CHAPTER 3

MODEL BASED RANKING OF SCHOOLS

ITA G. G. KREFT* and JAN DE LEEUW†

*Center for the Study of Evaluation
†Departments of Psychology and Mathematics, University of California, Los Angeles, U.S.A.

Abstract

The problem studied in this chapter is the ranking of schools in terms of quality. There are many different ways in which we can quantify if and in how far a school is better than another one. One simple way is to compare means of student outcomes at the end of the school career. But this can be extremely misleading, because it does not correct the differences in school input. If we want to correct for background characteristics of students, in order to find out how well the school does with the input material it receives, then we have to use statistical models in order to make clear on what assumptions our corrections are based.

We first study linear models with fixed regression parameters, such as ANCOVA and models with separate regression parameters for each school. In the first case the regression lines are parallel and the ranking of the schools is the same for all backgrounds. In the second case schools may be ranked quite differently for boys and girls, for blue collar and white collar children, and for high-IQ and low-IQ students. In our example, a 1959 data set of 1290 children in 37 schools in the city of Groningen, it turns out that indeed slopes differ, and thus rankings of schools differ for different backgrounds. We investigate how important and how real this effect is.

More recently random coefficient models have been proposed for school effect analysis by Aitkin & Longford, Raudenbusch & Bryk, Mason, Wong, & Entwistle, De Leeuw & Kreft, and Goldstein. In these random coefficient models schools are considered to be a random sample from a population of schools, and we want to make statements about this population and not necessarily about the individual schools in our study, interpreted as subpopulations. In random coefficient models the residuals in the regression equations are not independent for students in the same school. As a consequence if we predict the outcome for a student in a particular school, we also have to take the other students in the school into account. This means that predictions will generally be more conservative than in fixed coefficient models (the so-called shrinkage to the mean). It is argued in the paper that random coefficient models are more appropriate for school effects analysis. Rankings of the schools in the example are also computed, both for random intercept models (which are the random coefficient version of ANCOVA) and for random slope models (which correspond with nonparallel slopes models). Criteria are discussed which one can use to decide which rank orders are the most appropriate ones.

Introduction

The empirical example analyzed in this paper is from the field of school effectiveness research. The analysis model is a multilevel linear model. These two choices are not accidental. In educational research there is a clear link between substantial and methodological issues, and it is also clear that the validity of statistical inferences is enhanced when appropriate analytical models are used. Thus this chapter 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. We shall see that the rankings do not only depend on the measured properties of the schools, but also on the model chosen to represent the differences between the schools.

In this chapter 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 from very sucessful to very unsuccessful. 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 or not correlated at all. It is obvious that with so many choices a researcher is able to satisfy who ever she wants, just by choosing one model over the many others that will be most pleasing to the audience. 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 ways of schools are ordered. What we need are more objective criteria. Criteria that enable us to choose the best analysis method. *Best* in the sense that it provides us with the most reliable and useful estimates.

In this chapter 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 a starting point we introduce some of the well known traditional linear models, such as analysis of variance and covariance and multiple regression techniques. These three techniques have basically the same underlying assumptions. It will be made clear that these statistical tools can and in fact are applied here to situations for which they are not designed, and this makes them less than optimal. Later in this paper we introduce models that are designed for the school effectiveness situation in educational research. Since school effectiveness research is complicated, this will lead to a more complicated statistical model (see Aitkin & Longford, 1986; De Leeuw & Kreft, 1986; Raudenbush & Bryk, 1986).

Description of the Data

All the different techniques we present have the same goal: to order schools from very successful to very unsuccessful. For our example we use the Dutch GALO data (described by Peschar, 1975). These data contain information about primary school leavers in 37 schools in the city of Groningen in 1959. This is the same data as used in our 1986 article (De Leeuw & Kreft, 1986). In a sense this chapter is an extension of the earlier article, in which we compared different estimation procedures. It is also interesting to compare our

analysis with the similar work in the article of Aitkin & Longford (1986). We also compare different ways of analysis in order to show that these lead to different solutions. To do this we use variables which are measured at the pupil level. The dependent variable is *advice* (of the head teacher about the most appropriate school for a particular student after primary education) with seven categories (for the seven possibilities a student can choose from in secondary education in the Netherlands). The seven categories are scored 1–7, and the resulting variable ADV is used as a numerical variable. This is clearly not exactly appropriate, but given the availability of statistical techniques there is very little else we can do. As predictors for our operationalization of school success we used the three variables SEX, IQ and fathers occupation (SES and six categories, again treated as a numerical variable). Another predictor is the distinction between the 37 schools as 37 groups. The number of students varies per school from as low as 11 to as high as 66 (see Table 3.1a).

Table 3.1a
Ranking of Schools by Variables

##	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 3.1b
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

The First Ordering of Schools by Way of Means

To rank order the schools we start in a very simple way. We simply 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 3.1a gives the rank numbers of the 37 schools on all four variables. Column 1 of Table 3.1a gives the school number, column 2 the school size, and the remaining four columns 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 advice in Table 3.1a is closely related to an ordering by mean IQ or mean SES. This can be seen in Table 3.1b, which gives (Spearman) rank correlations between the columns of Table 3.1a including school size (SIZE).

Observe that Table 3.1b shows a significant positive correlation between SIZE and the three variables IQ, SES, ADV. This is clearly because of schools such as #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.

The data can be used to illustrate another interesting phenomenon. At individual level the (Spearman) rank correlation between ADV and IQ is .743, while the correlation between mean ADV and mean IQ calculated at school level is .881. This *blowing up* of the correlation coefficient when calculated over aggregated variables is a well established fact, known as the *Robinson effect*, named after Robinson (1950), who was one of the first to describe and explain this. Another example is the correlation between SES and ADV, which is .305 at the individual level and is blown up to .714 at the school level. The higher correlations at the aggregate level explain why an ordering by mean ADV is also an ordering by mean SES, but even more so by mean IQ.

It is clear from our results so far that the ordering of schools from highest to lowest mean advice 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 advice 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 if 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 can be formulated in terms of using linear models. This makes it

possible to talk about assumptions, and it also points our various natural alternatives. In our models index j is used for schools, index i is used for students, who are nested in schools. Throughout we use the convention of writing random variables in boldface notation.

The first model we use is:

$$\mathbf{y}_{ij} = \alpha_j + \tau_{ij}, \tag{3.1}$$

where the disturbances are normal, independent, centered, and homoscedastic (this last assumption means that they are assumed to have the same variance σ^2 for each individual). This is the one way analysis of variance (ANOVA) model, in which we have a single parameter α_j for every school. From the linear model point of view (3.1) 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 3.1a.

A method to test if schools by themselves have an influence on achievement is to *partial out* the influence of the individual traits and see if this leaves something to be explained by schools. In other words, does the difference between schools disappear if the influence of the student characteristics is taken into account. There are several ways to do this. The first and most simple one is to perform an analysis of covariance (ANCOVA), where schools are the groups, and where SES, SEX and IQ are the covariates. Model (3.2) is the analysis of covariance model, with the dependent variable advice:

$$\mathbf{y}_{ij} = \alpha_j + \beta_1 \mathrm{SEX}_{ij} + \beta_2 \mathrm{IQ}_{ij} + \beta_3 \mathrm{SES}_{ij} + \tau_{ij}. \tag{3.2}$$

This is a substantial improvement over model (3.1) as far as the unexplained part of the variation is concerned, which is decreased from .815 to .346. The estimated α's can be used again to order the schools. We can use the residual variances, which are respectively 2.07 and .88, to test the difference between models (3.1) and (3.2), i.e., we can test the hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ within the model (3.2). The likelihood ratio chi square is $1290 \times (\ln 2.07 - \ln 0.88) = 1103.44$. With only three degrees of freedom this is clearly highly significant.

If we compare the ANOVA and ANCOVA columns in Table 3.2a we clearly see a different order. The rank correlation between them is $r = .667$, indicating only a moderate agreement between the two orderings. The conclusion so far is that controlling for background characteristics does make a difference. Using only the aggregated means is misleading. But using ANCOVA as the way to avoid a bias in the direction of the school population characteristics has its own problems.

The ANCOVA model is based on at least one critical assumption, which is a priori unlikely to be true. This is the *homogeneity of slopes*, 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 (3.2), but now for each school separately (3.3). Thus model (3.3) allows each school to have its own estimates for the regression coefficients:

$$\mathbf{y}_{ij} = \alpha_j + \beta_{1j} \mathrm{SEX}_{ij} + \beta_{2j} \mathrm{IQ}_{ij} + \beta_{3j} \mathrm{SES}_{ij} + \tau_{ij}. \tag{3.3}$$

Table 3.2a
Ranking of Schools by Fixed Models

##	SIZE	ANOVA	ANCOVA	I_-S_-	I_-S_+	I_+S_-	I_+S_+
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 3.2b
Rank Correlations between Rank Orders

	ANOVA	ANCOVA	I_-S_-	I_-S_+	I_+S	I_+S_+
ANOVA	1.000	.669	−.036	−.062	.434	.383
ANCOVA	.669	1.000	.257	.226	.624	.621
I_-S_-	−.036	.257	1.000	.734	.450	.315
I_-S_+	−.062	.226	.734	1.000	.099	.563
I_+S_-	.434	.624	.450	.099	1.000	.598
I_+S_+	.383	.621	.315	.563	.598	1.000

If we fit this model the residual variance is .78, which means that the percentage of unexplained variance drops to .307. This corresponds to a likelihood ratio statistic (for testing the homogeneity of slopes) of $1290 \times (\ln 0.88 - \ln 0.78) = 155.61$, with $36 \times 3 =$

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, although significant. Although there is some evidence of heterogeneous slopes, it is not very strong.

Let us ignore this statistical information for the moment, and act as if the 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) × 7 (SES) × 85 (IQ from 60–144) = 1190 different conceivable students, the number of possible comparisons is large. To illustrate this we picked, rather arbitrarily, four different types of hypothetical students and ordered the schools according to their success with those students. This produces four different orderings. We code them by using I_A for low IQ, I_+ for high IQ, S_- for low SES, and S_+ for high SES. The first group (code I_A S_+) are girls with an IQ of 90 and a blue collar worker as father (SES-category 2). The second (I_-S_+) are girls with the same IQ, but fathers who are working as businessmen, or who own a small business (SES-category 4). The third (I_+S_-) and fourth (I_+S_+) ordering are girls with the same SES backgrounds, but with the difference that their IQ is now considerably higher (it is equal to 110).

The columns in Table 3.2a give, next to the orderings from models (3.1) and (3.2), the orderings by using model (3.3) in the last four columns as follows:

(I_-S_-) predicted advice $= \alpha_j + \beta_{1j}\,(SEX = 2) + \beta_{2j}(IQ = 90) + \beta_{3j}(SES = 2)$
(I_-S_+) predicted advice $= \alpha_j + \beta_{1j}\,(SEX = 2) + \beta_{2j}(IQ = 90) + \beta_{3j}(SES = 4)$
(I_+S_-) predicted advice $= \alpha_j + \beta_{1j}\,(SEX = 2) + \beta_{2j}(IQ = 110) + \beta_{3j}(SES = 2)$
(I_+S_+) predicted advice $= \alpha_j + \beta_{1j}\,(SEX = 2) + \beta_{2j}(IQ = 110) + \beta_{3j}(SES = 4)$

Replacing the α's and β's with the different estimated values per school produces 37 outcomes for the prediction of advice and thus for the rank order for all schools. The results are indeed different for different types of students, as we can see in Table 3.2a when comparing the last four columns. Each row in Table 3.2a contains the rank numbers for one single school over the six different methods of ordering. Table 3.2b has the (Spearman) rank correlations between the six rank orders.

Some examples of the difference between rankings a school gets if we compare different students in those schools are: schools #1, #21 and #23 (see Table 3.2a). School #1 scores high for the IQ-110 students (columns I_+S_- and I_+S_+) by occupying a third place, but does poorly (28th and 24th place) for the IQ-90 girls (columns I_-S_- and I_-S_+). The same is true for school #21, which scores high for 110-IQ students but does a lot 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 fairly poorly on most scales but is up to 9th and 13th place for higher SES girls (see columns I_-S_+ and I_+S_+). This shows an interaction effect between student characteristics and the school.

The correlations between the orderings, with different students in model (3.3) as the ranking criteria, are moderate to low. The highest correlation is only .734. This is the one between I_-S_- and I_-S_+, 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: $r = .669$. The correlation between orderings for low-SES girls which only differ in IQ, orderings $I_-S_-$and I_+S_- is $r = .450$. The largest discrepancy is between the orderings

I_-S_+ and I_+S_- of girls with different IQ's and with fathers with different occupations, being as low as $r = .099$. The correlations between ANCOVA and the last four orderings is moderately high between the two groups with high IQ girls ($r = .624$ or $r = .621$), and low for the low IQ girls ($r = .257$ or $r = .226$). 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, if compared with the other orderings. Partly this is related to the size of the school. For low IQ students the average predicted advice 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. The orderings given by ANOVA and ANCOVA agree more with the high IQ orderings I_+S_- and I_+S_+. This is unfortunate, since often research in school effectiveness is interested in the success of schools with under-privileged students. In our case, and probably in a lot 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 school are more successful. The orderings based on model (3.3), 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 fits are also bad compared with the ANCOVA model).

But still we encounter problems in using model (3.3). This is already clear from the fact that the differences between the slopes are not very significant (compare the likelihood ratio test earlier), while the ordering of the schools for the various typical individuals we have used is wildly different. If we order schools by way of estimating different models for different schools and compare outcomes we do that by taking the coefficients at face value. By 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 worst on IQ, ADV, SES. If we use ANCOVA to correct for background it moves up one place, but if we let the school determine its own regressions coefficients strange things happen. For high SES students this turns out to be one of the best schools there is, but of course high SES students do not really occur on this school. Taking such things into account leads to 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 independent sampling. We have reasons to doubt the validity of this assumption, which means that students are randomly sampled from an infinitely large population of students. This is shown in the equation of all fixed models where it is stated that the individual errors τ_{ij} are uncorrelated, have a mean of zero and a constant variance σ^2. But 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 good reasons to assume that

individual (student) error terms of students in the same group are correlated. The error term contains next to random measurement error various influences that are not measured, the influence of the variables not in the model (Kreft, 1987). Since students in the same class share a lot of hours per day of common experiences it is somewhat unrealistic to assume (as is done in the fixed models) that unmeasured influences are unsystematic. More realistic is to assume that the error term contains a systematic part which points in the same direction instead of cancelling out by being random.

Another problematic aspect is the use of fixed effect models such as AN(C)OVA. ANOVA as well as 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. Starting from the traditional random effect model and comparing it with the fixed effect model we find that the main difference is that we are no longer dealing with means or point estimators, but with variance components. The variance due to treatment is estimated instead of estimating effects directly by taking differences of treatment means from the grand mean. In random coefficient models it is assumed that the errors 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 correlation between the individuals within the group. This also makes the chance of type I errors in the random model smaller compared to 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 which 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 as fathers and businessmen as fathers, the actual slopes for IQ and SES are very different for different schools. This makes allowing for the possibility that schools have different slopes 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 into 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; Raudenbush, 1988). This results in more efficient and reliable estimates due to a shrinkage factor applied 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 which is a result of small groups (see Raudenbush, 1988). Each LS estimate β_j is weighted proportional 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. Summarizing 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 small schools in the data set are 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 treatments designs (see for instance the SAS module VARCOMP for random analysis of variance model), but these are often not useful in education 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 (1988), 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 overall mean. This makes the number of parameters to be estimated much smaller than in the separate models for the separate schools method. In the last method the schools are considered independent of each other and for each school separate and independent parameters are estimated. In our example of 37 schools, using model (3.3), this leads to $37 \times 4 = 148$ parameters (one intercept and three slopes for each school). With the error variance σ^2 this leads to 149 parameters. In the examples in the next paragraph, 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 leads, in our case, to the estimation of only four means and four variances (for the intercept and the three slopes) plus an individual error variance, altogether giving ten estimates, a lot less than in the fixed model. Some models specify extra estimates for the covariance between the slopes and intercept, which adds maximally a total of 6 covariances to our model and brings the number of parameters to 16, still much less than in the fixed model above. The random coefficient model is more parsimonious 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, that schools are just a sample from a well-defined population and secondly 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 which have the intention to benefit specific minority groups of 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 (3.1), (3.2), and (3.3). The ANOVA model (3.1) becomes:

$$\mathbf{y}_{ij} = \alpha_j + \tau_{ij}. \tag{3.4}$$

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

$$\alpha_j = \alpha + \gamma_j. \tag{3.5}$$

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 τ_{ij}, and of the school level disturbances of other schools. If we substitute (3.5) in (3.4) we find:

$$\mathbf{y}_{ij} = (\alpha + \gamma_j) + \tau_{ij}. \tag{3.6}$$

This implies that $\mathbf{y}_{ij}$ 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 $\rho = \omega^2/(\sigma^2 + \omega^2)$. The model has only three free parameters (α, ω^2, σ^2), in contrast to the 37 + 1 = 38 free parameters in ANOVA (3.1). 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 + 0.38) = 0.15. This deviates significantly from zero.

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

$$\mathbf{y}_{ij} = (\alpha + \gamma_j) + \beta_1 \mathrm{SEX}_{ij} + \beta_2 \mathrm{IQ}_{ij} + \beta_3 \mathrm{SES}_{ij} + \tau_{ij}. \tag{3.7}$$

Only the intercept (the overall effect) is random in model (3.7). This model only needs 6 parameters to be estimated. For the estimates of σ^2 and ω^2 we now find .91 and .04, which is a correlation between errors of children in the same school of only .04 (still significant, though). For the fixed ANCOVA model the estimate of the individual level error variance was .88. Testing the random ANOVA within the random ANCOVA model, i.e., testing that $\beta_1 = \beta_2 = \beta_3 = 0$ in (3.7), 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 analog of the heterogeneous regression model (3.3). It is:

$$\mathbf{y}_{ij} = \alpha_j + \beta_{j1}\,\mathrm{SEX}_{ij} + \beta_{j2}\,\mathrm{IQ}_{ij} + \beta_{j3}\mathrm{SES}_{ij} + \tau_{ij}. \tag{3.8}$$

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 τ_{ij}. This decomposition is shown in (3.5).

$$\alpha_j = \alpha + \gamma_j. \tag{3.9a}$$

$$\beta_{jk} = \beta_k + \eta_{jk}. \tag{3.9b}$$

If we substitute these terms in (3.8) we get the equation:

$$y_{ij} = (\alpha + \gamma_j) + (\beta_1 + \eta_{j1})\,SEX_{ij} + (\beta_2 + \eta_{j2})\,IQ_{ij} + (\beta_3 + \eta_{j3})\,SES_{ij} + \tau_{ij}. \tag{3.10}$$

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 .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 estimating of the random intercepts as well.

In Table 3.3a we show the new orderings obtained with the random coefficient models. The columns are defined in the same way as those in Table 3.2. Thus the first two columns are school number and school size, the third one is the random ANOVA (3.4), the fourth one the random ANCOVA (3.7), and the last four columns are for the general heterogeneous random slopes model (3.8), with ordering for $I_- S_-$, $I_- S_+$, $I_+ S_-$, and $I_+ S_+$ girls. Table 3.3b gives the correlations between the seven rank orders. In Table 3.4 we compare the six rankings of the fixed model with the six rankings of the random model.

It is very clear from Table 3.3b that the random coefficient model beautifully takes care of the variability of the school level regression coefficients, which caused the low correlations in Table 3.2b. Slopes really do not make a difference any more, and thus the rank order for our four types 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, both in terms of fit, and in terms of interpretability). This may be because allowing the slopes to vary forces the estimating procedure to shrink them towards zero, which makes (3.8) like (3.4), and not like (3.7).

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 .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 .994. In Table 3.4 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 both in the fixed and random case the ANCOVA model (or the random intercept model) are much to be preferred. Moreover they both seem to give basically the same information, in a somewhat

Table 3.3a
Ranking of Schools by Random Models

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

Table 3.3b
Rank Correlations between Rank Orders

	ANOVA	ANCOVA	I_-S_-	I_-S_+	I_+S_-	I_+S_+
ANOVA	1.000	.661	.925	.926	.926	.926
ANCOVA	.661	1.000	.645	.647	.644	.647
I_-S_-	.925	.645	1.000	1.000	1.000	1.000
I_-S_+	.926	.647	1.000	1.000	1.000	1.000
I_+S_-	.926	.644	1.000	1.000	1.000	1.000
I_+S_+	.926	.647	1.000	1.000	1.000	1.000

Table 3.4
Rank Correlations between Rank Orders.
Fixed Models in Rows, Random Models in Columns

	ANOVA	ANCOVA	I_-S_-	I_-S_+	I_+S_-	I_+S_+
ANOVA	.999	.663	.928	.928	.928	.928
ANCOVA	.667	.994	.654	.654	.652	.654
I_-S_-	−.031	.291	.000	.002	−.002	.002
I_-S_+	−.060	.265	.061	.064	.060	.064
I_+S_-	.441	.639	.354	.353	.352	.353
I_+S_+	.384	.650	.460	.459	.457	.459

different form. We will not answer the question here if the random or fixed ANOVA model has the better fit of the two, because clearly fit measures cannot be directly compared. 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 & Kreft, Aitkin & Longford, and Bryk & Raudenbush have also found, the random intercept or random ANCOVA model is a better way to present our data. Models in which slopes are allowed to vary can be very misleading, in the case of fixed coefficients, because the betas bounce all over the place and lead to wildly different conclusion about the ranking of schools. In the random slopes model the variation in the betas is suitably depressed, but the complicated estimation problems in this case do not seem to be completely solved here. 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 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 which 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 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, we think on the basis of the appropriateness and the parsimony arguments 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, in the sense of De Leeuw (1989). What we have shown here is, 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 (in press).

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

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

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,* **43**, 805–820.

Kreft, G. G. (1987). Models and methods for the measurement of school effects. 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.

Peschar, J. L. (1975) Milieu-School-Beroep. (*Social background school, and occupation*). Groningen: Tjeenk Willink.

Raudenbush, S. W. (1987). Educational applications for hierarchical linear models. A review. *Journal of Educational Statistics,* **13**, 85–116.

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*, 1988.

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.* Los Angeles: BMDP Statistical Software, Inc.