 149 parameters. In the next paragraph, in the examples with random coefficient models, we do not estimate parameters but distributions around a mean with a certain variance. In the random coefficient model, with all four coefficients random, this produces, in our case, the estimation of only four means and four variances (for the intercept and the three slopes) plus an individual error variance: ten estimates altogether, many fewer than those in the fixed model. Some models specify extra estimates for the covariance between the slope and intercept, which adds a maximum total of 6 covariances to our model and brings the number of parameters to 16, still many fewer than those in the fixed model. The random coefficient model is more parsimoneous than the fixed model in this sense.

More Rankings

If a researcher has reasons to believe—or, if one insists, if she has a theory—(a) that schools are just a sample from a well-defined population and (b) that slopes may be different between schools and error terms within schools are correlated, then the random coefficient model applies. For instance, when a researcher wants to evaluate policy measures that are intended to benefit specific groups of minority students, a random coefficient model may have to be used in order to measure the effect of the school policy on the slope of SES. In order to estimate the effectiveness of the schools in our own data, we choose three models from the class of random coefficient models.

We study the random coefficient versions of Models 1, 2, and 3. The ANOVA model (Model 1) becomes

$$\mathbf{y}_{ij} = \boldsymbol{\alpha}_j + \boldsymbol{\epsilon}_{ij}. \qquad [4]$$

Observe that we now use bold face for α_j, because it is random. It can be decomposed as

$$\boldsymbol{\alpha}_j = \alpha + \boldsymbol{\gamma}_j. \qquad [5]$$

The assumption is that the disturbances γ_j, which are the same for all individuals in the same school, are normally distributed with expectation zero and constant variance ω^2 for all schools. It is also independent of the error term ϵ_{ij}, and of the school level disturbances of other schools. If we substitute Model 5 in Model 4 we find

$$\mathbf{y}_{ij} = (\alpha + \boldsymbol{\gamma}_j) + \boldsymbol{\epsilon}_{ij}. \qquad [6]$$

This implies that y_{ij} is normally distributed with mean α, and with variance $\sigma^2 + \omega^2$. Outcomes for individuals in different schools are independent, but for individuals in the same school they have a covariance of ω^2, and thus a correlation of $\rho = \omega^2/(\sigma^2 + w^2)$. The model has only three free parameters (α, ω^2, s^2), in contrast to the 37 + 1 = 38 free parameters in ANOVA (Model 1). If we fit the model we find estimates of the two variances equal to 2.13 and 0.39, and thus a correlation of 0.39/(2.13 + 0.39) = 0.15. This deviates significantly from zero.

We also fitted the random intercept (fixed slope) model, which makes a comparison possible with the fixed effect model ANCOVA. This is

$$\mathbf{y}_{ij} = (\alpha + \boldsymbol{\gamma}_j) + \beta_1 SEX_{ij} + \beta_2 IQ_{ij} + \beta_3 SES_{ij} + \boldsymbol{\epsilon}_{ij}. \qquad [7]$$

Only the intercept (the overall effect) is random in Model 7. This model only needs six parameters to be estimated. For the estimates of σ^2 and ω^2 we now find 0.91 and 0.04, which is a correlation between errors of children in the same school of only 0.04 (still significant, though). For the fixed ANCOVA model the estimate of the individual level error variance was 0.88. Testing the random ANOVA within the random ANCOVA model (i.e., testing that $\beta_1 = \beta_2 = \beta_3 = 0$ in Model 7), produces a chi square of 4706.54 - 3572.24 = 1134.30, which is highly significant with three degrees of freedom (compare the chi square of 1103.44 when comparing the fixed ANOVA and ANCOVA models).

The most general model is the random coefficient analogue of the heterogeneous regression Model 3. It is

$$\mathbf{y}_{ij} = \boldsymbol{\alpha}_j + \boldsymbol{\beta}_{j1} SEX_{ij} + \boldsymbol{\beta}_{j2} IQ_{ij} + \boldsymbol{\beta}_{j3} SES_{ij} + \boldsymbol{\epsilon}_{ij}. \qquad [8]$$

All regression parameters are now random. The random intercept and the random slopes consist of a fixed part and disturbances. These disturbances are again at the group level with expectation zero and independent of the individual error variances ϵ_{ij}. This decomposition is shown in Model 5.

$$\boldsymbol{\alpha}_j = \alpha + \boldsymbol{\gamma}_j, \qquad [9a]$$

$$\boldsymbol{\beta}_{jk} = \beta_k + \boldsymbol{\eta}_{jk}. \qquad [9b]$$

If we substitute these terms in Model 8 we get the equation

$$y_{ij} = (\alpha + \gamma_j) + \\ + (\beta_1 + \eta_{j1})SEX_{ij} + (\beta_2 + \eta_{j2})IQ_{ij} + (\beta_3 + \eta_{j3})SES_{ij} + \epsilon_{ij}. \qquad [10]$$

The number of parameters in this model is 4 (mean regression parameters) + 4 (variance regression parameters) + 6 (covariance regression parameters) + 1 (individual level error variance) = 15. Fitting this model produces an estimate of σ^2 of 0.89, which is not much smaller than the value of 0.91 for the random ANCOVA model. The likelihood ratio chi square for testing the random ANCOVA within the random heterogeneous slopes model is 5.85, which is clearly nonsignificant with 15 - 6 = 9 degrees of freedom. Again, this indicates that there is no significant variation in the slopes in these data. The heterogeneous slopes model basically fits the same structure as the random ANCOVA model, but, because of the additional parameters, it does this with much less stability. Actually, the estimated random slopes (the posterior means of the random effects) show very little variation around the origin, and this seems to have a detrimental effect on the estimation of random intercepts as well.

In Table 3a we show the new orderings obtained with the random coefficient models. The columns are defined in the same way as those in Table 2a. Thus the first two columns are school number and school size, the third one is the random ANOVA (Model 4), the fourth one the random ANCOVA (Model 7), and the last four columns are for the general heterogeneous random slopes (Model 8), with ordering for is, IS, Is, and IS girls. Table 3b gives the correlations between the seven rank orders.

Table 3a
Ranking of Schools by Random Models

Number	SIZE	ANOVA	ANCOVA	is	iS	Is	IS
1	29	6	4	3	3	3	3
1	29	6	4	3	3	3	3
2	33	33	31	31	31	31	31
3	31	30	35	23	25	19	20
4	66	34	32	26	26	25	27
5	39	4	17	6	6	6	6
6	45	24	36	17	17	12	13
7	39	25	29	18	19	18	18
8	31	18	30	27	27	26	25
9	53	29	6	21	20	22	22
10	31	9	16	30	29	30	29
11	30	7	12	7	7	7	7
12	36	20	25	15	16	14	15
13	52	23	13	5	5	5	5
14	29	28	33	29	30	28	28
15	33	16	22	12	12	13	12
16	65	37	37	37	37	36	36
17	57	31	23	22	23	24	26
18	31	11	19	14	15	17	17
19	26	13	28	10	10	8	9
20	27	14	20	24	22	23	23
21	25	5	1	2	2	2	2
22	27	19	9	32	32	32	32
23	26	21	18	20	21	20	21
24	36	2	2	4	4	4	4
25	11	1	5	25	24	27	24
26	27	17	7	8	8	9	8
27	15	26	10	9	9	10	10
28	27	8	11	13	13	15	14
29	20	22	8	1	1	1	1
30	32	12	27	11	11	11	11
31	49	32	26	33	33	33	33
32	57	36	24	36	36	37	37
33	37	27	21	34	34	34	34
34	39	35	34	35	35	35	35
35	35	15	14	28	28	29	30
36	28	10	15	16	14	16	16
37	16	3	3	19	18	21	19

Table 3b
Rank Correlations between Rank Orders

	ANOVA	ANCOVA	is	iS	Is	IS
ANOVA	1.000	.661	.587	.612	.549	.591
ANCOVA	.661	1.000	.532	.560	.454	.491
is	.587	.532	1.000	.997	.990	.992
iS	.612	.560	.997	1.000	.986	.990
Is	.549	.454	.990	.986	1.000	.996
IS	.591	.491	.992	.990	.996	1.000

In Table 4 we compare the six rankings of the fixed model with the six rankings of the random model.

Table 4
Rank Correlations between Rank Orders
Fixed Models in Rows, Random Models in Columns

	ANOVA	ANCOVA	is	iS	Is	IS
ANOVA	.999	.663	.588	.613	.551	.592
ANCOVA	.667	.994	.541	.569	.464	.502
is	-.031	.291	-.371	-.361	-.462	-.442
iS	-.060	.265	-.317	-.301	-.388	.390
Is	.441	.639	.514	.518	.436	.458
IS	.384	.650	.527	.537	.461	.457

It is very clear from Table 3b that the random coefficient model beautifully takes care of the variability of the school level regression coefficients, which caused the low correlations in Table 2b. Slopes really do not make a difference any more, and thus the rank order for our four categories of girls is almost completely identical. We also see the remarkable fact that the four random slope rank orders now correspond more closely with the random ANOVA than with the random ANCOVA solution (which is a far better solution in terms of fit and interpretability). This may be because allowing the slopes to vary forces the estimating procedure to shrink them towards zero, which makes Model 8 like Model 4, and not like Model 7.

Comparing the correlations between some of these "same" models leads to interesting conclusions. Comparing the fixed effect ANOVA with the random effect ANOVA shows a correlation of 0.999. The posterior means are virtually equal to the school means. The same thing is true for the fixed and random ANCOVA models, in which the correlation is 0.994. In Table 4 we once again see the serious defects of the heterogeneous slopes model in the case of fixed effects, and the reasonable performance in the case of random effects (although we then seem to fit the random ANCOVA model in a very inefficient way, making the resulting rank orders closer to random ANOVA, i.e., to the uncorrected means). From the point of view of fit and interpretability, it is clear that the ANCOVA models, both fixed and random, are much preferred. Moreover, they both seem to give basically the same information in a somewhat different form. We will not answer the question of whether either of the two ANOVA models has a <u>better</u> fit, because it is clear that fit measures cannot be compared directly. There is no simple residual variance in the random intercept model, and a direct comparison of likelihoods is also not quite appropriate (because the models are not nested).

Conclusion

There are two important outcomes of our analysis. In the first place we find (again) that variation in the slopes in school effectiveness models is not systematic, and only marginally significant (if at all) in the statistical sense. As De Leeuw and Kreft (1986), Aitkin and Longford (1986), and Raudenbush and Bryk (1986) have found, the random intercept or random ANCOVA model is a better way to present our data. When coefficients are fixed, the application of models in which slopes are allowed to vary can be very misleading because the betas bounce all over the place and lead to wildly different conclusions about the ranking of schools. In the random slopes model the variation in the betas is suitably depressed, but the complicated estimation problems encountered here do not seem to be solved completely. The likelihood surface is presumably very flat. The different ways in which the fixed and random slopes models handle the bouncing beta problem is another important outcome of our analysis.

There are two alternatives that remain if we want to rank schools. The obvious one, using school means, is very stable. It gives virtually the same results for random and fixed models. If we do not interpret it in a purely descriptive way and do think of it as a model-based ranking, then the corresponding model is thoroughly discredited (chi squares near 1100 with three degrees of freedom). If we rank schools in terms of output only, we do not measure their effectiveness, because if we want to measure effectiveness we also have to take input into account. After correcting for input we have a rank order that is quite different, although still moderately correlated with the output rank order.

We have discussed the differences between the random and fixed models and we have given rational arguments for why the random model is a good alternative. The model is built on more realistic assumptions: random effects and random slopes and correlated error terms within groups. Because the random coefficient model is based upon the knowledge of the sampling of schools and the shared history of the students within the same school, the stability of the estimates is increased. Although we cannot show statistically that the random model is preferable to the fixed model, on the basis of appropriateness and parsimony arguments, we think it is preferable to think in terms of random coefficient models. Of course, it still is the responsibility of every individual researcher to consider the choice of her tool. She is the one who is supposed to know if certain assumptions are realistic and if they apply to her situation. She has to choose the tool, as discussed by De Leeuw (1989). In this report we have shown that the choice of the tool can really make a difference.

References

Aitkin, M.A., & Longford, N.T. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*, 1-43.

Burstein, L., Kim, K., & Delandshere, G. (1988). Multilevel investigations of systematically varying slopes: Issues, alternatives and consequences. In R.D. Bock (Ed.), *Multilevel analysis of educational data*. Cambridge, MA: Academic Press.

De Leeuw, J. (1989). *Data modeling and theory construction* (Statistics Series No. 10). Los Angeles: UCLA Institute for Social Science Research.

De Leeuw, J., & Kreft, G.G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*(1), 57-86.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Charles Griffin & Company, Ltd.

Hays, W.L. (1974). *Statistics for the social sciences*. London: Rhinehart and Winston

Jennrich, R.I., & Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 42*, 805-820.

Kreft, G.G. (1987). *Models and methods for the measurement of school effects*. Unpublished doctoral dissertation. University of Amsterdam, The Netherlands.

Kreft, G.G., & De Leeuw, E.D. (1988). The see-saw effect: A multilevel problem? *Quality & Quantity, 22*, 127-137 .

Longford, N.T. (1988a). Fisher scoring algorithm for variance component analysis of data with multilevel structure. In R.D. Bock (Ed.), *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Longford, N.T. (1988b). *VARCL users manual*. Princeton, NJ: Educational Testing Service.

Peschar, J.L. (1983). *School, milieu, beroep* (School, social background, and profession). Groningen: Tjeenk Willink.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, Summer.

Raudenbush, S.W., & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1-17.

Raudenbush, S.W., & Bryk, A.S. (1988). Methodological advances in studying effects of schools and classrooms on student learning. *Review of Research in Education*

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 315-357.

Schluchter, M.D. (1988). *BMDP 5V: Unbalanced repeated measures models with structured covariance matrices*. BMDP Statistical Software, Inc. Los Angeles.