EXTENDING THE SCOPE OF ANALYSING DATA OF IEA STUDIES: APPLYING MULTILEVEL MODELLING TECHNIQUES TO ANALYSE TIMSS DATA

Leonidas Kyriakides, University of Cyprus, Cyprus
Charalambos Charalambous, University of Cyprus, Cyprus

Abstract

The traditional approach in analysing IEA data has been to use single-level statistical models. This paper argues that multilevel modelling offers a wider spectrum in analysing IEA data because of multi-stage sampling, since students are nested within classes, classes within schools, and schools within countries. Two illustrative examples are used to provide support to this argument. In the first example, TIMSS 1999 data concerning students' achievement and their self-perceptions in two subjects (i.e., Mathematics and Science) were analysed using single and multi-level analyses. While single-level analysis resulted in positive correlations at the micro-level and negative correlations at the macro-level, the findings of multi-level analysis did not support such contradictory relationships. In the second example multi-level modelling techniques were used to analyse TIMSS data related to student achievement in Mathematics. A large number of variables taken from the student, teacher and school questionnaires and which were associated with the theoretical models of Educational Effectiveness Research were added to the model, explaining about 40% of the variance of student achievement. The results of the study are discussed in terms of the importance of using multi-level modelling in analysing IEA data. Methodological issues concerning the design of comparative studies in education are also raised.

INTRODUCTION

Since 1959, the International Association for the Evaluation of Educational Achievement (IEA) has undertaken a large number of comparative studies in different subjects. The ultimate goal of these studies has been to isolate those factors related to student learning which could be manipulated through policy changes in curriculum, resource allocation or instructional practice (Martin, 1996; Yang, 2003). It has been expected that information that arises from such investigations could help policy-makers, curriculum specialists and researchers better understand the performance of their educational systems (Martin et al., 2000; Mullis et al., 2000). Thus, the weaknesses and strengths of educational systems at micro or macro level could be identified and intervention programs attempting to improve educational effectiveness could be developed. Specifically, Schmidt & Valverde (1995) argue that:

> *By looking at the educational systems of the world we challenge our own conceptions, gain new and objective insights into education in our own country, and are thus empowered with a fresh vision with which to formulate effective educational policy and new tools to monitor the effects of these new policies* (p. 7).

In this perspective, one could identify common research interests between IEA studies and those related to the Educational Effectiveness Research (EER), since both attempt to identify factors at pupil, teacher and school level which are associated with student achievement. However, it was not until the end of 1980s that statisticians developed multi-level modelling techniques for analysing hierarchical data, which could help researchers in either of these two domains identify teacher and school effectiveness factors.

This paper argues that multilevel modelling, a methodology for the analysis of data with a focus on nested sources of variability, offers a wider spectrum in analysing IEA data. The first part of the paper raises issues related to the importance of using multilevel techniques in analysing IEA's data, whereas the second part provides an example where single and multi-level analyses of the same set of TIMSS data result in different relations among student achievement (i.e., the dependent variable) and exploratory variables situated at the macro level (i.e., country level). The third part provides an illustrative example of analysing TIMSS data by using multi-level modelling in order to identify variables at different levels which are associated with student achievement. Finally, suggestions concerning the design of comparative evaluative studies are provided.

STATISTICAL TREATMENT OF CLUSTERED DATA: THE USE OF MULTILEVEL ANALYSIS

A common procedure in social research with two-level data is to aggregate the micro-level data to the macro-level. The simplest way to do this is to work with the averages for each macro-unit. There is nothing wrong with aggregations when the researcher is only interested in macro-level propositions, although it should be taken into account that the reliability of an aggregated variable depends, among other things, on the number of micro-level units in a macro-level unit. However, in cases where the researcher is interested in macro-micro or micro-level propositions, aggregation may result in gross errors, such as the "shift of meaning" (Huttner & van den Eeden, 1995) and the "ecological fallacy" (Alker, 1969).

The first potential error arises from the fact that a variable being aggregated to the macro level refers to the macro units and not directly to the micro units. For instance, although student intelligence is an index of the extent to which he/she is able to perform a learning task, the classroom average of student intelligence comprises an index of how demanding the intended curriculum should be in order to meet the learning needs of the classroom. The second potential error has to do with the fact that a correlation between macro-level variables cannot be used to make assertions about micro-level relations. For example, the average classroom ability may be negatively related to the frequency of praising students, since teachers tend to provide positive feedback more frequently to low achieving pupils (Brophy, 1992). However, this does not provide any clue about the micro-level (i.e., student-level) relationship between frequency of providing positive feedback and achievement. On the contrary, studies within educational psychology have shown that positive feedback may improve student achievement (Walberg, 1986). Aggregating data also prevents researchers from investigating cross-level interaction effects. For instance, it does not allow us find out whether the effect of aptitude on achievement varies according to the extent to which ability grouping of pupils within classroom is used (Kyriakides, 2004). It is also not possible to examine how variables measured at one level affect relations occurring at another (Bryk & Raudensbush, 1992).

However, the most significant error is the neglect of the original data-structure (Raudenbush & Bryk, 1986). Single-level analyses require the researcher to assume incorrectly that individuals within similar subunits share no common characteristics. Such an approach leads to the possibility of biased regression coefficients and associated standard errors. Specifically, the group (e.g., class) and its members (e.g., students) both influence and are influenced by the group membership. Students bring certain skills and attitudes with them; at the same time they are clustered in classes and schools with certain characteristics. Therefore, treating students as if they were independent of the class and school grouping ignores the complexity in the data (Snijders & Bosker, 1999). To ignore this relationship risks overlooking the importance of group effects, and as a consequence, important relationships may be overlooked and erroneous conclusions may be drawn (Heck & Thomas, 2000). This can be attributed to the fact that individuals in a group or context tend to be more similar on many important variables than individuals in different contexts.

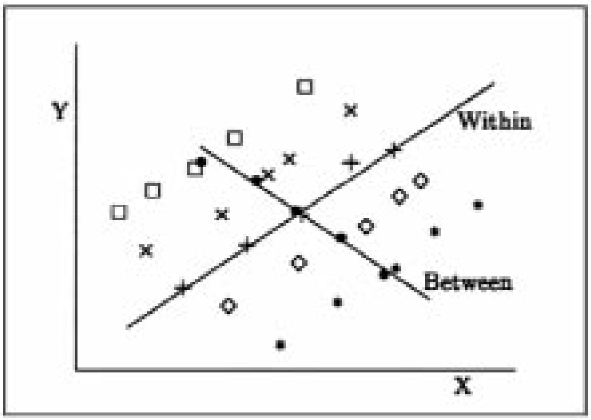
Multi-level modelling techniques, a methodology for the analysis of data with complex patterns of variability, can meet the deficiencies of single-level analysis mentioned above. Specifically, multi-level analysis explicitly models the manner in which students are grouped within classes or schools and, therefore, has several advantages (Goldstein, 2003). First, multi-level analysis takes into account the hierarchically structured data and the variability associated with each level, since there is variability between students as well as between classes. One may draw wrong conclusions if any of these sources of variability is ignored (Snijders & Bosker, 1999; Opdenakker & Van Damme, 2000). Second, multi-level analysis provides a means of partitioning the outcome variable's variance into different levels (within and between units). This enables researchers to compare the teacher and school effects. Third, it yields better-calibrated estimates for the variance of standard errors. Fourth, it offers a single framework that combines the information within and across units to produce more accurate explanations and outcomes. Finally, clustering information provides correct standard error confidence intervals and significance tests, which are more conservative than the traditional ones that are obtained simply by ignoring the presence of clustering. By allowing the use of covariates measured at any of the levels of a hierarchy, the researcher is able to explore the extent to which the differences in average achievement results between schools are accountable for by factors related either to the school, class or student characteristics.

COMPARING RESULTS EMERGING FROM SINGLE AND MULTI-LEVEL ANALYSIS: AN EMPIRICAL INVESTIGATION USING TIMSS DATA

To further illustrate the importance of using multi-level analysis in comparative studies, in this section we focus on the most significant error of conducting single-level analysis (i.e., ignoring the hierarchical structure of data and thus using aggregated data). First a theoretical example is provided, drawn from Snijders and Bosker (1999). To show that such an error is not hypothetical, data from TIMSS 1999 are analysed using both single and multi-level analyses and different results are obtained.

Figure 1 presents the situation of five groups, for each of which five observations are taken. The groups are indicated by the shapes □, x, +, ◊ and *, and the five group means are indicated by ● .

Figure 1: Micro-level Versus Macro-level Relations Of The Five Groups Measurements (copied with permission from Snijders and Bosker, 1999, p. 14)



Analysis at micro-level reveals that the regression line of Y on X has a positive slope and that all observations of groups indicated by □ and x are above the regression line, whereas the observations of the groups ◊ and * are below the line. If someone aggregates the data and regresses the average $\bar{Y}$ on average $\bar{X}$ for each group (depicted by •), a regression line with a negative slope emerges. We can also observe that the averages of all groups are almost perfectly on the regression line, thus leading to the impression that there are almost no differences between the five groups' average $\bar{Y}$

after adjusting for the average $\bar{X}$. Snijders and Bosker (1999) conclude that "although the situation [as described above] is an idealized example, it clearly shows that working with aggregated data is dangerous at best and disastrous at worst" (p. 15).

Obviously, a question that arises is whether this example could represent a real situation. A review of various secondary analyses of TIMSS data enable us to identify an example where a single level analysis was used, resulting in positive relations at micro-level and negative relations at macro-level (Shen, 2002). Specifically, the author used a data set related to student achievement in Mathematics and Science and three measures of their self-perception (i.e., how much they like the two subjects, their self-perceived competence in the subjects and their perceived easiness of the subjects). Correlation analyses were conducted at the student level for each of the 38 participating countries to examine the relationships between achievement in the two subjects and the three measures of self-perceptions. For most countries, the correlation coefficients were positive, except those referring to the relations of perceived easiness in Science and achievement which were found to be negative. Then correlation analyses at the country level were conducted for the two subjects separately, resulting in high negative significant correlation coefficients (i.e., below -0.60 in the case of Mathematics and below -0.45 in the case of Science).

The aforementioned data set was also used in the present study. However, cases with incomplete data were eliminated from further analysis, since multilevel analysis cannot cope with missing data (Goldstein, 2003). Specifically, three countries in the case of Mathematics (i.e., Jordan, Morocco, and Turkey) and four countries in the case of Science (i.e., Jordan, Morocco, Finland and Bulgaria) were eliminated from subsequent analysis, since less than 85% of their student data were complete. Data from the Netherlands were also discarded, since there was no data related to students' liking the two subjects. Correlation analysis of achievement and self-perception measures of the remaining data set on Mathematics (n=151168) and Science (n=152630) confirmed the presence of the patterns initially identified by Shen (2002). Specifically, the majority of the correlation coefficients of the three variables examined in Shen's study with student achievement in each subject were statistically significant and higher than 0.25. When aggregated data at country level were used, statistically significant but negative relations were found (the range of the absolute values of the correlation coefficients were 0.61 to 0.72 in the case of Mathematics and 0.50 to 0.73 in the case of Science).

Data were then submitted to multi-level analysis for each outcome measure in Mathematics and Science separately. The first step in the analysis was to determine the variance at individual, class, school and country level without explanatory variables (empty model or model 0). The variance at each level reached statistical significance (p<.05) and this implies that MLwiN could be used to identify the explanatory variables which were associated with achievement in each subject. Table 1 illustrates the parameter estimates and the standard errors deriving from the multi-level analysis of achievement in Mathematics and Science. We can observe that 31.6% of the variance in Mathematics achievement was at the pupil level, 20.5% of the variance was at the class level, 5.3% at the school level and 42.6% at the country level. Similarly, 39.6% of the variance in Science achievement was at the pupil level, 21.5% at the class level, 9.7% at the school level and 29.2% at the country level. This implies that for both subjects, classrooms had a higher effect on student achievement than schools. This finding is in line with the current findings of EER which emphasise the significance of the classroom effect (e.g., Kyriakides, Campbell, & Gagatsis, 2000; Teddlie & Reynolds, 2000). We can also observe that the country effect is much higher than the effect of classroom and school.

Table 1: Parameter Estimates And (Standard Errors) For The Analysis Of Mathematics And Science Achievement (Pupils Within Classes, Within Schools, Within Countries)

SUBJECT	*MATHEMATICS*			*SCIENCE*		
Factors	*Model 0*	*Model 1*	*Model 2*	*Model 0*	*Model 1*	*Model 2*
FIXED PART **(Intercept)**	49 554 (1169)	38 750 (1270)	116 316 (11 780)	52 464 (524)	43 827 (533)	96 256 (4 838)
Pupil level						
Self-perceived competence		2517.5 (23.3)	2517.4 (23.3)		22.3 (0.3)	22.3 (0.3)
Attitude towards the subject		1040.7 (21.6)	1040.8 (21.6)		9.9 (0.3)	9.9 (0.3)
Perceived easiness of the subject		216.1 (20.6)	216.2 (20.6)		-2.9 (0.2)	-2.9 (0.2)
Country level						
Self-perceived competence			NSS			NSS
Attitude towards the subject			NSS			NSS
Perceived easiness of the subject			-24 887 (4169)			22 158.1 (2 021.9)
RANDOM PART: Variances						
Country	42.6%	42.4%	24.4%	29.2%	29.2%	12.4%
School	5.3%	5.2%	5.1%	9.7%	9.5%	9.5%
Class	20.5%	18.7%	18.7%	21.5%	18.9%	18.9%
Student	31.6%	26.4%	26.4%	39.6%	36.3%	36.3%
Absolute	108 915 061	100 964 261	80 669 539	91 175 374	85 659 263	70 259 743
Explained		7.3%	25.4%		6.1%	22.9%
Significant tests						
X^2	3 070 521	3 043 875	3 043 843	3 106 641	3 093 825	3 093 744
Reduction		26 646	32		12 816	81
Degrees of freedom		3	1		3	1
p-value		.001	.001		.001	.001

NSS= Not statistically significant effect.

In model 1 the three variables at student level which were used in Shen's (2002) analysis were added to the empty model of each subject. The likelihood statistic (X^2) showed a significant change between the empty model and model 1 ($p<.001$) for both subjects. Moreover, a small percentage of the variance of student achievement in each subject was explained (i.e., Mathematics: 7.3% and Science: 6.1%) and almost all of the explained variance was at the level of student and classroom. We can also observe that all exploratory variables were associated with achievement. In model 2, the three explanatory variables at country level concerning student self-perceptions were added. All of these were aggregated from the student level data. It was found that only the perceived easiness of the subject had a significant effect on achievement in each subject. However, the effect of perceived easiness in Science was negative both at the student and country level whereas in Mathematics this pattern did not occur. This reveals that the findings of multi-level analysis are generally not in line with the attempt of Shen to aggregate the data at country level and identify relevant correlations. Specifically, multi-level analyses show that two aggregated variables at country level (i.e., self-perceived competence and attitude towards the subject) are not associated with student achievement and only perceived easiness in Mathematics had a positive effect when taken as a student variable, and a negative effect when aggregated at country level. It can therefore be claimed that the comparison of the single analyses and multi-level analyses investigating the relationship of achievement with these three affective variables measured at both the student and country level reveal the importance of using multi-level modelling techniques in analysing hierarchical data, as is the case of the IEA data.

IDENTIFYING VARIABLES OF TIMSS STUDY ASSOCIATED WITH STUDENT ACHIEVEMENT BY TAKING INTO ACCOUNT THEORETICAL MODELS OF EDUCATIONAL EFFECTIVENESS

This part presents the results of an attempt to analyse TIMSS 1999 data by using multi-level modelling techniques. The purpose of this analysis is to examine the extent to which the main theoretical models of EER could help us identify variables which were measured through the TIMSS study and are associated with student achievement. However, it should be stressed that the empirical studies which were conducted in order to test the main assumptions of EER models (e.g., Jong & Westerhof 1998; Reezigt, Guldemond, & Creemers, 1999; Driessen & Sleegers, 2000; Kyriakides et al., 2000) used value-added techniques and thereby were focused on student progress rather than final outcomes. It has already been recognized that one significant limitation of the IEA studies relates to the absence of a measure of students' prior knowledge or intelligence (van der Linden, 1998). Therefore, the results of this analysis should better be seen as informative for designing future comparative evaluation studies rather than providing empirical support to the EER models.

In an attempt to identify variables which might be associated with student achievement, we mainly drew on the comprehensive model of educational effectiveness (Creemers, 1994), which is one of the most influential theoretical constructs in the field (Kyriakides, 2003; Teddlie & Reynolds, 2000). This model is an extension of Carroll's model of school learning (Carroll, 1963), which is considered as the starting educational productivity model, and states that the degree of mastery is a function of the ratio of amount of time students actually spend on learning tasks to the total amount of time they need. His model consists of five categories of variables that are expected to explain variations in educational achievement: aptitude, opportunity to learn, perseverance, quality of instruction and ability to understand instruction. Numerous studies and meta-analyses have confirmed the validity of Carroll's model (e.g., Doyle, 1985; Stallings, 1985) but, as Carroll (1989) pointed out 25 years after the construction of his model, the one factor in his model that needed further elaboration was 'quality of instruction'. In this context, Creemers (1994) developed Carroll's model of learning by adding to the general concept of "opportunity", the more specific "opportunity to learn". In Creemers' model time and opportunity are discerned both at the classroom level and at the school level. In this way Creemers made a distinction between available, and actually used, time and opportunity.

Creemers' model was also based on four assumptions. First, time-on-task and opportunity used at the student level are directly related to student achievement. Second, quality of teaching, the curriculum, and the grouping procedures influence the time on task and opportunity to learn. For example, some teachers spend more time actually teaching than others, who spend more time on classroom management and keeping order. Teachers are therefore the central component in instruction at the classroom level. Third, teaching quality, time and opportunity at the classroom level are also influenced by factors at the school level that may or may not promote these classroom factors. Thus, quality, time and opportunity are not the key concepts at the classroom level only, but also at the school level. Finally, it is acknowledged that, although teachers are able to influence time for learning and opportunity to learn in their classrooms through the quality of their instruction, it is the students who decide how much time they will spend on their school tasks and how many tasks they will complete. Thus, achievement is also determined by student factors such as aptitudes, social background and motivation.

Identifying explanatory variables from TIMSS data

Based on the main assumptions of Creemers' model presented above, a selection was made of all TIMSS variables which could be categorized as: context, time, opportunity and quality factors. The variables derived from the student, teacher and school questionnaires are presented below.

(A) Explanatory variables at student level

- *Student Background Factors:* Information was collected on three student background factors: age, sex (0=girls, 1=boys), and socio-economic status (SES). Three SES variables were available: father's and mother's education level and the possession of the four items included in student questionnaires in all countries: calculator, computer, dictionary and study desk.
- *Expectations:* Expectations were measured through two items of the questionnaire asking each student to indicate the extent to which his/her mother and friends

believe that it is important to do well in mathematics. Another item asking students to indicate the extent to which they believe that it is important to do well in mathematics was also taken into account.

- *Time factors:* Time on task for each student is not easy to assess because it is difficult to observe what is going on in a person's mind. Therefore, we used a proxy for this variable. Specifically, time on task during classes was measured through an item asking students to indicate how frequently they skip a class during a month. Moreover, students were asked to indicate whether their peers skip the class and this was seen as another measure of student attentiveness (i.e., peers-rated attentiveness).
- *Opportunity factors:* Time spent doing homework and time spent on private tuition were seen as measures of the opportunity factor. Information was collected from two relevant items administered to students.

(B) Explanatory variables at teacher level:

- *Contextual factors:* Variables concerned with the context of each classroom, such as the average of each of the three SES indicators mentioned above, and the gender percentages were measured. The contextual factors were aggregated from the student level data. We were also able to use some variables concerning the characteristics of teachers, especially background characteristics such as sex, age and length of teaching experience. As far as teachers' subject and pedagogy knowledge is concerned, we collected data on whether teachers held a first degree in mathematics and/or mathematics education or whether they had relevant postgraduate qualifications.
- *Quality of teaching:* At the classroom level, quality of instruction is considered as one of the main variables which account for learning on the part of student. Relevant information was collected through items included in the student and teacher questionnaires. On the student questionnaire, we used 20 items asking students to indicate how often a number of activities take place in the mathematics lessons. The data collected were aggregated at the teacher level. Similar items concerning the grouping of students during a lesson, and the frequency with which specific teaching tasks are employed (e.g., students explain the reasoning behind an idea, practice computational skills) were used. Items concerning the way teachers make use of information gathered from assessment were also taken into account.
- *Time factors:* Although the actual time spent on teaching mathematics was not measured by TIMSS, it was possible to use a proxy of this variable based on student responses to an item concerning the extent to which their teachers get interrupted. For each teacher an average of his/her students' responses was calculated.
- *Opportunity to learn:* Opportunity to learn was measured through three set of items included on the teacher questionnaire that were concerned with the amount of homework their students are usually asked to undertake, the type of homework students are assigned and the way teachers proceed with completed homework.

(C) Explanatory variables at country level

- *Context:* In order to investigate the country effect, contextual factors concerning the average SES indicators were taken into account. These factors were all aggregated from the student level data. Four items from the school questionnaire concerning the admittance policy and especially those related to the performance criteria were also considered as indications of the context of the school. Finally, items asking headteachers to specify the frequency and severity of student behaviors that cause problems in school life were examined.
- *Quality factors:* Conditions for the quality of instruction at school and country level can be measured by investigating the extent to which each school has developed rules and agreements about aspects of classroom instruction in Mathematics. Therefore two items included in the school questionnaire, asking headteachers to indicate whether their school has its own written statement of the curriculum content to be taught, and whether they had developed relevant instructional materials to address the curriculum were used. Quality factors were also measured by collecting data from 10 items concerning the extent to which parents were involved in school life. Finally, the extent to which the school has a policy on differentiation was measured by taking information from a relevant set of items from the school questionnaire.
- *Time factors:* The measurement of time factors at school level emerged from an item asking headteachers to evaluate the extent to which students are absent on a typical day.
- *Opportunity to learn:* Consensus about the "mission" of the school can be seen as an indication of the conditions for the opportunity to learn at the school level. Thus, items asking headteachers to indicate the extent to which teachers share ideas, materials, and discuss instructional goals and issues were taken into account.

MLwiN (Rasbash, Browne, Goldstein, Yang, Plewis, Healy, Woodhouse, Draper, Langford, & Lewis, 2002) was used in order to examine the extent to which the variables mentioned above are associated with student achievement. At the first stage, we examined which levels had to be considered in order to reflect the hierarchical structure of the data. Empty models with all possible combinations of the levels of analysis (i.e., student, class, teacher, school and country) were established and the likelihood statistics of each model were compared (Snijders & Bosker, 1999). It was found that an empty model consisting of student, teacher and country level represented the best solution. Therefore, this model was used as a basis for subsequent analyses. Table 2 illustrates the parameter estimates and standard errors of the explanatory variables which were found to be associated with student achievement. All models were estimated without the variables that did not have a statistically significant effect.

Table 2: Parameter Estimates And (Standard Errors) For The Analysis Of Mathematics Achievement Using Exploratory Variables Adopted From Educational Effectiveness Research (Pupils Within Classes, Within Countries)

Factors	*Model 0*	*Model 1*	*Model 2*	*Model 3*	*Model 4*	*Model 5*
FIXED PART: (Intercept)	487.4 (11.6)	445.5 (11.3)	497.5 (12.2)	252.3 (49.0)	252.8 (48.2)	258.2 (48.5)
Pupil level						
Sex		8.2 (0.4)	8.3 (0.4)	8.5 (0.4)	8.4 (0.4)	8.2 (0.5)
Mother's Educational Background		4.4 (0.2)	4.4 (0.2)	3.6 (0.2)	3.6 (0.3)	3.5 (0.3)
Father's Educational Background		5.6 (0.2)	5.7 (0.2)	5.2 (0.2)	5.3 (0.3)	5.5 (0.3)
Possessions		10.8 (0.3)	10.6 (0.3)	10.1 (0.3)	10.0 (0.3)	10.1 (0.3)
Time factors						
Student Perceived attentiveness			-12.7 (2.3)	-12.3 (2.1)	-12.7 (2.3)	-12.7 (2.5)
Peers Perceived attentiveness			-17.0 (1.6)	-13.9 (1.4)	-13.3 (1.5)	-9.0 (1.7)
Teacher level						
Mother's Educational Background				46.2 (1.2)	42.2 (1.3)	40.7 (1.4)
Possessions				78.7 (15.7)	74.1 (15.3)	74.6 (15.6)
Quality of teaching						
Students work in small groups (APD)					-12.7 (1.4)	-9.8 (1.6)
Writing equations to represent relations					2.6 (0.9)	2.0 (1.0)
Use graph calculators to solve problems					4.0 (1.0)	3.2 (1.1)
Opportunity to learn						
Students are given homework (APD)					17.1 (1.6)	15.3 (1.8)
Homework related to writing definitions					-1.7 (0.8)	NSS
Time factors						
Time spent on re-teaching and clarifying content					-0.3 (0.1)	-0.3 (0.1)
Country level (ASD)						
Contextual factors						
Admittance policy						8.5 (2.6)
Severity of behaviours causing problems						-4.3 (1.3)
Time factors						
Absenteeism						-0.7 (0.2)
Quality factors						
Parents active involvement						9.4 (2.6)
RANDOM PART: Variances						
Country	44,1%	39.0%	39.0%	22.7%	21.6%	20.9%
Teacher	24.7 %	20.0%	18.7%	14.5%	13.8%	13.3%
Student	31.2%	29.6%	29.4%	29.4%	29.4%	29.4%
Absolute	11624.3	10 299.1	10 124.8	7 741.8	7 532.5	7 393.1
Explained		11.4%	12.9%	33.4%	35.2%	36.4%
Significance test						
X^2	2 011 548	1 362 420	1 260 443	1 259 016	1 137 382	924 362
Reduction		649 128	101 977	1427	121 634	213 020
Degrees of freedom		4	2	2	6	4
p-value		.001	.001	.001	.001	.001

APD = aggregated pupil data, ASD = aggregated school data, NSS = Not Statistically Significant Effect

Of the total variance, 3441.2 {Standard Error (SE)=14.3}, 2278 (SE=44.1) and 4528.4 (SE=1068.7) was accounted for at the individual, teacher and country level, respectively. Thus, the variance at each level reached statistical significance (p<.001) and this implies that MLwiN could be used to identify the explanatory variables which are associated with mathematics achievement (Goldstein, 2003). We can also observe that 31.2% of the variance was at the student level, 24.7% of the variance was at the teacher level and 44.1% was at the country level. In model 1 the context variables at student level were added to the empty model. The following observations arise from the figures of the third column of Table 2. First, model 1 explained 11.4% of the total variance and the likelihood statistic (X^2) showed a significant change between the empty model and model 1 (p<.001). Second, the effects of all contextual factors were significant. Specifically, adding gender background improved model 1 and boys achieved higher scores than girls in Mathematics. Furthermore, all variables related to student SES were positively related to their achievement. In model 2 all the variables at the student level (i.e., 'expectations', 'student's self-rated attentiveness', 'peers-perceived attentiveness', 'homework' and 'private tuition') were added but only the two indicators of time factors were found to be associated with student achievement. We can also observe that model 2 explained 12.9% of the variance, all of which was at the student and teacher level.

In model 3 the contextual factors at teacher level were entered. The following observations arise from the fifth column of Table 2. First, the two contextual factors at classroom level showing statistically significant effect were the two SES indicators aggregated at the classroom level. Second, none of the variables concerning teacher background characteristics had a statistically significant effect. Third, this model explained 33.4% of the total variance and the inclusion of SES indicators at both student and classroom level helped us explain almost 50% of the unexplained variance at country level. In model 4 all the explanatory variables at classroom level which are included in Creemers' model and taken into account by TIMSS study were entered. Although only six variables were found to have a statistically significant effect on student achievement, these variables refer to all three main categories of variables included in Creemers' model (i.e., time factors, quality of teaching and opportunity to learn). Specifically, the amount of homework given to students was positively related to student achievement. However, a negative relation between student achievement and the frequency of assigning homework related to writing definitions was identified. It is also important to note that asking students to write equations in order to represent relations and use graph calculators to solve problems were positively associated with achievement, whereas working in small groups was negatively associated with achievement.

In model 5 the variables at country level were entered. The following observations arise from the figures in the seventh column of Table 2. First, the only contextual factor at country level which had a statistically significant effect was the existence of a school admittance policy. It is important to stress than none of the SES indicators at country level was found to be associated with student achievement. Second, the only measure of quality of instruction associated with student achievement was related to the existence of a policy on active parental involvement. Although no form of parental involvement other than active involvement was found to be associated with student achievement, this finding is in line with the results of a meta-analysis which was conducted in order to synthesize the quantitative literature on the relationship between parental involvement and students' academic achievement (Fan & Chen, 2001). This study revealed that active involvement has a more noticeable effect than other dimensions of parental involvement on students' academic achievement. Third, none of the variables concerning opportunity to learn had significant effects and only one indicator of time factors was found to be associated with student achievement. Finally, model 5 explained only 36.4% of the variance and 29.4% remained unexplained at the student level, 13.3% at the teacher level and 20.9% at the country level. These figures reveal that most of the explained variance was at the country and teacher level rather than student level.

DISCUSSION

The results reported in this paper reveal the importance of using multi-level modelling techniques to analyse TIMSS data. This argument is indicated through the comparison of the findings that emerged from single and multi-level analyses provided in the second part of the paper. The findings of multi-level analysis indicate that the errors associated with ignoring the hierarchical structure of data are not hypothetical, especially since variables which were aggregated at a higher level (i.e., in our example the country level) were not found to have statistically important effects on achievement. These inconsistencies arise from the fact that single level analysis does not take into account certain sources of variability within and between units. The findings presented in this paper also indicate that multi-level analysis of TIMSS data provides a means for partitioning the variability of student achievement into different levels and, therefore, allows researchers to compare effects at those levels. For instance, in the multi-level analyses reported in the second part of the paper, it was found that classroom effects were more important than the school effects for both subjects. Country effects were even more important than those either at the classroom or school level. Thus, multi-level analysis provides a single framework for identifying variables at different levels that explain variance in student achievement and thereby produces more accurate explanations of the phenomena under study.

The latter argument is further supported by the analysis reported in the third part of the paper where TIMSS variables at different levels were found to explain part of the variance of student achievement. However, most of the total variance of student achievement remained unexplained (63.6 %). Nevertheless, this can be attributed to the fact that no measure of aptitude was taken into account. Various studies in different countries (e.g., Kyriakides, 2002; Teddlie & Reynolds, 2000; Townsend, Clarke & Ainscow, 1999) reveal that variables such as prior knowledge or intelligence have the most significant effect on student achievement, and therefore, need to be taken into account in comparative studies. Such an approach could also aid in measuring students' progress rather than final outcomes. As van der Linden (1998) supports,

> *Schools differ in the conditions under which they have to operate as well as the input of students from previous programs or local families they have to accept. Therefore, ...a better practice would be to compare them only for the "added value", that is for the knowledge and skills schools add to whatever students already know and can do when they enter the program, given their individual background* (p.572).

It is also important to note that the inclusion of aptitude variables in IEA studies could lead to more coherent conclusions, since their effect in the aforementioned studies was found to be even stronger than the effect of student SES variables, which in the present study were found to have the most significant effects. These significant effects of SES indicators could be misinterpreted, especially if we considered how the results of research on equality of opportunity in education conducted in the USA in the 1960s and 1970s (Coleman et al., 1966, Jencks et al., 1972) were used to claim that pupil attainment can entirely be determined by social background characteristics. In this perspective, it can be argued that the inclusion of such variables in future IEA studies should be seriously taken into account, in order to encourage researchers to focus both on student progress and on the factors associated with this progress.

All three multi-level analyses reported here revealed a large variation at country level. This provides support to previous arguments that TIMSS data should not be used for summative reasons (i.e., ranking countries according to their students outcomes) (Cogan and Schmidt, 2002; Stedman, 1997; Valverde & Schmidt, 2000), but for identifying factors associated with student achievement at various levels. Such an approach could be much more constructive, since these findings could help policy makers design intervention programmes that are focused on those specific factors found to be associated with achievement and which are the most critical for each country. In this way, analysis of TIMSS data will be conducted mainly for formative rather than summative reasons.

However, one might assert that the large variation at country level identified in the present paper suggests that the educational systems of different countries bear little resemblance. Therefore international comparative studies should give their place to national studies that would be illuminating in terms of the specific needs of the educational system of each country. We argue that the adoption of this approach would lead the pendulum to the other end, fostering the ethnocentric character of research in the domain of educational effectiveness. It could also deprive researchers and policy makers from information that cannot be gained from single-systems studies alone (Schmidt, Jakwerth & McKnight, 1998). Thus, in the future international comparative studies should try to identify both school and teacher effectiveness factors that are present in different educational contexts (i.e., that "travel" across countries) (Reynolds, Creemers, Stringfield, Teddlie, & Schaffer, 2002) and factors that are unique to specific countries. Moreover, factors that operate differently in different educational settings need to be highlighted. Moving a step forward, researchers should also address questions related to the reasons that make some of these factors universal and some specific, and the rationale for the differential effectiveness of some factors in different countries. Such information should be helpful in two ways. On the one hand it could contribute to the development of both generic and differentiated comprehensive theoretical models of educational effectiveness (Campbell, Kyriakides, Muijs & Robinson, 2003), and on the other hand it could be much more useful for research into the improvement of educational effectiveness both at micro and at macro level (Kyriakides, 2003).

Implications of the above argument for the design of international comparative studies could also be drawn. The fact that most of the variance in student achievement was not explained even after adding a large number of variables included in the student, teacher and school questionnaires of TIMSS underlines the need for reconsidering the design of IEA studies so as to carefully select appropriate effectiveness factors. Since the need for theoretical frameworks that precisely define the significant variables in the process of education has been highlighted (Bos & Kuiper, 1999; Lassibille & Gomez, 2000), we argue that theoretical models of EER (e.g., Creemers, 1994; Scheerens, 1992; Stringfield & Slavin, 1992) could serve as a source for developing relevant instruments. This argument is supported by the fact that the secondary analysis of TIMSS data presented in the third part of this paper has shown that variables linked to the three main factors of Creemers' model (i.e., time, opportunity and quality) were associated with student achievement. It could, therefore, be argued that had the instruments of IEA studies provided more information concerning these three factors more variance would have been explained. This especially holds for the country level, since variables associated with different educational policies that may affect the three main factors of Creemers' model at school or teacher level were absent. For example, future IEA studies could be directed towards investigating the effect of providing educators training and support that promote effective instruction on student achievement gains.

In order to provide further support to the suggestions mentioned above, additional secondary analyses of IEA studies could be conducted through multi-level modelling. First of all, it is important to conduct a similar analysis to the one reported in the third part of the paper on Science data. Employment of multi-variate multi-level analyses could also be conducted and might help us examine whether effectiveness factors operate similarly in both subjects. Moreover, the extent to which there is criterion-consistency in school effectiveness can be identified by examining if schools and countries are equally effective in Mathematics and Science. A similar question can be addressed in relation to the effectiveness of each country. Another possible way (using existing TIMSS data) to identify generic and context-specific effectiveness factors is to conduct separate analyses for different countries with notable differences in the way support is provided to schools in order to implement curriculum as well as in the core beliefs of their societies concerning appropriate educational aims and best practices. In this context, we currently attempt to compare results of multi-level analyses of both subjects (i.e., Mathematics and Science) using the data from Japan, Singapore, England and Canada. The selection of these countries was based on the claim that in some countries (e.g., East Asian countries) virtually all educational professionals adopt the same values about what should happen in a classroom or a school, whereas in other countries (e.g., the English-speaking countries) there is huge variation in what is seen as appropriate or at least acceptable teaching practice that might reflect unresolved value debates at national level (Alexander, 2000; Reynolds et al, 2002). It can be claimed that these attempts could further contribute to extending the scope of analysing TIMSS data and would also develop a better understanding of education by which all countries participating in the IEA studies would benefit.

References

Alexander, R. (2000). Culture and Pedagogy. Oxford: Basil Blackwell.

Alker, H.R. (1969). A typology of ecological fallacies. In M. Dogan, & S. Rakkan (Eds.), *Quantitative ecological analyses in the social sciences, 69-86.* Cambridge, Mass.: The M.I.T. Press.

Bos, K. & Kuiper, W. (1999). Modelling TIMSS Data in a European Comparative Perspective: Exploring Influencing Factors on Achievement in Mathematics in Grade 8. *Educational Research and Evaluation, 5* (2), 157-179.

Brophy, J. (1992). Probing the subtleties of subject matter teaching. *Educational Leadership, 49*, 4-8.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models.* Newbury Park: CL: SAGE.

Campbell, R.J., Kyriakides, L., Muijs, R.D., & Robinson, W. (2003). *Assessing Teacher Effectiveness: A Differentiated Model.* London: Routledge Falmer.

Carroll, J. B. (1963). *A model of school learning. Teacher College Record, 64*, 723-733.

Carroll, J. B. (1989). The Carroll Model: A 25 year retrospective and prospective view. *Educational Researcher, 18*, 26-31.

Cogan, L. S., & Schmidt, W. H. (2002). "Culture Shock" – Eighth Grade Mathematics from an International Perspective. *Educational Research and Evaluation, 8* (1), 13-39.

Coleman, J.S., Campbell, E., Hobson, C., McParttland, J., Mood, A., Weinfield, F. & York, R. (1966). *Equality of Educational Opportunity.* Washington DC: US Government Printing Office.

Creemers, B.P.M. (1994). *The effective classroom.* London: Cassell.

Doyle, W. (1985). Effective secondary classroom practices. In M.J. Kyle (Ed.), *Reading for excellence: An effective schools sourcebook.* Washington, DC: U.S. Government Printing Office.

Driessen, G. & Sleegers, P. (2000). Consistency of Teaching Approach and Student Achievement: An empirical test. *School Effectiveness and School Improvement*, 11 (1), 57-79.

Fan, X. & Chen, M. (2001). Parental Involvement and Students' Academic Achievement: A Meta-Analysis. *Educational Psychology Review, 13* (1), 1-22.

Goldstein, H. (2003) (3rd Edition). *Multilevel statistical models.* London: Edward Arnold.

Heck, R.H. & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques. Mahwah*, NJ: Lawrence Erlbaum Associates.

Huttner, H.J.M. & van den Eeden, P. (1995). *The Multilevel Design: A guide with an Annotated Bibliography, 1980-1993.* Westport, Conn.: Greenwood Press.

Jencks, C., Smith, M.S., Ackland, H., Bane, M.J., Cohen, D., Grintlis, H., Heynes, B. & Michelson, S. (1972). *Inequality.* New York: Basic Books.

Jong, R. & Westerhof, K. J. (1998). Empirical evidence of a comprehensive model of school effectiveness: a multi-level study in Mathematics in the first year of junior general education in the Netherlands. *Paper Presented at the International Congress for School Effectiveness and Improvement 10.* Manchester, UK.

Kyriakides, L. (2004). A Study on Differential Teacher and School Effectiveness: Some Implications For Policy Evaluation. *Paper presented at the International Congress for School Effectiveness and Improvement 17.* Rotterdam, the Netherlands.

Kyriakides, L. (2003). A theoretical framework for school effectiveness research based on Creemers' model: An empirical study. *Paper presented at the 84th Annual Meeting of the American Educational Research Association.* Chicago, USA.

Kyriakides, L. (2002). A research based model for the development of policy on baseline assessment. *British Educational Research Journal, 28* (6), 805-826.

Kyriakides, L., Campbell, R.J., & Gagatsis, A. (2000). The significance of the classroom effect in primary schools: An application of Creemers comprehensive model of educational effectiveness. *School Effectiveness and School Improvement, 11* (4), 501-529.

Lassibille, G., & Gomez, L.N. (2000). Organization and Efficiency of educational Systems: some empirical findings. *Comparative Education, 36* (1), 7-19.

Martin, M.O. (1996). Third International Mathematics and Science Study. In M.O. Martin, & D.L. Kelly (Eds.), *TIMSS technical report, vol.1* (p. 1.1-1.19). Boston College: IEA.

Martin, M. O., Mullis, I.V.S., Gonzalez E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K. M., (2000). *TIMSS 1999 - International Science Report.* The International Study center at Boston College: IEA.

Mullis, I.V.S., Martin, M. O., Gonzalez E.J., Gregory, K.D., Garden, R.A., O'Connor, K. M., Chrostowski, S.J., & Smith, T. A. (2000). *TIMSS 1999- International Mathematics Report.* The International Study center at Boston College: IEA.

Opdenakker, M.C., & van Damme, J. (2000). The Importance of Identifying Levels in Multilevel Analysis: An Illustration of Ignoring the Top or Intermediate Levels in School Effectiveness Research. *School Effectiveness and School Improvement, 11*(1), 103-130.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., & Lewis, T. (2002). *A users' guide to MLwiN.* United Kingdom: University of London, Institute of Education.

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

Reezigt, G. J., Guldemond, H., & Creemers, B.P.M. (1999). Empirical Validity for a Comprehensive Model on School Effectiveness and School Improvement. *School Effectiveness and School Improvement, 10* (2), 193-216.

Reynolds, D., Creemers, B., Stringfield, S., Teddlie, C. & Schaffer, G. (Eds.) (2002). *World Class Schools*. London: Routledge Falmer.

Scheerens, J. (1992). *Effective Schooling: Research, Theory and Practice.* London: Cassell.

Schmidt, W., Jakwerth, P., McKnight, C.C. (1998). Curriculum sensitive assessment: Content *does* make a difference. *International Journal of Educational Research*, 29, 503-527.

Schmidt, W. & Valverde, G.A. (1995). *National Policy and Cross-National Research: United States Participation in the Third International and Science Study. East Lansing*, MI: Michigan State University, Third International Mathematics and Science Study.

Shen, C. (2002). Revisiting the Relationship Between Students' Achievement and their Self-perceptions: a cross-national analysis based on TIMSS 1999 data. *Assessment in Education, 9* (2), 161-184.

Snijders, T. & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* London: Sage.

Stallings, J. (1985). Effective Elementary classroom practices. In M. J. Kyle (Ed.), *Reaching for excellence: An effective schools sourcebooks.* Washington DC: US Government Printing Office

Stedman, L. C. (1997). International Achievement Differences: An Assessment of a New Perspective. *Educational Researcher, 26* (3), 4-15.

Stringfield, S.C. & Slavin, R.E. (1992). A hierarchical longitudinal model for elementary school effects. In B.P.M. Creemers & G.J. Reezigt (Eds.), *Evaluation of Educational Effectiveness*, Groningen: ICQ.

Teddlie, C. & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research.* London: Falmer Press.

Townsend, T., Clarke, P. & Ainscow, M. (1999) (Eds.), *Third Millennium Schools: A World of Difference in Effectiveness and Improvement.* Lisse: Swets and Zeitlinger.

Valverde, G.A., & Schmidt, W. H. (2000). Greater expectations: learning from other nations in the quest for "world-class standards" in US school mathematics and science. *Journal of Curriculum Studies, 32* (5), 651-687.

van der Linden, W.L. (1998). A discussion of some methodological issues in international assessments. *International Journal of Educational Research, 29*, 569-577.

Walberg H.J. (1986). Syntheses of research on teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 214-229). New York: Macmillan.

Yang, Y. (2003). Dimensions of Socio-economic Status and their Relationship to Mathematics and Science Achievement at Individual and Collective Levels. *Scandinavian Journal of Educational Research*, 47 (1), 21-41.