compared with alternative model formulations in order to reach valid inferences on the subject matter.

9.4 MULTI-LEVEL MODELLING IN AN EDUCATIONAL SURVEY

Multi-level modelling on hierarchically structured data with a continuous response variable is used in a study problem concerning students' literacy in a multinational educational survey. Cluster sampling has been used with schools as clusters, reflecting the hierarchical structure of the population. The sampling design introduces strong intra-cluster correlation for the response variable, and this is a property that should be taken into account in the analysis. The disaggregated approach introduced here provides an alternative to the methods for the nuisance or aggregated approach, which is the main approach in this book. We apply the disaggregated approach by fitting a two-level linear model separately for data from a number of countries. The results are also compared with those from an analysis ignoring the design complexities.

PISA: An International Educational Survey

The data are from the OECD's Programme for International Student Assessment (PISA). The first PISA Survey was conducted in 2000 in 28 OECD member countries and 4 non-OECD countries. The PISA 2000 Survey covered three subject-matter areas: reading literacy, mathematical literacy and scientific literacy. We discuss here the area of reading literacy. We selected from the PISA database the following countries: Brazil, Finland, Germany, Hungary, Republic of Korea, United Kingdom and United States. Our selection of countries is deliberate; countries with varying clustering effects were chosen, keeping, however, in mind a good regional representativeness. The survey data set from these 7 countries comprised a total of 1388 schools and 32 101 pupils.

A highly standardized survey design was used in the PISA 2000 Survey, including standardization of basic concepts, procedures and tools, such as measurement instruments, sampling design, data-collection procedures and estimation and analysis procedures. This was to guarantee as far as possible the international comparability of results.

Sampling of Schools and Students

In the sampling design for an educational survey, it is natural to utilize the existing administrative and functional structures of the school system. There, the schools can be taken as basic units, which are grouped by areas of school administration or

similar administrative criteria. On the other hand, the teaching is organized into teaching groups or school classes, composed of the students and the teacher. In educational surveys, a school is often taken as the primary unit of data collection because of economical and other practical reasons. From the sampled schools, students are selected as the secondary units. There is thus a natural hierarchy in the population, which is a property that is utilized both in the sampling design and in the modelling procedures for this case study.

Stratified two-stage cluster sampling was used in most PISA countries. The first stage consisted of sampling individual schools in which 15-year-old students were enrolled. Schools were sampled with systematic PPS sampling (see Section 3.2), the measure of size being a function of the estimated number of eligible (15-year-old) students enrolled. In most cases, the population of schools was stratified before sampling operations. A minimum of 150 schools was selected in each country (where this number existed), although the requirements for national analyses often required a somewhat larger sample.

In the second stage, samples of students were selected within the sampled schools. Once the schools were selected, a frame list of each sampled school's 15-year-old students was prepared. From this list, 35 students were then selected with equal probability. All 15-year-old students were selected if fewer than 35 were enrolled.

A minimum response rate of 85% was required for the schools initially selected. A minimum participation rate of 80% of students within participating schools was required. This minimum participation rate had to be met at the national level, not necessarily by each participating school (OECD 2001, 2002a).

Weighting Schemes

Appropriate sampling weights were constructed for each national sample data set. The element weight consisted of factors reflecting school selection probabilities, student selection probabilities within schools and school and student nonresponse adjustments. For each country, the weight w_{ik} for student k in school i can be expressed as follows:

$$w_{ik} = w_{1i} \times w_{2ik} \times f_i, \qquad i = 1, \ldots, m \text{ and } k = 1, \ldots, n_i,$$

where

$w_{1i} = 1/(\pi_i \hat{\theta}_i)$ is the reciprocal of the product of the inclusion probability π_i and the estimated participation probability $\hat{\theta}_i$ of school i;

$w_{2ik} = 1/(\pi_{k|i} \hat{\theta}_{k|i})$ is the reciprocal of the product of the conditional inclusion probability $\pi_{k|i}$ and estimated conditional response probability $\hat{\theta}_{k|i}$ of student k from within the selected school i;

f_i is an adjustment factor for school i to compensate any country-specific refinements in the survey design, and
m is the number of sample schools in a given country and n_i is the number of sample students in school i.

The student-level element weights, rescaled to sum up to the actual size of the available sample data set in each country, were used in the analyses. In a given country, the mean of the rescaled weights is one, but there are differences between countries in the variation of the weights. The smallest standard deviation of the rescaled weights is 0.143 and the largest is 0.983. A more detailed description of weighting procedures is given in OECD (2002b).

Reading Literacy in Selected Countries

The outcome variable y is the student's combined reading literacy score (or to be exact, the first of five plausible values of combined reading literacy), scaled so that the common mean over the participating OECD countries is 500 and the standard deviation is 100. We call the response variable the combined reading literacy score. Descriptive statistics on reading literacy in the selected countries are presented in Table 9.8. Means and standard errors of the combined reading literacy score have been calculated by techniques presented in Chapter 5. Therefore, the estimates are design-based and account properly for the complexities (weighting, stratification and clustering) of the sampling design used in a given country. There are two different design effects in the table. The overall design effect accounts for weighting, stratification and clustering. The second design effect

Table 9.8 Descriptive statistics for combined reading literacy score in the PISA 2000 Survey by country (in alphabetical order).

	Combined reading literacy score					Number of observations in data set	
Country	Mean	Standard error	Overall design effect	Design-effect accounting for stratification and clustering	Effective sample size of students	Students	Schools
Brazil	402.9	3.82	8.33	5.17	476	3961	290
Finland	550.7	2.15	2.79	2.74	1600	4465	147
Germany	497.4	5.68	13.47	11.68	305	4108	183
Hungary	485.7	6.02	20.00	16.20	231	4613	184
Republic of Korea	526.6	3.66	12.99	11.67	351	4564	144
United Kingdom	531.4	4.08	14.08	7.16	564	7935	328
United States	517.0	5.16	6.93	5.46	354	2455	112

Data source: OECD PISA database, 2001.

accounts for stratification and clustering and allows for a comparison with the weighted SRS analysis option. Both design effects indicate a strong clustering effect for most countries. In some cases, the difference between the first and second design-effect estimates is substantial, indicating a large variation in the weights.

The effective sample sizes of students are calculated by dividing the number of students by the overall design effect. The effective sample size is the equivalent sample size needed to achieve the same precision in estimation if simple random sampling from a student population without any clustering were used. If the observations are not independent from each other, the effective sample size decreases: the higher the design effect, the smaller the effective sample size. Though the nominal sample sizes of students are large (several thousands) in all countries, some of the effective sample sizes are quite small (only a few hundred).

Design-effect estimates also indicate that standard errors calculated under an erroneous assumption of simple random sampling would be much smaller than the design-based standard error estimates for most countries.

Fitting a Two-level Hierarchical Linear Model

In the analysis, the outcome variable y is the combined reading literacy score. The variation of the outcome variable is explained with two school-level and four student-level variables. The school-level explanatory variables are school size (SSIZE) and teacher autonomy (AUTONOMY). School size is a measure formed from the actual number of students in the school, divided by 100. School principals were asked to report who had the main responsibility for several tasks in the school. Teacher autonomy was derived from the number of categories that principals identified as being mainly the responsibility of teachers. Both variables were standardized so that the common mean over the participating OECD countries was zero and the standard deviation was one.

The student-level explanatory variables are student's gender (recoded so that one is for females and zero is for males, and named FEMALE), socioeconomic background (SEB), engagement in reading (ENGAGEMENT) and achievement press (ACHPRESS). The index of SEB was derived from students' responses on parental occupation. The index of engagement in reading was derived from students' level of agreement with several statements concerning reading habits and attitudes, and the index of achievement press was derived from students' reports of the pressure they feel from their teacher. These three indices were again standardized so that the common mean over the participating OECD countries was zero and the standard deviation was one.

The two-level regression model for the combined reading literacy score y, with explanatory variables and random variation at both levels, is given by

$$y_{ik} = \text{INTERCEPT} + \gamma_1 \times \text{SSIZE}_i + \gamma_2 \times \text{AUTONOMY}_i + \beta_1 \times \text{FEMALE}_{ik} + \beta_2 \times \text{SEB}_{ik} + \beta_3 \times \text{ENGAGEMENT}_{ik} + \beta_4 \times \text{ACHPRESS}_{ik} + u_i + e_{ik},$$

where the index k refers to the level-1 unit (student) and i to the level-2 unit (school). The fixed effects γ and β denote regression coefficients of the school- and student-level variables respectively. Residual u_i is the random effect of school i assumed normally distributed with mean zero and variance σ_u^2, whereas e_{ik} is the student-level residual assumed normally distributed with mean zero and variance σ_e^2. The random effects u_i and e_{ik} are assumed independent. The student-level rescaled weights were used in the analyses.

Units within naturally existing clusters, such as schools, tend to be more similar or homogeneous with respect to the variable of interest than units selected at random from the population. This means that the level-1 units (students) cannot be assumed statistically independent within schools, and the study variable tends to be positively intra-cluster correlated. In the context of multi-level modelling, the intra-cluster correlation is estimated by (Skinner *et al.* 1989; Goldstein 2002; Snijders and Bosker 2002) as

$$\hat{\rho}_{\text{int}} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}^2},$$

where the estimated total variance $\hat{\sigma}^2$ of the study variable is divided into two components, the between-school variance $\hat{\sigma}_u^2$ and the within-school variance $\hat{\sigma}_e^2$. The intra-cluster correlation coefficient measures the pair-wise correlation between values of level-1 units (students) in the same level-2 group (school) and is called the intra-school correlation coefficient. In a model-based context, the coefficient is estimated from the variance components of the null model, i.e. the multi-level model with only intercept and residuals at both levels. For example, the estimated intra-school correlation coefficient for Hungary in Table 9.9 is $6093.7/(6093.7 + 3148.3) = 0.659$. The coefficient can also be estimated from the variance components of the model including explanatory variables, in which case it is called the residual intra-school correlation coefficient. The residual intra-school correlation coefficient for Hungary in Table 9.10 is $4744.2/(4744.2 + 2897.4) = 0.621$. Note that the concept of intra-cluster correlation is used in a design-based context earlier in this book (see Section 3.2).

Variance components were estimated by restricted maximum likelihood (REML), and the fixed effects were estimated by generalized least squares (GLS) given these variance estimates (Bryk and Raudenbush 1992). These estimates are accompanied by standard error estimates that account for the clustering effect (see, for example, the 'sandwich' form in Section 8.4).

Table 9.9 Estimates of two-level variance component models (null models) for combined reading literacy score in the PISA 2000 Survey by country (ordered by the size of the estimated intra-school correlation coefficient).

Country	Intra-school correlation coefficient	Variance components		Intercept	Standard error
		School level	Student level		
Hungary	0.659	6093.7	3148.3	464.1	5.84
Germany	0.553	5572.2	4507.8	496.1	5.61
Brazil	0.428	3146.9	4201.4	387.9	3.61
Republic of Korea	0.375	1828.6	3043.0	520.9	3.74
United States	0.241	2318.2	7315.5	503.3	4.97
United Kingdom	0.212	1917.5	7126.5	529.0	2.88
Finland	0.063	470.7	6960.9	550.6	2.18

Data source: OECD PISA database, 2001.

Table 9.9 presents results for basic two-level variance component models, i.e. null models without explanatory variables. In these models, one fixed effect, the intercept, and the school-level random intercepts are estimated. The total variance is divided into between-schools and within-schools variance components, which are used to calculate the intra-school correlation coefficient. Estimated coefficients vary considerably between the selected countries, with a minimum value of 0.063 and a maximum value of 0.659.

In a given country, the intercept in Table 9.9 is the estimated average of school intercepts. The intercepts are somewhat different from the country means in Table 9.8. Standard error estimates of estimated intercepts are also different because they are calculated using the estimated multi-level model.

Estimated two-level models for combined reading literacy score are presented in Table 9.10. In school-level variables, the effect of school size is statistically significant in some countries. The second school-level variable, teacher autonomy, does not have statistically significant effects in any of the countries.

In student-level explanatory variables, the effects of socioeconomic background and engagement in reading are statistically significant at least at the 5% level in every country. The effect of socioeconomic background varies greatly between countries. The higher the socioeconomic background score, and the more he or she is engaged in reading, the better tends to be his or her reading proficiency score. The strength and direction of the effect of achievement press varies greatly. In most cases, the gender effect was statistically significant.

The estimated models explain a considerable amount of school- and student-level variation in reading literacy as is indicated by the proportional reduction figures. However, there is substantial variation in the degree of reduction gained by the fitted model, when compared to the null model. In most countries, the

Table 9.10 Estimates of two-level models for combined reading literacy score in the PISA 2000 Survey by country.

		Hungary	Germany	Brazil	Republic of Korea	United States	United Kingdom	Finland
Fixed effects:								
Coefficient								
Intercept	γ_0	471.2	496.4	382.0	506.8	496.6	524.9	531.6
	s.e	6.36	4.58	4.56	6.29	6.05	3.38	4.91
	t-test	74.14	108.37	83.75	80.53	82.12	155.06	108.27
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
School-level variables:								
	γ_1	30.6	27.4	2.4	7.1	1.0	3.8	5.9
School size	s.e	9.00	9.22	1.47	3.44	2.54	3.14	7.35
	t-test	3.41	2.97	1.64	2.07	0.38	1.20	0.80
	p-value	0.001	0.003	0.100	0.039	0.705	0.232	0.426
Teacher	γ_2	4.8	−7.1	−3.1	2.5	4.1	−2.3	2.8
autonomy	s.e	5.62	5.22	4.24	5.39	3.63	2.61	2.68
	t-test	0.86	−1.37	−0.74	0.47	1.14	−0.89	1.06
	p-value	0.392	0.171	0.459	0.641	0.256	0.374	0.291
Student-level variables:								
Female	β_1	6.4	3.6	3.1	15.9	14.9	9.8	19.6
	s.e	2.22	2.41	2.54	2.49	3.71	2.64	2.43
	t-test	2.89	1.50	1.21	6.38	4.00	3.71	8.09
	p-value	0.004	0.133	0.228	0.000	0.000	0.000	0.000
Socioeconomic	β_2	6.0	11.5	9.9	2.2	16.7	23.3	15.8
background	s.e	1.09	1.53	1.35	0.92	2.22	1.32	1.34
	t-test	5.56	7.50	7.34	2.40	7.51	17.70	11.78
	p-value	0.000	0.000	0.000	0.016	0.000	0.000	0.000
Engagement in	β_3	19.5	19.0	19.5	16.6	28.9	31.5	33.9
reading	s.e	1.04	0.98	1.51	1.04	1.99	1.40	1.26
	t-test	18.68	19.36	12.87	15.94	14.49	22.59	27.05
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Achievement	β_4	0.9	−1.6	3.4	3.4	−3.3	−7.2	−3.7
press	s.e	0.93	1.16	1.44	0.89	2.04	1.59	1.40
	t-test	0.92	−1.35	2.36	3.85	−1.62	−4.52	−2.65
	p-value	0.356	0.176	0.018	0.000	0.106	0.000	0.008
Random effects:								
Variance component								
School level		4744.2	3501.6	2730.5	1387.3	1770.6	999.6	394.8
Student level		2897.4	3981.9	3830.6	2809.6	6094.1	5779.0	4984.3
Residual intra-school correlation coefficient		0.621	0.468	0.416	0.331	0.225	0.147	0.073
Proportional reduction in variance components, compared to null model (%)								
School level		22.1	37.2	13.2	24.1	23.6	47.9	16.1
Student level		8.0	11.7	8.8	7.7	16.7	18.9	28.4
Total		17.3	25.8	10.7	13.8	18.4	25.0	27.6

Data source: OECD PISA database, 2001.

unexplained school-level variation is still large, compared to the unexplained total variation, which can be seen from the residual intra-school correlation coefficient figures.

Only linear effects of explanatory variables were included in the models. The possible quadratic effects could also be studied for some variables (e.g. school size). All the coefficients of the level-1 explanatory variables are also considered as fixed effects, although there may exist between-school variation in the coefficients, in which case also random coefficient regression models could be used.

Comparison with Weighted SRS Analysis

We finally compare the results of the multi-level modelling exercise with those obtained ignoring the clustering effects. We use the weighted SRS analysis option (see Section 8.2) corresponding to an assumption of independence of the observations. Under this option, a fixed-effects linear model is fitted for the outcome variable, using similar explanatory variables as for the two-level model. Estimation under the weighted SRS option uses the weighted least squares method (see Section 8.4). We selected the German data for comparison (Table 9.11).

The response variable in the German data is highly intra-school correlated, and, as a consequence, the standard-error estimates of the estimated fixed level-2 effects are too small in the model fitted under the weighted SRS option. One of the two school-level effects, teacher autonomy, would be mistakenly considered as statistically significant if the weighted SRS analysis option were used, and the effect of school size would be estimated as being too small. From the level-1 explanatory variables, the effects of socioeconomic background and engagement in reading are much larger compared to the estimates from the two-level model. Achievement press would also appear as a statistically significant effect.

Summary

This case study shows that for data obtained by cluster sampling, an analysis assuming independent observations may be grossly misleading, since the positive intra-cluster correlation of observations will be ignored. Only if the clustering effect were not indicated would the results of an analysis with a two-level model and a weighted SRS-based analysis be similar.

We used here a 'disaggregated' approach in which the hierarchical structure of the population was explicitly modelled by a two-level model. An alternative way to analyse hierarchically structured data is to use design-based methods, as described in Chapter 8. There, instead of modelling the hierarchical structure, the clustering effect induced by the data structure was considered as a nuisance.

Table 9.11 Comparison of estimated coefficients of a two-level model for combined reading literacy score and a fixed-effects model fitted under the weighted SRS analysis option (the German data are used as an example).

Coefficient		Two-level model	Weighted SRS option
Intercept	γ_0	496.4	497.5
	s.e	4.58	1.93
	t-test	108.37	258.08
	p-value	0.000	0.000
School size	γ_1	27.4	20.1
	s.e	9.22	1.74
	t-test	2.97	11.52
	p-value	0.003	0.000
Teacher autonomy	γ_2	−7.1	−7.3
	s.e	5.22	1.38
	t-test	−1.37	−5.26
	p-value	0.171	0.000
Female	β_1	3.6	3.3
	s.e	2.41	2.74
	t-test	1.50	1.20
	p-value	0.133	0.229
Socioeconomic background	β_2	11.5	31.5
	s.e	1.53	1.38
	t-test	7.50	22.9
	p-value	0.000	0.000
Engagement in reading	β_3	19.0	28.9
	s.e	0.98	1.17
	t-test	19.36	24.6
	p-value	0.000	0.000
Achievement press	β_4	−1.6	−4.7
	s.e	1.16	1.31
	t-test	−1.35	−3.64
	p-value	0.176	0.000

Data source: OECD PISA database, 2001.

Thus, in a design-based analysis, we try to 'clean out' the clustering effect from the estimation and testing results to obtain valid inferences.

From a substance matter point of view, the extra contribution of multi-level modelling is that it provides explicit information about the differences between clusters, and thus more information is obtained for the interpretation of the results.