ANNEX A8
TECHNICAL NOTES ON MULTILEVEL REGRESSION ANALYSIS

Method

The multilevel analysis described in Chapter 5 was performed using Hierarchical Linear and Nonlinear Modelling (HLM).[1] A three-level regression analysis was carried out, with students serving as level 1, schools as level 2, and countries/economies as level 3. The model coefficients and statistics were estimated using a full maximum likelihood procedure.[2] Normalised student final weights were used, so that the sum of the weights was equal to the number of students in the dataset, and each country contributed equally to the analysis. Five plausible values for the students' science performance served as the outcome variable.

Data

The data file used for the multilevel analysis included 387 769 students from 14 052 schools in 55 countries/economies.[3] Three data sources were used:

- The PISA 2006 student and school questionnaires for the majority of indicators used in the multilevel analysis.
- *Education at a Glance 2006* (OECD, 2006e), for data on the age of first selection in school systems in the OECD countries.
- The PISA 2006 system-level questionnaire, for additional system-level data in partner countries/economies, concerning in particular the use of standards-based external examinations in science and the age of first selection in the school system. The system-level questionnaire was filled out by the National Project Manager of each partner country/economy.

Data preparation

Selecting and recoding variables

Based on both theoretical considerations and previous empirical findings, several school and system-level explanatory variables were selected in order to examine their association with student performance and the impact which socio-economic background has on student performance. The variables were sorted into six groups:

- Admitting, grouping and selecting
- School management and funding
- Parental pressure and choice
- Accountability policies
- School autonomy
- School resources (human, material and educational)

For each of these six groups, a few variables were selected, mainly from the PISA 2006 database, but also some derived from education system-level information. Indices were preferred over single-item statements whenever they were available since more information could be combined in one index and the problem of measurement error was less severe for indices than for single items (see Annex A1 for examples of indices). For some of the analysis it was possible to choose from several similar variables. In these cases, variables with the lowest rate of missing data and the highest correlation with performance were selected.

Demographic and socio-economic background variables, which are less likely to be amenable to school and system-level factors, were selected based on previous empirical findings. These background variables were included in the net models (that is, models accounting for background factors) in order to examine the net effects of the school and system-level variables. The variables used in the model were:

1. The commercial software HLM 6.04 (developed by Raudenbush, Bryk and Congdon) was used.
2. In this method, both the regression coefficients and the variance components are included in the likelihood function.
3. France and Qatar were not included.

- Student level
  - PISA index of economic, social and cultural status (ESCS) and its squared value
  - Gender
  - Language spoken at home
  - Immigrant status
- School level
  - School average PISA index of economic, social and cultural status
  - School size and its squared value
  - School location
- System level
  - Country average PISA index of economic, social and cultural status

These selected background and explanatory variables were re-coded where necessary. The descriptive statistics for all variables are listed in Box A8.1. The variables with "X" or "Y" in the first or second place are school-level or system-level variables, respectively. A detailed SPSS syntax for recoding variables is included in Annex A9, available on line at *www.pisa.oecd.org*.

Box A8.1 **Descriptive statistics of background and explanatory variables**

Variable description	Type	Variable name	Mean	S.D.	Min.	Max.	% missing
STUDENT LEVEL							
Student's PISA index of economic, social and cultural status (ESCS)	B	ESCS; MESCS	-0.22	1.08	-5.67	3.35	0.95%
Student's PISA index of economic, social and cultural status squared	B	ESCS2	1.22	1.81	0.00	32.14	0.95%
Student is female	B	FEMALE	0.50	0.50	0.00	1.00	<0.00%
Student has no immigration background (student and parents were born in the country of assessment)	B	NATIVE; MNATIVE	0.88	0.33	0.00	1.00	2.47%
Student speaks the test language or other national language most of the time or always at home	B	SAMELANG; MSAMELANG	0.93	0.26	0.00	1.00	3.44%
SCHOOL LEVEL							
School located in a small town or village (fewer than 15 000 people)	B	XRURAL; MXRURAL	0.33	0.47	0.00	1.00	1.49%
School located in a city (with over 100 000 people)	B	XCITY	0.36	0.48	0.00	1.00	1.49%
School size	B	XSCHSIZE; MXSCHSIZE	8.47	7.44	0.03	100.00	2.61%
School size squared	B	XSCHSIZ2	127.10	382.99	0.00	10000.00	2.61%
School average PISA index of economic, social and cultural status	B	XESCS; MXESCS	-0.22	0.74	-3.67	1.97	<0.00%
Admitting, grouping and selecting							
School with ability grouping for all subjects within school	E	XABGR; MXABGR	0.19	0.40	0.00	1.00	3.33%
School with low academic selectivity of school admittance	E	XLOSELE	0.32	0.47	0.00	1.00	2.33%
School with high academic selectivity of school admittance	E	XHISELE; MXHISELE	0.23	0.42	0.00	1.00	2.33%
School management and funding							
School being privately managed	E	XPRIVMAN; MXPRIVMAN	0.18	0.39	0.00	1.00	3.11%
School with high proportion of school funding from government sources	E	XGOVFUND; MXGOVFUND	82.75	26.77	0.00	100.00	7.47%
Parental pressure and choice							
School with high level of competition	E	XSCHLCOMP; MXSCHLCOMP	0.73	0.45	0.00	1.00	2.66%
School with high levels of perceived parental pressure	E	XPRESSPA; MXPRESSPA	0.65	0.48	0.00	1.00	2.53%
Accountability policies							
School informing parents of children's performance relative to other students in the school	E	XACC1; MXACC1	0.61	0.49	0.00	1.00	2.31%
School informing parents of children's performance relative to national benchmarks	E	XACC2; MXACC2	0.45	0.50	0.00	1.00	2.84%
School informing parents of students' performance relative to other schools	E	XACC3; MXACC3	0.28	0.45	0.00	1.00	3.20%
School posting achievement data publicly	E	XACC4; MXACC4	0.37	0.48	0.00	1.00	2.92%
School using achievement data for evaluating principals	E	XACC5; MXACC5	0.37	0.48	0.00	1.00	3.69%
School using achievement data for evaluating teachers	E	XACC6; MXACC6	0.55	0.50	0.00	1.00	2.93%
School using achievement data for allocating resources to schools	E	XACC7; MXACC7	0.35	0.48	0.00	1.00	3.52%
School with achievement data tracked over time	E	XACC8; MXACC8	0.66	0.47	0.00	1.00	3.33%
Any accountability variable missing		MXACC					7.82%
School autonomy							
School autonomy in staffing	E	XFACS; MXFACS	0.00	1.00	-1.24	1.66	1.15%
School autonomy in budgeting	E	XFACB; MXFACB	0.00	1.00	-2.31	0.87	1.12%
School autonomy in educational content	E	XFACC; MXFACC	0.00	1.00	-1.93	1.09	1.11%
Any school autonomy variable missing		MXFAC					1.16%

. . .

School resources							
School average number of students per teacher (student-teacher ratio)	E	XSTRATIO; MXSTRATIO	14.75	6.69	0.27	100.33	6.84%
School-level index of teacher shortage	E	XTCSHORT; MXTCSHORT	0.05	1.06	-1.06	3.62	2.01%
School average number of computers for instruction per student	E	XIRATCOMP; MXIRATCOMP	0.12	0.12	0.00	1.47	4.32%
School-level index of quality of school educational resources	E	XSCMATED; MXSCMATED	-0.28	1.13	-3.43	2.14	1.62%
School average students' learning time for regular lessons in school	E	XLTSCTOT; MXLTSCTOT	10.24	2.37	0.00	21.00	0.31%
School average students' learning time for out-of-school lessons	E	XLTOSTOT; MXLTOSTOT	2.77	1.57	0.00	13.00	0.31%
School average students' learning time for self-study or homework	E	XLTSTTOT; MXLTSTTOT	5.27	1.64	0.00	19.00	0.31%
School providing opportunity of taking science	E	XANYSCIE; MXANYSCIE	81.73	22.29	0.00	100.00	0.27%
School average index of school activities to promote students' learning of science	E	XSCIPROM; MXSCIPROM	0.23	1.01	-2.27	1.64	1.79%
SYSTEM LEVEL							
Country average PISA index of economic, social and cultural status	B	YESCS	-0.22	0.51	-1.52	0.77	-
Admitting, grouping and selecting							
System with early selection (each additional year between the first age of selection and the age of 15)	E	YYRSSEP	1.20	1.62	0.00	5.00	-
System-level number of school types and distinct educational programmes available to 15-year-olds	E	YNRTRACK	2.33	1.21	1.00	5.00	-
Parental pressure and choice							
System with high proportion of competitive schools	E	YSCHLCOM	74.61	16.15	27.81	98.76	-
Accountability policies							
System with standards-based external examinations	E	YSCENTEX	0.56	0.47	0.00	1.00	-
Percentage of observations with at least one missing value on one variable							28.21%

Note: In the second column "B" refers to background variable and "E" refers to explanatory variable. For the computation of the percentages of missing values, equal country weights were used. The first letter "M" in variable names signifies a missing dummy. As only four (unweighted) cases are missing in the variable "female", the missing value was imputed to 0 and a missing dummy was not created for this variable.

StatLink http://dx.doi.org/10.1787/142050165315

Treatment of missing data

The proportion of missing values for the variables considered in the analysis is presented in Box A8.1. Even though the missing rate was less than 5% for most of the variables, a listwise deletion of all observations that have a missing value for at least one variable would have reduced the sample size by 28.21%, since more than 30 variables were included in the models. Therefore, missing values were imputed in order to include the maximum number of cases in the analysis.

Since the missing rates were not high for most of the variables, a simple imputation approach was used to circumvent the problem of missing data: predictors at the individual and school level were imputed using a dummy variable adjustment (Cohen and Cohen, 1985). Due to the small number of observations (n=55) at the system level, system-level variables were not imputed.

It is known that this imputation method generally produces biased estimates of coefficients (Jones, 1996), and that standard errors of those variables that contain missing values are underestimated since they do not account for the uncertainty introduced through imputation. However, given the fact that only on 2 of 33 variables, more than 5% of the data were missing (Box A8.1), this bias was considered negligible.

As a first step of the imputation, a so-called "missing dummy" variable was created for all variables with missing values regardless of whether a variable was continuous, categorical or dichotomous. A missing dummy variable was set to 1 if the data were missing on that variable and it was set to 0 if the data were not missing. The first letter "M" in variable names in Box A8.1 signifies a missing dummy.

As a second step, missing values were imputed for continuous variables. Missing values were replaced by the weighted school average of the variable. If all data on the respective variable were missing in one school such that the weighted school mean could not be computed, the weighted country mean was imputed. If all data on the respective variable were missing in a country, the weighted international mean was imputed. When a missing value was replaced by the country or school mean, the weights were proportional to the sampling probability (weighting factor W_FSTUWT from the PISA 2006 dataset). When a missing value was replaced by the international mean, equal country weights were used, *i.e.* each country was given an equal weight of 1 000 cases.

Categorical variables were re-coded into a set of dummy variables.[4] For each category or for combined categories, a dummy variable was created with the value of 1 if the observation belongs to the respective category and 0 otherwise. Missing values in dummy and dichotomous variables were replaced by 0.

Student weights

For the multilevel analysis, data files were weighted at the student level with "normalised student final weights", which were computed based on the student final weights (W_FSTUWT)[5] in the PISA 2006 dataset. Normalised student final weights were developed at two different levels according to the purpose of the analysis:

- *At the country level for the two-level regression analysis* The student final weights (W_FSTUWT) were normalised at the country level to *i)* make the sum of the weights within each country equal to the number of students within the country in the dataset (*i.e.* the sample size of the country); and *ii)* maintain the same proportion of weights as in the student final weights (W_FSTUWT within each country).
- *At the international level for the three-level regression analysis* The student final weights (W_FSTUWT) were normalised at the international level including 55 of the 57 participating countries to *i)* make the sum of the weights across the 55 countries equal to the number of students across the 55 countries in the dataset; *ii)* maintain the same proportion of weights as in the student final weights (W_FSTUWT) within each country; and *iii)* ensure that each individual country's contribution to the analysis is equal by introducing a country factor (*i.e.* the sum of the weights within each country is the same for all 55 countries).

The SPSS syntax for computing the normalised student final weights is found in Annex A9, available on line at *www.pisa.oecd.org*.

Modelling student performance

This section outlines the modelling strategy used in the multilevel analysis of school and system-level variables related to educational performance.

For building the multilevel model, a step-by-step approach was adopted, starting from the student level upwards to the country level, following an approach suggested by Raudenbush and Bryk (2002). This resulted in the background variables at all three levels listed above and the background model presented in Box A8.2 (see Model 0b in Table 5.19g).

Box A8.2 **Background model for student performance**

Level-1 Model

$$Y = P0 + P1*(ESCS) + P2*(MESCS) + P3*(ESCS2) + P4*(FEMALE) + P5*(NATIVE) + P6*(MNATIVE) + P7*(SAMELANG) + P8*(MSAMELAN) + E$$

Level-2 Model

$$P0 = B00 + B01*(XESCS) + B02*(MXESCS) + B03*(XRURAL) + B04*(XCITY) + B05*(MXRURAL) + B06*(XSCHSIZE) + B07*(XSCHSIZ2) + B08*(MXSCHSIZ) + R0$$

Level-3 Model

$$B00 = G000 + G001(YESCS) + U00$$

Note: See Box A8.1 for the definition of the variables.

4. The number of dummy variables created from a categorical variable is smaller than the number of categories of the variable since one or more categories are used as a reference group.

5. The detailed description of student final weights is in *PISA 2006 Technical Report* (OECD, forthcoming).

Throughout the multilevel analysis, unless otherwise specified, an effect is considered statistically significant if the p-value is below 0.1 at country level and below 0.005 at school level. Different criterion values were chosen for the two levels to balance between significance and statistical power. In particular, at the country level, where there are only 55 cases, statistical power is rather low, which is why a higher significance level was chosen. In contrast, there are more than 14 000 observations at the school level and so a rather low significance level of 0.005 was chosen.

Model 0b in Table 5.19g summarises the results for the background model, *i.e.* the model in which only the background variables are considered.

Explanatory variables considered

The impact of selected system and school-level variables on science performance was analysed using multilevel models, before and after accounting for the demographic and socio-economic background variables.[6] As described in an earlier section of this annex, the six groups of variables were selected based on both theoretical considerations and previous empirical findings. One or more variables were selected as indicators measuring each of these groups. All background and explanatory variables for each of the six groups are listed in Box A8.1.

In the analysis of the impact of system and school-level variables on learning outcomes, a two-step procedure was applied. The system and school-level variables listed in Box A8.1 were grouped into six groups: Admitting, grouping and selecting; school management and funding; parental pressure and choice; accountability policies; school autonomy; and school resources (human, material, and educational).

In the first step, the effects of the variables of each of the six groups were examined in turn, estimating separate models for each group. In the second step, from each of the separate models run in the first step, only the significant predictors were combined in a model. When a predictor turned out to be no longer significant in the combined model, it was dropped from the analysis.

This two-step procedure was applied following the model specification suggested by Raudenbush and Bryk (2002), as well as by Snijders and Bosker (1999). The opposite, backward approach of entering all possible predictors at a time, and then removing the non-significant ones was not feasible due to the large number of variables and multicollinearity problems.[7]

The final net combined model is shown in Box A8.3. The results of this model are presented in the Model 2N in Table 5.19g as well as in Box 5.8 in Chapter 5.

Box A8.3 **Final net combined model for student performance**

Level-1 Model

$$Y = P0 + P1*(ESCS) + P2*(MESCS) + P3*(ESCS2) + P4*(FEMALE) + P5*(NATIVE) + P6*(MNATIVE) + P7*(SAMELANG) + P8*(MSAMELAN) + E$$

Level-2 Model

$$P0 = B00 + B01*(XESCS) + B02*(MXESCS) + B03*(XRURAL) + B04*(XCITY) + B05*(MXRURAL) + B06*(XSCHSIZE) + B07*(XSCHSIZ2) + B08*(MXSCHSIZ) + B09*(XLTSTTOT) + B010*(XLTSCTOT) + B011*(XLTOSTOT) + B012*(XSCIPROM) + B013*(MXSCIPRO) + B014*(XFACB) + B015*(MXFACB) + B016*(XACC4) + B017*(MXACC4) + B018*(XLOSELE) + B019*(XHISELE) + B020*(MXHISELE) + B021*(XABGR) + B022*(MXABGR) + B023*(MXLTTOT) + R0$$

Level-3 Model

$$B00 = G000 + G001(YFACB) + G002(YESCS) + U00$$

Note: See Box A8.1 for the definition of the variables.

6. A gross model is defined as the model without accounting for the background variables. A net model is defined as the model accounting for the background variables.

7. Multicollinearity exists when two or more independent variables are highly correlated.

Besides the final net combined model for student performance depicted here, which is a model where background variables are included, the same two-step analysis strategy was used for gross models, *i.e.* models where background variables are not accounted for. The results of the final gross combined model are depicted in the Model 2G in Table 5.19g as well as in Box 5.8 in Chapter 5.

Fixed/random effects and centring

In the models for examining the impact of selected system and school-level variables on science performance, all slopes were fixed and only the intercept was randomised at all three levels.

All variables including both background and explanatory variables were centred on the grand mean. The grand mean centring is a linear transformation of variables by subtracting the overall mean of all 55 countries from the value proper. Note that, for fixed slopes, it does not make a difference for the estimated slope whether a variable is grand-centred or not centred. Only the interpretation of the intercept changes when centring by the grand mean. In all models, the intercept is to be interpreted as the achievement score in science for a student who has the international mean in all variables included in the model.

Modelling the impact of socio-economic background on student performance

In investigating the roles that variables at the school and system level play with respect to the impact which socio-economic background has on student performance, a two-step procedure similar to the models for student performance was conducted. In the first step, the effects of the variables for each of the six groups listed above were examined in turn. Then, in the second step, only the variables with statistical significance in the first step were included in the combined model.

To look at the net effects of these factors, background variables at the student, school, and system levels were included in the analyses. The background variables are exactly the same as those used in the models described in the preceding section. The detailed background model specification is presented in Box A8.4.

In Box A8.4, the first letter "M" in variable names signifies a missing dummy and "X" and "Y" in the first or second place denote school and system-level variables, respectively. The only difference to the first set of models (models for student performance) is that the ESCS slope and the XESCS slope were given a random slope.

In Table 5.20g, the Model 0b summarises the results for the background model for the impact of socio-economic background on student performance, *i.e.* the model in which only the background variables are considered without any explanatory variables.

Box A8.4 **Background model for the impact of socio-economic background**

Level-1 Model

$$Y = P0 + P1*(ESCS) + P2*(MESCS) + P3*(ESCS2) + P4*(FEMALE) + P5*(NATIVE) + P6*(MNATIVE) + P7*(SAMELANG) + P8*(MSAMELAN) + E$$

Level-2 Model

$$P0 = B00 + B01*(XESCS) + B02*(MXESCS) + B03*(XRURAL) + B04*(XCITY) + B05*(MXRURAL) + B06*(XSCHSIZE) + B07*(XSCHSIZ2) + B08*(MXSCHSIZ) + R0$$

$$P1 = B10 + R1$$

Level-3 Model

$$B00 = G000 + G001(YESCS) + U00$$

$$B01 = G010 + U01$$

$$B10 = G100 + U10$$

Note: See Box A8.1 for the definition of the variables.

Explanatory variables considered

In the first step of the analysis, the variables at the school level were added to the background model for each of the six groups separately, estimating the slope of the student-level PISA index of economic, social and cultural status, as well as the intercept. As an example, Box A8.5 contains the group of the impact of socio-economic background for the group of school resources variables. All variables from the school resources group were introduced to the equations signifying the slope of the student-level PISA index of economic, social and cultural status (P1), in addition to the intercept (P0). The focus is on the coefficients (and corresponding significance values) for the slope, P1. The missing dummies for the variables (starting with the letter M) were included only in the estimation of the intercept, but not in the slope estimation.

Box A8.5 **Model of the impact of socio-economic background: "school resources"**

Level-1 Model

Y = P0 + P1*(ESCS) + P2*(MESCS) + P3*(ESCS2) + P4*(FEMALE)
+ P5*(NATIVE) + P6*(MNATIVE) + P7*(SAMELANG) + P8*(MSAMELAN) + E

Level-2 Model

P0 = B00 + B01*(XESCS) + B02*(MXESCS) + B03*(XRURAL) + B04*(XCITY)
+ B05*(MXRURAL) + B06*(XSCHSIZE) + B07*(XSCHSIZ2) + B08*(MXSCHSIZ)
+ B09*(XSTRATIO) + B010*(MXSTRATI) + B011*(XTCSHORT) + B012*(MXTCSHOR)
+ B013*(XIRATCOM) + B014*(MXIRATCO) + B015*(XSCMATED) + B016*(MXSCMATE)
+ B017*(XLTSTTOT) + B018*(XLTSCTOT) + B019*(XLTOSTOT) + B020*(XANYSCIE)
+ B021*(MXANYSCI) + B022*(XSCIPROM) + B023*(MXSCIPRO) + B024*(MXLTTOT) + R0

P1 = B10 + B11*(XSTRATIO) + B12*(XTCSHORT) + B13*(XIRATCOM) + B14*(XSCMATED)
+ B15*(XLTSTTOT) + B16*(XLTSCTOT) + B17*(XLTOSTOT) + B18*(XANYSCIE)
+ B19*(XSCIPROM) + R1

Level-3 Model

B00 = G000 + G001(YESCS) + U00

B01 = G010 + U01

B10 = G100 + U10

B110 = G1100

Note: See Box A8.1 for the definition of the variables.

The model for the group of accountability practices variables is presented in Box A8.6. The variables at the system level, such as the existence of a standards-based external examination (SCENTEXA) were included in the estimate of the intercept (B00) and of two slopes, namely the slope of the student-level PISA index of economic, social and cultural status (ESCS; B10) and the slope of school-level PISA index of economic, social and cultural status (XESCS; B01). The coefficients of interest are G011 and G101.

Box A8.6 Model of the impact of socio-economic background: "accountability policies"

Level-1 Model

$$Y = P0 + P1*(ESCS) + P2*(MESCS) + P3*(ESCS2) + P4*(FEMALE) + P5*(NATIVE) + P6*(MNATIVE) + P7*(SAMELANG) + P8*(MSAMELAN) + E$$

Level-2 Model

$$P0 = B00 + B01*(XESCS) + B02*(MXESCS) + B03*(XRURAL) + B04*(XCITY) + B05*(MXRURAL) + B06*(XSCHSIZE) + B07*(XSCHSIZ2) + B08*(MXSCHSIZ) + B09*(XACC1) + B010*(XACC2) + B011*(XACC3) + B012*(XACC4) + B013*(XACC5) + B014*(XACC6) + B015*(XACC7) + B016*(XACC8) + B017*(MXACC) + R0$$

$$P1 = B10 + B11*(XACC1) + B12*(XACC2) + B13*(XACC3) + B14*(XACC4) + B15*(XACC5) + B16*(XACC6) + B17*(XACC7) + B18*(XACC8) + R1$$

Level-3 Model

$$B00 = G000 + G001(YSCENTEX) + G002(YESCS) + U00$$

$$B01 = G010 + G011(YSCENTEX) + U01$$

$$B10 = G100 + G101(YSCENTEX) + U10$$

Note: See Box A8.1 for the definition of the variables.

In the second step of the modelling procedure of the socio-economic impact, the variables that were statistically significant in the first-step estimations were jointly submitted to the analysis, yielding the combined model contained in Box A8.7.

Box A8.7 Final combined model for the impact of socio-economic background

Level-1 Model

$$Y = P0 + P1*(ESCS) + P2*(MESCS) + P3*(ESCS2) + P4*(FEMALE) + P5*(NATIVE) + P6*(MNATIVE) + P7*(SAMELANG) + P8*(MSAMELAN) + E$$

Level-2 Model

$$P0 = B00 + B01*(XESCS) + B02*(MXESCS) + B03*(XRURAL) + B04*(XCITY) + B05*(MXRURAL) + B06*(XSCHSIZE) + B07*(XSCHSIZ2) + B08*(MXSCHSIZ) + B09*(XLTSCTOT) + B010*(MXLTSCTO) + R0$$

$$P1 = B10 + B11*(XLTSCTOT) + R1$$

Level-3 Model

$$B00 = G000 + G001(YYRSSEP) + G002(YESCS) + U00$$

$$B01 = G010 + G011(YYRSSEP) + U01$$

$$B10 = G100 + G101(YYRSSEP) + U10$$

Note: See Box A8.1 for the definition of the variables.

The results from the final combined model are given in Table 5.20g (Model 2) as well as in Box 5.9 in Chapter 5.