Methodology for Conducting Country-Level Longitudinal Analyses: A Review and Comparison of Procedures

Kajsa Yang-Hansen & Jan-Eric Gustafsson
Department of Education, University of Gothenburg

Abstract

One of the most challenging tasks in secondary analysis of data from international studies is to make credible causal inference about the impact of different factors on educational achievement. The trend design of the major IEA studies has provided a basis for conducting longitudinal studies at the country level, and this approach has shown great promise in preliminary analyses based upon simple methods of analysis. However, within other scientific fields there is greater experience with these kinds of approaches. The purpose of the paper is to review methods for country-level longitudinal analysis, such as the time-series cross-section method used in political science and methods for panel analysis used in econometrics. The different methods are illustrated using data from the PIRLS studies.

Keywords: PIRLS, country-level comparison, time-series cross-section model, panel data, STATA

Background and Study Purposes

One of the most challenging tasks in secondary analysis of data from international studies is to make credible causal inference about the impact of different factors on educational achievement. This is because these data typically are cross-sectional. In such data selection bias and omitted variables form two main threats to valid causal inference. Selection bias implies that subgroups of cases are not comparable with respect to entry-level characteristics, which for example may be due to compensatory or competitive resource allocation strategies, or to self-selection into different groups. The problem of omitted variables implies that the independent variable under study is correlated with other variables, which are the actual causal factors.

As was demonstrated by Gustafsson (2007) a simple analysis may be performed when there are only two waves of measurement, in which change in the level of achievement at country level is related to change in one or more independent variables. With this approach many of the problems related to omitted variables can be avoided. This is because most country-level characteristics keep constant over time, which cause them to disappear when change over time is investigated. The problem of selection bias also disappears when data are analyzed at country level, because there are no mechanisms of compensatory or competitive resource allocation or self-selection that operate across country boarders.

However, this analytical approach is statistically unsophisticated because it does not take into account the variability in the country level observations that is due to sampling and measurement error. The limited number of observations at country-level may also cause the analysis to have low power. Another problem is that it is not easily applicable to situations when there are observations at more than two time points.

Within other disciplines, such as political science and econometrics, country level data for a large number of countries and a large number of time points have been available, and in these disciplines methodology has been developed to deal with such data. One purpose of the paper is to review these techniques, with a particular eye towards their applicability to analyses of the IEA trend data at country level, and another purpose is to make empirical comparisons between what seems to be the most useful techniques.

Methodology and Data Sources

Data

The current study relies on data from the 28 school systems from 26 different countries that took part in both PIRLS 2001 and PIRLS 2006 (Mullis, Martin, Kennedy & Foy, 2007). Table 1 presents the number of students participating in the two PIRLS studies.

Insert Table 1 about here

All together the two studies comprised some 242 000 students, fairly evenly distributed over the two occasions. Like in the PIRLS international report (Mullis et al., 2007), the results from Kuwait were not included in the longitudinal analysis.

The current study focuses on a limited set of variables. The total reading scores from the 2001 and 2006 assessments are, of course, of central interest. These measures are available in the form of 5 plausible values, from which a mean has been computed for each individual (READTOT).

Most of the analyses focus one single independent variable, namely the age of the students. Gustafsson (2007) reported analyses of this variable using the TIMSS 1995 and 2003 data. Analyses of the relation between student age and achievement within each of these two cross-sectional data sets did not show any correlation between age and achievement. This non-intuitive result indicates that these analyses are influenced by omitted variables that disturb the positive relation between age and achievement. For example, an important determinant of the age variability at the time of testing is the school start age, which varies over groups of countries. It is thus reasonable to assume that this variable is associated with a large set of cultural and educational factors, which may bias the relations between age and achievement. However, in both the TIMSS grade 4 and grade 8 data analyses of the relation between change in achievement and change in mean student age between 1995 and 2003 resulted in quite strong positive estimates, agreeing with results from other approaches such as the between-grade regression discontinuity design. This was interpreted as supporting the general approach to analyze country-level change. One reason for this is that the variability in student mean age over time is determined by non-systematic factors, such as at which particular point in time it is suitable to conduct the survey within different countries. If the variability over time in the independent variable is more or less random, this variation cannot be correlated with other variables, which should make it possible to interpret effects of the age variable in causal terms.

Table 2 presents the country level means of the reading score and the students' mean age for the two occasions. The values in the table agree with those presented in the international report (Mullis et al., 2007).

Insert Table 2 about here

For most of the countries the mean student age is about the same for the two testing occasions. However, in Russia the students were half a year older in 2006 than in 2001, and in Singapore they were a quarter of a year older. For some countries the 2006 samples had a mean age that was at least two months lower than the mean for the 2001 samples (Morocco, Iran, Hong Kong, Lithuania and Norway). Given the results obtained by Gustafsson (2007), it may be expected that these age differences between the samples account for some of the differences in reading performance between the two occasions.

Methods of analysis

As has already been made clear the primary aim of the current paper is to compare different methods of analysis for dealing with data from country-level comparisons. These data are longitudinal at the country level, but they are not longitudinal at the individual level. This unbalanced structure makes it impossible to analyze these data with methods normally used for analyzing longitudinal data, such as growth curve modelling (e. g., Singer. & Willett, 2003), because these methods assume that the same set of individuals is followed over several occasions of measurement.

However, within other scientific fields, methods have been developed to deal with unbalanced longitudinal data. Within political science, for example, much attention has since the 1980s been focused upon what is called the time-series cross-section model (TSCS). This seems primarily to be due to an important paper by Stimson (1985), which argued that comparisons between countries should be combined with analyses of development over time. He pointed out that,

> Pooling data gathered across both units and time points can be an extraordinarily robust research design, allowing the study of causal dynamics across multiple cases, where the potential cause may even appear at different times in different cases. Many of the possible threats to valid inference are specific to either cross-sectional or time-serial design, and many of them can be jointly controlled by incorporating both space and time into the analysis. (Stimson, 1985, p. 916,)

The TSCS approach offers strengths in analyses of causal relations, and several statistical techniques have been applied in the analysis of such data. However, “for pooled analyses, insofar as they are known at all, are known for special statistical problems.” (Stimson, 1985, p. 916). Moreover, the TSCS design in the field of political science typically is applied to data with a large number of time points (normally more than 15) and in the IEA data there are only a few time points.

Within the field of econometrics designs similar to the TSCS design have also been used, which typically are referred to as panel designs. A commonly used method for analyzing data from two time points is to use a so called differences-in-differences approach (DID, see e. g., Meyer, 1995). The basic idea of the DID approach is to attribute the change of average outcome in the treatment group over a time interval to the change in the control group. The DID estimates focus on the differences over time, which solves the problem of omitted variables and selection effects in the ordinary regression method with cross-sectional data (see also Gustafsson, 2008).

Depending, among other things, upon the nature of the variables involved, a DID analysis of the IEA data can be implemented in different ways. Suppose that we have measured a continuous independent variable at two occasions (e. g., AGE01 and AGE06) along with a continuous dependent variable at two occasions (e. g., READT01 and READT06). One possibility is to aggregate all the information to the country level, and then compute the differences AGE06-AGE01=AgeDiff and READT06-READT01=ReadDiff. We then regress ReadDiff on AgeDiff to get a straightforward implementation of the DID approach. In this analysis we take away all the characteristics of the countries which remain the same between then two occasions of measurement, and in this way many of the omitted variables are controlled for. Econometricians refer to this kind of analysis as a “fixed unit” analysis.

This straightforward implementation, however, may become tedious when several measurement occasions are included. In addition, while this analysis is simple both conceptually and technically, it has the disadvantage that the aggregation of the data to the country level implies a loss of information. Thus, in our case we only have 28 observations,

which are treated as equal in spite of the fact that the number of individuals upon which these are based varies over countries.

There is, however, another way to implement a fixed unit analysis at country level, while still analyzing individual data. To do that we first combine the data from the two occasions in a single file, putting each individual on a single row, all individuals having the same variables (e. g., AGE and READTOT). In order to identify the two occasions of measurement, a dummy variable is added for each individual, which has the value of 0 for the observations from 2001 and the value 1 for the observations from 2006. Another set of dummy variables is added to identify which particular country each observation comes from. For the PIRLS data we thus need 27 dummy variables to identify the 28 countries (see, e. g., Cohen, Cohen, West & Aitken, 2003). We can then perform an analysis in which READTOT is regressed upon the age variable aggregated to the country level and we also enter the occasion dummy and the country dummies. This also is called fixed-effects model, and the cross-country differences over time is captured by the differences in the intercept of the least squares dummy estimation (see e. g., Hsiao, 2003). In other words, by including the dummy variables (i. e., fixed effects) the average differences across countries in any of the predictors are controlled for. This analysis should yield approximately the same regression coefficient as the analysis based on aggregated data, but the standard errors will be dramatically different.

It is known that DID analyses tend to underestimate the standard errors, because there is correlation between individuals belonging to the same unit. However, there is also the problem that in the IEA studies cluster sampling typically is used. Under such sampling designs, the actual amount of information is a function not only of the number of individual observations, but also of the cluster size and the amount of intra-cluster similarity, as expressed by the intraclass correlation. These dependencies among observations can be corrected for in different ways, such as with the so called sandwich estimator (or Huber-White estimator, see Huber, 1967, White, 1982) implemented, for example, in STATA and Mplus (Muthén & Muthén, 1998-2007). This estimator does not make distributional assumption nor does it assume the model specification being correct. It investigates the actual amount of dependencies among observations within explicitly defined clusters, such as schools or classes, and corrects the estimated standard errors accordingly. It is often called a cluster robust estimator of standard errors Yet another approach to take the cluster structure of the data into account is to conduct a multilevel analysis (see, e. g., Snijders & Bosker, 1999).

There is also a third main approach to analyze these data, which involves application of multilevel analysis. This approach is based on what is called random coefficients modeling, which implies that a slope at individual-level varies across higher level unit, and such variation is captured by a random coefficient variable (see e. g., Hox, 2002). In the current context, the countries are not considered to be a set of fixed units, but rather as a sample. For this sample of countries, characteristics such as the mean and standard deviation of READTOT may be estimated with a latent variable model, and the estimate of the regression of the latent variable on AgeDiff, or a latent variable capturing the change in age, may also be computed.

These different approaches may be implemented in different ways, and the models may be estimated with different software systems. In the current paper, we explore similarities and differences among the different analytic approaches. For the approaches analyzing individual data we investigate different ways of dealing with the dependencies among observations caused by the sampling design and the panel design. We also compare the behavior of different computer programs, focusing in particular on the STATA and Mplus.

Findings and Discussion

Below we report the main findings from application of the different analytical approaches. The results from the different models are summarized in Table 3.

Take in Table 3 about here

Regression of aggregated difference scores

In the first step of the analysis the simple and intuitive method of aggregating all data to the country level was applied, in the same way as was done by Gustafsson (2007). Thus, the values of ReadDiff presented in Table 2 were regressed upon AgeDiff in an ordinary least-squares regression analysis (Model 1). This analysis resulted in an unstandardized regression coefficient of 47.7. The standard error was 15.4, and the t-value 3.10 ($p <. 005$). The beta-coefficient (or the correlation between ReadDiff and AgeDiff) was 0.52, which implies that the country differences in age over time accounted for 27 % of the country differences in level of reading performance over time. This is quite a substantial effect, and it agrees quite well with what was found in the analyses of TIMSS data by Gustafsson (2007).

The unstandardized coefficient implies that an age change of one year is associated with an increase in achievement of 47.7 points. This estimate reflects the combined effect of becoming one chronological year older and going to school one more year. It is interesting to observe that in the Swedish PIRLS 2001 data, which included samples from both grade 3 and grade 4, there was a performance difference of 42 score points, which agrees reasonably well with the regression estimate. There is, however, a subtle difference between the meanings of these estimates. The difference in level of achievement between grades three and four results from the combined effect of a twelve-month chronological year and a school year that is typically around nine months. In contrast, the unstandardized regression coefficient is based on the observed age variation within a school year, which, when expressed in terms of the expected effect of a one-year difference, captures the combined effect of a twelve-month chronological year and twelve months of schooling. We thus would expect the regression estimate based on the regression procedure to be somewhat higher than the observed performance difference between two adjacent grades, which is indeed the case.

As has already been pointed out the analytical approach based on aggregated difference scores has a number of disadvantages. In this analysis no account is taken of the fact that the sample sizes differ greatly between countries and occasions. Furthermore, the aggregation causes loss of information, because data on more than 240 000 individuals is being reduced to data on 28 countries. The analysis also assumes a simple linear relation between measures of change in achievement and change in the independent variables. It is also not quite clear how analyses of different subgroups, such as boys and girls, or different socio-economic groups, could be accomplished with this approach.

Fixed unit analysis with dummy variables (Fixed-Effects Models)

In the second step models were fitted in which the fixed unit analysis was implemented through analyzing individual data, with a vector of dummy variables to represent fixed country effects. While the dependent variable was measured at the individual level, the independent age variable was aggregated to the country level separately for each occasion. In addition, the individual level age variable was included in some models. Data were analyzed with STATA 10, and cases were, when possible, weighted with HOUWGT.

There are several regression commands available in the STATA. The REGRESS command is the basic procedure for conducting regression analysis and it was used first. In Model 2 the independent variables were aggregated age, a dummy variable to represent occasion

(PIRLS2006 where observations from 2006 had the value 1, and observations from 2001 had the value 0), and one dummy variable for each participating country except one.

The unstandardized regression coefficient for country-level age was 45.9, with SE=2.23 and t=20.55. The regression coefficient is close to the estimate obtained in Model 1, but the standard error is much lower. The small standard error is largely due to the fact that the regression procedure applied to individual data severely underestimates the amount of uncertainty in the parameter estimate, because dependencies among observations are not properly taken into account. This is primarily because the data have been sampled with a cluster model, typically based on the school as a unit. Since the students within a school tend to be more similar to one another than any two randomly sampled students, this causes loss of information. One way in which this cluster effect can be controlled is to empirically determine the actual amount of dependencies in the data and compute cluster robust standard errors on the basis of this information. This can be done with the STATA REGRESS command, through adding the VCE option, along with a cluster variable.

VCE (*Cluster Variable*) specifies that the observations are independent across clusters but not necessarily within clusters. VCE option affects only standard errors and the variance-covariance matrix of the estimates, not the estimated coefficients. The cluster variable used here is a school variable, which uniquely identifies the 9219 schools involved in the two samples. Using the VCE option (Model 3) left the parameter estimates unchanged, but caused the SE for the aggregated age variable to increase from 2.23 to 7.39 (i. e., by a factor of 3.3). There was, thus, a very substantial effect on the estimate of the SE for the aggregated age variable.

We could also consider extending the regression model by adding individual level variables. One example of such a variable could be the individual age variable (ASDAGE). Adding this model to the regression model (Model 4) gave a highly significant negative regression coefficient for the individual age variable of -16.5, while the regression coefficient for the country level age variable increased to 62.4. Thus, when the individual age variable was added, the parameter estimate for the aggregated variable was strongly affected. It may, however, be doubted that this is a correctly specified model. One indication of this is that the individual level age variable is so strongly negatively related to achievement. This is probably because this variable is influenced by selection effects which cause an overrepresentation of older students with a low level of performance. However, it does not make much sense to have such non-causal effects, which probably also differ over countries, included in a model which has as its main aim to avoid the influence of selection effects. The individual level age variable will therefore not be included in any further model.

Multilevel mixed effects models

Other, more advanced regression models, such as multilevel mixed effects model, also are available, which mix estimation of fixed parameters, as in the ordinary regression model, with estimation of random effects. Estimation of random effects implies that parameters of distributions, such as mean and standard deviation (sd) at the higher level are estimated, rather than the individual parameter values. Such models are extremely useful to approach data with a nested hierarchical data structure, such as when schools are sampling units within countries.

The STATA system offers the XTMIXED command to estimate models which involve random effects. Regrettably, however, with this command, case weights cannot be applied, and cluster robust standard errors are not available. Nevertheless, we will for our methodological purposes demonstrate use of this command.

It may first of all be noted that if we apply the XTMIXED command in the same way as the REGRESS command in Model 3 (except weights), not specifying any nested observational structure, we do get identical estimates as with REGRESS (Model 4). For the aggregated age variable the unstandardized regression coefficient was 45.2, so we may note that use of case weights does not in this case influence the results in any important manner.

Normally, however, we would use the XTMIXED command in such a way that we specify a nested observational data structure. One way to take advantage of this possibility is to keep the fixed-unit, dummy-variable, specification, but add the information that schools are nested under countries. In this model (Model 5) the parameter estimate of country-level age increased to 50.1 (SE = 5.64).

We can also treat the countries as a sample, rather than as a set of fixed units, by specifying that both countries and schools are nested. In this model (Model 6) the regression coefficient for country-level age was 46.6 (SE = 5.50). It is interesting to note that the parameter estimate in this model, which assumes countries to be a random sample, is very close to what was obtained in the fixed-unit, dummy variable regression model.

The XTREG command in STATA program also offers support for analysis of longitudinal and panel data. This command makes it possible for modeling trends and changes over time under different assumptions. It thus is possible to analyze data under, among other things, a fixed-effects model and a random-effects model. However, limitations also apply to these models. Thus, under the random-effects model cluster robust standard errors are not available, and case weights cannot be applied. However, under the fixed-effects model cluster robust standard errors can be obtained, but not case weights. It is, therefore, interesting to compare the results obtained when we use XTREG in this way with the results from the other procedures.

From this model (Model 7) the estimate of the regression coefficient for the country-level age variable was 45.2 (SE = 6.56). Here too we obtain estimates which are close to the results of the other methods. In fact, we can see that the estimated regression coefficient is identical with what was obtained with the dummy-variable REGRESS command without case weights. It also may be noted that the standard error obtained here is larger than what has been obtained with the other random-effects methods, and close to what was produced by the fixed-unit, dummy-variable method with cluster robust standard errors.

Two-level modeling

The different commands for estimating random-effects models in STATA are useful for dealing with data with a nested observational structure. However, they still are limited in the sense that only models with a single dependent variable may be estimated. In contrast, the Mplus program offers a set of multilevel, multivariate, modeling procedures, which allow a considerably greater flexibility. Thus, these methods have a potential for analyzing country-level longitudinal data. Given the complexity of these procedures, the principles and procedures are described in some detail.

To apply these procedures the data should be organized in the “long” format, in the same manner as when fixed units are analyzed with the dummy variable approach. A cluster variable must be identified, which in this case is a variable identifying country. In the present example there thus are 28 clusters.

The dependent variable is change in achievement, and in the Mplus implementation of the random coefficient model, we can estimate the mean and variance of the latent distribution of country level change through invoking a latent variable, which functions as a “container” variable. To define the latent container variable we use a special statement in Mplus, which

may be written as follows:

```
readch | READTOT ON PIRLS06;
```

This statement introduces the latent variable *readch* through regressing READTOT on the dummy variable PIRLS06 which has a value of 1 for those individuals who took in PIRLS 2006 and a value of 0 for those who took part in PIRLS 2001. The regression coefficient expresses the amount of change within each country when estimated at the country level and the mean and variance of the *readch* variable expresses the estimated mean and variance of the change for the 9 countries. Estimating a model with this statement only, yielded a mean of 3.0 and a variance of 207.2 in the *readch* variable. The variance of *readch* was significant (t=3.38).

This model may be extended through adding for each country an aggregated variable expressing the change in mean age (AgeDiff). Regressing the latent *readch* variable on AgeDiff (Model 8) a regression coefficient of 47.0 (SE = 15.4) was obtained, with a t-value of 3.06. The residual variance of *readch* was 152.6, which implies that 26 % of the variance in *readch* was explained by AgeDiff.

In this model AgeDiff was entered as a fixed, manifest, variable. However, with the random coefficient modeling approach allowed by Mplus, we can also estimate the change in age at country level from individual data, using the same latent variable modeling approach as was used to estimate *readch*. We can then regress the *readch* variable upon this *agech* latent variable. In this model (Model 9) the regression coefficient for *readch* on *agech* was 47.5 (SE = 15.3), and the residual variance in *readch* was 151.2. It is also quite interesting to note that the estimate of the regression of *readch* on *agech* is very close to the estimates obtained with the fixed unit, dummy variable approach.

So far we have not commented the standard errors. However, given that the Mplus model used above does not take into account dependencies among observations that are due to cluster effects, we may suspect the standard errors to be underestimated in this model. With the Mplus 5 program this problem can, in principle, be solved through using the complex sample option, specifying school as a cluster variable. However, when this was attempted for the current models, the estimation procedure failed to converge.

We can note, however, that not only the parameter estimates, but also the standard errors in these two-level analyses are very similar to those obtained in the aggregated country-level analysis of difference scores (Model 1). What is done in the two-level analysis is indeed something highly similar to the aggregated analysis, the major difference being that we estimate random coefficients with latent variables rather than compute observed differences in aggregated scores. Above all, in both analytic approaches it is recognized that the number of observations is the same as the number of countries. Furthermore, with the large sample sizes available here, there is great stability in the country-level estimates, which causes the differences in results between the approaches to be small. This also makes it reasonable to expect that taking the effects of cluster sampling within countries into account will not appreciably affect the estimates of the standard errors of the estimated regression coefficient because these are only to a very small extent affected by a change in the estimate of the precision of the country-level means. The fact that Mplus failed to estimate the model which corrects for the effects of the cluster sampling is thus probably not of any importance, because the standard errors obtained in the regular analysis may be expected to be very much the same.

Discussion

As has already been observed there is little variation in the estimated regression coefficients

for the different models. It may be noted also that inclusion of weights has only a marginal effect on the parameter estimates. It also is interesting to note the both the parameter estimates and the standard errors of the very simple analysis in Model 1 agree perfectly with the results of the quite complicated analysis in Model 9.

While the parameter estimates do not vary much over the different analytic models, the standard errors do. It is quite obvious that in the models with low estimated standard errors the estimates are biased, because dependencies among observations are not accounted for. When this is done, either through computing cluster robust standard errors (Models 3 and 7) or through estimating random effects associated with schools (Models 5 and 6), the standard errors increase by a factor around 3. The effects are somewhat stronger when the former approach is taken than when the latter approach is used.

Still, however, the standard errors in Models 3 and 7 are only about half as large as those obtained in the country-level analysis, and it may be asked if there are other factors which cause bias in the standard errors estimated from individual level data. The fact that there are repeated observations on the same set of countries is one such source of dependencies among observations which may not properly be accounted for in the fixed-unit analytical models. It may, however, also be the case that the analyses conducted on the basis of individual level data yield a higher level of power. This issue should be clarified in further research.

The analyses conducted at the country level (Models 1, 8 and 9) have the widest margins of error. There is little reason to suspect that these are biased, even though they may be on the conservative side, because the data is not fully exploited in these analyses. However, given the need to guard against the risk of falsely rejecting the null hypothesis in these kinds of analyses, there may be reason to stay with safe procedures.

Conclusions and Implications

The results of this study indicate that the simple, intuitive, method of doing the analysis on variables expressing differences between two time points and which are aggregated to the country level is a useful approach, which gives the same results as the more complex two-level random coefficient approach. However, the simple method is limited to deal with situations when there are only two points of measurement, while the two-level approach can be extended in various ways. Furthermore, in spite of its technical complexity, the two-level random coefficient approach is quite simple conceptually which makes it attractive. Another advantage of this approach is that it allows fitting of relations at individual level as well. Thus, the two-level random coefficient approach may be recommended as a main tool for conducting longitudinal country-level analyses.

The results presented here also indicate that the regression model controlling for fixed units with dummy variables, and correcting the standard errors for cluster effects, is a useful approach for dealing with country-level longitudinal data. Within the STATA system it seems that the standard regression command has advantages for conducting these kinds of analyses, primarily because it supports case weights and because it computes the cluster robust standard errors. It is, indeed, somewhat paradoxical that the more advanced analytical models implemented with the XTREG and XTMIXED commands are more limited in these respects than is the standard REGRESS command.

Given that more and more data suitable for longitudinal analyses at country level are becoming available it will be quite interesting to see which substantive results can be obtained with these analytical methods.

References

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd Ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gustafsson, J.-E. (2007). Understanding causal influences on educational achievement through analysis of differences over time within countries. In T. Loveless (Ed.) *Lessons Learned: What International Assessments Tell Us about Math Achievement*, (pp. 37-63). Washington, DC: The Brookings Institution.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal*, *7*(1), 1-17..

Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge University Press.

Hox, J. J. (2002). *Multilevel Analysis: Techniques and Applications.* Lawrence Erlbaum Associates, Mahwah, NJ.

Huber, P. J. (1967). The behavior of maximum likelihood estimation under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, LeCam, L. M. and Neyman, J. editors. University of California Press, pp. 221- 233.

Meyer, B. D. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics, 13*, 151-161.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007) *IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthén, B. O. & Muthén, L. (1998-2007). Mplus User's Guide.

Singer., J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Stimson, James (1985). Regression in Space and Time: A Statistical Essay. *American Journal of Political Science, 29,* 914–947.White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1- 25.

Table 1. Number of participating students

	PIRLS 2001	PIRLS 2006
Bulgaria	3460	3863
Canada, Ontario	5252	3988
Canada, Quebec	3001	3748
England	3156	4036
France	3538	4404
Germany	7633	7899
Hong Kong SAR	5050	4712
Hungary	4666	4068
Iceland	3676	3673
Iran, Islamic Republic	7430	5411
Israel	3973	3908
Italy	3502	3581
Latvia	3019	4162
Lithuania	2567	4701
Macedonia, Rep of	3711	4002
Marocko	3153	3249
Moldova, Rep of	3533	4036
Netherlands	4112	4156
New Zealand	2488	6256
Norway	3459	3837
Romania	3625	4273
Russia	4093	4720
Scotland	2717	3775
Singapore	7002	6390
Slovak Republic	3807	5380
Slovenia	2952	5337
Sweden	6044	4394
United States	3763	5190
Total	114382	127149

Table 2. Country level means for the dependent and independent variables

	Readtot01	Readtot06	ReadDiff	Age01	Age06	AgeDiff
Bulgaria	550	547	-3.5	10.9	10.9	-0.05
Canada, Ontario	548	555	7.5	9.9	9.8	-0.12
Canada, Quebec	537	533	-4.3	10.2	10.1	-0.15
England	553	539	-13.4	10.2	10.3	0.08
France	525	522	-3.6	10.1	10.0	-0.10
Germany	539	548	8.5	10.5	10.5	-0.07
Hong Kong SAR	528	564	36.0	10.2	10.0	-0.19
Hungary	543	551	7.7	10.7	10.7	-0.01
Iceland	512	511	-1.8	9.7	9.8	0.07
Iran, Islamic Republic	414	421	7.1	10.4	10.2	-0.23
Israel	509	512	3.5	10.0	10.1	0.06
Italy	541	551	10.7	9.8	9.7	-0.15
Latvia	545	541	-3.7	11.0	11.0	-0.08
Lithuania	543	537	-6.4	10.9	10.7	-0.20
Macedonia, Rep of	442	442	0.8	10.7	10.6	-0.03
Marocko	350	323	-26.9	11.2	10.8	-0.38
Moldova, Rep of	492	500	8.1	10.8	10.9	0.02
Netherlands	554	547	-7.1	10.3	10.3	-0.01
New Zealand	529	532	2.9	10.1	10.0	-0.02
Norway	499	498	-1.2	10.0	9.8	-0.18
Romania	512	489	-22.2	11.1	10.9	-0.16
Russia	528	565	36.8	10.3	10.8	0.48
Scotland	528	527	-0.8	9.8	9.9	0.08
Singapore	528	558	30.3	10.1	10.4	0.26
Slovak Republic	518	531	12.7	10.3	10.4	0.02
Slovenia	502	522	20.0	9.8	9.9	0.06
Sweden	561	549	-11.7	10.8	10.9	0.05
United States	542	540	-2.2	10.2	10.1	-0.13

Table 3. Results from different models

Model	Estimate	Standard error	t-value
1. Regression on aggregated difference scores	47.7	15.40	3.10
2. Regression, country dummies, country means for age, weights	45.9	2.23	20.55
3. Regression, country dummies, country means for age, weights, robust se	45.9	7.39	6.21
4. Regression, country dummies, country means for age, no weights	45.2	1.85	24.41
5. Mixed-effects,, country dummies, country means for age, no weights, random schools	50.1	5.64	8.88
6. Mixed-effects, country means for age, no weights, random countries, random schools	46.6	5.50	8.47
7. XTREG, country means for age, no weights, fixed countries, robust se	45.2	6.56	6.89
8. Two-level, random coefficients, fixed age, weights	47.0	15.40	3.05
9. Two-level, random coefficients, weights	47.4	15.34	3.09