

**Understanding causal influences on educational achievement
through analysis of within-country differences over time**

Jan-Eric Gustafsson
Department of Education, Gothenburg University

Invited paper presented at the 2nd IEA Research Conference, Washington, November 8-
11, 2006

Preliminary version Oct 20, 2006. Do not cite or quote.

Address:

Jan-Eric Gustafsson

Department of Education

University of Gothenburg

P. O. Box 300

S-405 30 Gothenburg

SWEDEN

Jan-Eric.Gustafsson@ped.gu.se

Abstract

When the International Association for the Evaluation of Educational Achievement (IEA) was established in 1959 the basic idea was to use comparative analysis of country differences in achievement to take advantage of the world as an educational laboratory. A large number of comparative studies involving substantial numbers of countries has since then been conducted and much has indeed been learned. However, it has also been learned that causal inference from cross-sectional comparative data is a weak method for gaining knowledge about which factors are conducive to educational achievement, because of the impossibility to control for the large number of differences between countries. During the 1990s a new generation of IEA studies was launched. These studies (e. g. TIMSS) were designed to give information about within-country trends of achievement in addition to information about between-country differences. The paper proposes that analysis of within-country differences over time is a powerful method of finding out which educational factors are related to achievement, and particularly so when the analysis involves several countries. This suggestion is illustrated through analyses of data from the TIMSS study of mathematics achievement in 1995 and 2003 for grades 4 and 8, investigating effects of age and class size on achievement.

Introduction

The International Association for the Evaluation of Educational Achievement (IEA) was founded in 1959 by a small group of educational and social science researchers with the purpose of using international comparative research as a means to understand the great complexity of factors influencing the achievement of students in different subject matter fields. A popular metaphor was that they wanted to use the world as an educational laboratory.

The first study, which investigated mathematics achievement in 12 countries, was conducted in 1964 (Husén, 1967). During the more than 40 years which have passed since the publication of this study, different groups of researchers have under the auspices of the IEA published a large number of studies of educational achievement in different countries in a wide range of subject matter areas. For example, in the first round in 1995 of the TIMSS study (at that time TIMSS was an acronym for Third International Mathematics and Science Study) which investigated knowledge and skill in mathematics and science, no less than 39 countries participated (Beaton, Mullis, Martin, Gonzalez, Kelly and Smith, 1996; Mullis, Martin, Beaton, Gonzalez, Kelly and Smith, 1997). For the third round of TIMSS (TIMSS now stands for Trend in Mathematics and Science Study) (Mullis, Martin, Gonzalez, & Chrostowski, 2004) which was conducted in 2003, 50 countries participated, and for the fourth round in 2007 an even larger number of countries will participate. Not only has the number of participating countries increased dramatically, but also the frequency of repetition which is due to the fact that the studies of mathematics, science and reading are now designed to capture within-country trends of achievement, and are therefore repeated every fourth or fifth year.

The data collected in these studies has been used to generate a vast amount of knowledge about international achievement differences. Because the data has been made freely and easily available to all interested researchers they have been used in a large number of secondary analyses by researchers in many different fields such as economics, education, sociology, and didactics of different subject matter areas. However, voices of criticism

have also been raised against them. One line of criticism is that the international studies have primarily come to serve as a source of benchmarking data for purposes of educational policy and educational debate, thereby becoming a means of educational governance, reducing the importance and influence of national policy makers (Nóvoa & Yariv-Mashal, 2003). This benchmarking function of the international studies has grown in importance during the 1990s. One reason for this is that the methodology of the international studies has become more suited for efficient and unbiased estimation of country-level performance, basically through taking advantage of the advances in the National Assessment of Educational Progress (NAEP) in the United States, where a complex methodology based on item-response theory and matrix-sampling designs had been developed during the 1980s (see Jones & Olkin, 2004). Another reason is that the increasing number of participating countries made the benchmarking function more interesting. This became even more pronounced when the OECD also started international surveys of educational achievement through the PISA program (OECD, 2001). The OECD presence even more emphasized the economic importance of the educational results.

Another line of criticism of the international studies of educational achievement is that the “world educational laboratory” has not been particularly successful in disentangling the complex web of factors which are important in producing a high level of knowledge and skills among the students. Even though advances have been made, there is still a lot to be learned, and doubts have been expressed that cross-sectional surveys offer the appropriate methodology for advancing this kind of knowledge. Indeed, Allard (1990) argued that there is little evidence that comparative surveys in any field of social science have been able to generate knowledge about causal relations, pointing at the great complexity of the phenomena investigated, and at the uniqueness of different countries, as the reasons for this. The observation that cross-sectional surveys do not easily allow causal inference is made in many text-books on research design, so the methodological challenges are well known. Furthermore, the international studies of educational achievement are not based upon an elaborated theoretical framework, which makes it difficult to apply the analytical methods which have been developed for purposes of

causal inference from cross-sectional data (Williams , Williams, Kastberg, & Jocelyn, 2005).

It would be unfortunate if the aim of generating explanations in causal terms for patterns of results in the comparative educational surveys would be lost out of sight as a consequence of these difficulties. The search for explanations is one of the main aims of scientific research, and explanations also are needed if policymakers are to take full advantage of the benchmarking results. The present chapter argues that there might be reason for a somewhat more optimistic stance concerning the possibility to arrive at causal inferences if advantage is taken of the possibilities to do longitudinal analyses at the country level of the data generated by the trend design implemented in the latest generation of comparative educational studies. The reason for this is that with the longitudinal design many variables which in cross-sectional research cause spurious results are kept constant because they do not vary within countries over time.

The chapter thus has two main aims: to (a) discuss those methodological problems and pitfalls in the international studies which are of importance when the purpose is to make causal inferences; and (b) to show with two concrete examples how a longitudinal approach to analysis of country level trend data may be used to avoid some of these methodological problems.

Problems and pitfalls in causal inference from cross-sectional data

A typical international cross-sectional study measures achievement for samples of students in a set of participating countries, and information also is collected about student characteristics (e. g., social background, gender, and courses taken), teacher characteristics (e. g., experience and education), teaching characteristics (e. g., teaching methods, homework) , school characteristics (e. g., size, location, and relations with parents). In addition, country-level information is typically collected about, among other things, curriculum and institutional factors (e. g., organisation of the school-system,

decision-making power at different levels). Using different kinds of statistical analysis, such as regression analysis, these background and contextual factors are then used as independent variables in order to determine their amount of influence on achievement.

Let me take a concrete example of such a study. Mullis, Campbell and Farstrup (1993) used NAEP data collected in 1992 to investigate the effects of the amount of direct teaching of reading grade 4 students had obtained on their level of reading performance. When they correlated amount of instruction and achievement they found a significant negative correlation, showing that those students who had obtained more teaching had a lower level of reading achievement. It does not seem reasonable, however, to interpret the negative correlation as being due to a causal relation such that more teaching of reading causes students to read more poorly. On the contrary, it is more reasonable to interpret the negative correlation as being an expression of a compensatory educational strategy, such that poor readers had been provided with more teaching resources, either in regular education or in special education, than had proficient readers. Mullis et al. (1993) favored the latter interpretation.

This example suggests that it is easy to confuse the direction of causality in cross-sectional data, and that it therefore is necessary to be cautious when making causal statements on the basis of analyses conducted with such data. The problem of confused direction of causality is well known and it goes under different labels in different disciplines. Econometricians refer to this as the “endogeneity” problem, while sociologists and psychologists often talk about the problem of “reversed causality”. Another term used to describe the same problem is to say that there is “selection bias” such that groups of students who received different treatments were not comparable in terms of their level of performance before they received the treatments.

This problem, whatever label is used, seems inescapable in studies with a cross-sectional design, at least when the analysis is done at the individual level. This is because it is rarely, if ever, reasonable to assume that the amount and kind of instruction allocated to a student is independent of the characteristics of the student. As the example presented

above illustrates compensatory resource allocation is, for example, common in educational contexts, implying that more or better instruction is given to poor-performing students. Lazear (2001) has developed a theoretical model to explain the effects of variation in class-size, which model among other things implies that there will be a selection bias such that larger classes will tend to be populated by higher-performing students. Thus, a positive correlation may be expected between class-size and achievement.

One way to deal with the problem of selection bias is to statistically control for the differences between students that existed before the treatment was applied. However, this requires a measure of these pre-existing differences, and in a cross-sectional design such measures are generally not available. If it is possible to use a longitudinal design, measures of pre-existing differences can be obtained. In international studies of educational achievement longitudinal designs have not been used, mainly for practical reasons. Instead other, more readily available measures, such as indicators of socio-economic background, have been used to control for the pre-existing differences.

This brings me to another source of erroneous causal inference from cross-sectional data, namely the problem of omitted variables. When an independent variable is related to a dependent variable in a statistical model, and the estimated relation is interpreted in causal terms it is assumed that there are no other independent variables which are correlated with the independent variable in focus, and which have not already been included in the model. If there are such omitted variables they will cause bias in the estimated causal relation if they correlate with the residual of the dependent variable, and it may imply that we ascribe causality to other variables than the ones which are actually involved. One approach to try to solve the problem of omitted variables would be to try to measure and analyze all potentially relevant variables. It is, however, virtually impossible to exhaustively include all the relevant variables, even if strong theory is available to guide the selection.

To summarize the discussion above, two main threats to correct causal inference from cross-sectional data have been identified. The first is the problem of selection bias which makes us confuse the direction of causality, and of course also causes biased estimates of the strength of effects. The other problem is the problem of omitted variables, which makes us ascribe causality to other independent variables than the ones that are actually causally involved.

Before taking any further step in the discussion it may be useful to present a concrete example of an analysis of IEA data which aims to solve the problems of causal inference, namely the study by Wössman (2003).

Wössman's secondary analysis of TIMSS 1995

Wössman (2003) recently presented a very ambitious and interesting analysis of the TIMSS 1995 data for population B (i. e., grades 7 and 8). The main aim of the study was to investigate effects of schooling resources and institutional factors on student achievement.

Wössman (2003) argued that the data should be analyzed at the individual level, because:

The relevant level of estimation is the individual student (not the class, school, district or country), because this directly links a student's performance to the specific teaching environment. The estimation of such microeconomic education production functions provides the opportunity to control for individual background influences on student performance, to assess the effects of the relevant resource and teacher characteristics which the student faces, and to estimate the effects of institutional features below the country level which are relevant to the individual student. (p. 124).

Wössman (2003) thus constructed a student-level database comprising the more than 260 000 students in grades 7 and 8 from the 39 countries that participated in TIMSS 1995.

The database included the achievement scores in mathematics and science, along with the information collected through questionnaires administered to students, teachers, and principals. In addition information at country level collected in the TIMSS study was included, and combined with data from other sources on the institutional structure of the educational system, such as the level of decision-making, and use of centralized examinations.

The data was analyzed with regression models, in which student test scores were regressed on measures of student background, along with indicators of resources and institutional factors measured at the classroom, school and country level. In order to take into account effects of the hierarchically structured data on estimates of standard errors Wössman used a technique called clustering-robust linear regression, which empirically estimates the covariances of the error terms caused by the non-independence of individuals due to the cluster sampling design.

Wössman (2003) expressed confidence that effects of the country-level institutional factors were possible to estimate correctly with this model, because there should not be any endogeneity problems in the estimation of these effects. He also emphasized that estimation of these factors requires using the “world as an educational laboratory”:

The link between institutions and student performance could hardly be tested using country-specific data as there is no significant variation in many institutional features within a single country on which such an analysis could be based. Only the international evidence which encompasses many education systems with widely differing institutional structures has the potential to show whether institutions have important consequences for students. (p. 120)

The main conclusion of the analyses reported in the paper was that there were quite strong effects of institutional factors. It was, among other things, found that there were positive effects on performance of centralized examinations and control mechanisms,

school autonomy in making decisions about hiring of teachers, and individual teacher influence over teaching methods.

The results also showed that resource factors such as class-size had little effect on student achievement. However, Wössman (2003) noted that: “Given the single cross-section structure of the performance study, the potential endogeneity problems plaguing resource estimates cannot be fully overcome in the TIMSS data.” (p. 121). This makes it necessary to interpret the negative findings concerning effects of resource factors with caution.

It may be instructive to take a somewhat closer look at some of the estimates obtained, in order to judge whether they are reasonable or not. For mathematics it was found that attending an additional grade improved performance with 40 points, on the scale with a mean of 500 and standard-deviation 100. This estimate is well in line with other studies of the effects of attending another year at school (see below, and Cahán & Cohen, 1989), even though it is higher than what has been found in most other studies. For becoming one year older in chronological age a negative effect of -14 scores points was observed. This estimate is hard to accept as a correct determination of the effect of becoming one year older, because in the middle school years it is a well established finding that chronological age is positively associated with achievement in mathematics. Instead it is more reasonable to interpret the negative effect of age as being due to endogeneity problems: students of lower ability start school at older age and repeat classes more often than high-performing students. Obviously the available measures of student characteristics (primarily family background) did not adequately control for the selection bias introduced in this way.

For class size a fairly strong and significant positive effect was found, implying that students who attended larger classes outperformed students who attended smaller classes. This is counter-intuitive and does not agree with findings in a large number of other studies (see below). For this age-level the most well-established finding is that class-size does not have any effect on achievement (Gustafsson, 2003, Hoxby, 2000). Again, the most credible interpretation of this effect is that it is an expression of an endogeneity

problem (see Lazear, 2001), which cannot be properly controlled for due to the lack of entry-level measures of achievement in the TIMSS data.

There is also reason to discuss somewhat further the interpretation of the findings concerning institutional effects. One problem that has already been brought up, is that there may still be problems due to omitted variables, causing estimates of parameters to be biased, and incorrect inferences to be made about which independent variable is causing the effect. For example, one of Wössman's main findings was that use of centralized examinations has an effect on achievement (16 points for mathematics). However, the use or non-use of centralized examinations has not been randomly assigned to countries, but tends to be more common in certain categories of countries (Confucian Asian countries in particular), than in other categories of countries. The different categories of countries also are likely to differ in a large number of other respects, such as for example the value ascribed to education and how teaching in classrooms is organized. Such variables may also affect achievement, and if they are not controlled for, the effect may be incorrectly ascribed to use of centralized examinations. Given the great number of potential omitted variables it is an impossible task to argue that a particular study is not afflicted by the omitted variable problem.

Alternative approaches to causal inference

It thus seems that the two major methodological problems of endogeneity and omitted variables still remain to be solved in the analysis of data from international studies of educational achievement. How to do that is discussed below.

The lack of appropriate control over previous achievement makes it very difficult to deal with the problem of selection bias in the comparative studies of educational achievement. However, even though the selection bias problem seems virtually impossible to deal with at the individual level, it may be more approachable at higher levels of aggregation. Thus, even though there may be inextricable connections between student characteristics and the treatments that the students receive, there need not be such connections at the school or

national level. A concrete example of this is the suggestion by Wössman and West (2005) to study effects of class-size at school level rather than at class level. The reason is that the distribution of students over classes within a school is likely to be associated with endogeneity problems, while there is no a priori reason to suspect that mean class size for the schools is affected by endogeneity problems.

It is also easy to think of other examples where there is selection bias at individual level, but not necessarily at class level. Take, for example, amount of homework assigned to students. At the student level there is likely to be selection bias such that poorly performing students are assigned much homework, which at least some students spend quite a lot of time on. Motivated and high-achieving students also are likely to spend time and effort on home-work. The combined effect may be that there is little of a relation between amount of homework and student achievement. However, there may also be a variation between teachers in their readiness to assign homework to students, which is an expression of an instructional strategy, and which is independent of the composition of the group of students. One way to conduct such analyses would be to perform two-level modeling of the relations between amount of homework and achievement at individual level and at class-level using either regression techniques as implemented in HLM (Bryk & Raudenbush, 19xx), or two-level latent variable techniques as implemented in for example Mplus (Muthén & Muthén, 2006).

It is, of course, also possible to conduct the analysis at even higher levels of aggregation, such as the national level. At this level there will be even less risk of being misled by problems of selection bias, because it is not reasonable to expect that differences in use of homework, for example, between countries can be affected by endogeneity problems.

Analysis of data at higher levels of aggregation may thus be one way of solving the problem of selection bias. There may be reason to use different levels of aggregation (e. g., class, school, and country) for different substantive questions, but here I will assume aggregation to the country level. The main reason for this choice is that several of the international studies now are designed to yield estimates of change in achievement over

time. Such trend data on a panel of countries offer opportunities for longitudinal analysis, which may help solving at least a part of the omitted variable problem. Through relating within-country change over time in explanatory variables to within-country change in achievement a “fixed-country” analysis is performed which keeps most country characteristics fixed. The characteristics which are kept constant in this way cannot be candidates for being omitted variables.

It is thus proposed that the problems of selection bias and omitted variables can at least partially be solved through conducting longitudinal analyses of data aggregated to the country level. While this approach only allows investigation of a subset of the questions which are of interest in comparative educational research it may be worthwhile to see to what extent causal inference is possible. Below two examples of research issues are discussed, mainly from the methodological point of view, namely the effect of student age on achievement, and the effect of class size on achievement.

Example 1: Student age and mathematics achievement

The analytical idea will first be illustrated with a concrete example, focusing on an independent variable which should have a fairly simple relationship with mathematics achievement. One such variable is student age, which according to results in several different kinds of studies is positively related to student achievement.

The TIMSS 1995 study allowed inference to be made about the combined effect of another year of schooling and becoming one year older, because both for population 1 (9-10 year olds) and population 2 (13-14 years olds) adjacent grades were sampled. For population 1 most countries sampled grades 3 and 4, and for the upper grade the international mean was 57 points higher than for the lower grade. However, the increases in mean performance between the two grades varied between a high of 84 points in the Netherlands and a low of 46 points in Thailand. For population 2 the samples of most countries included students from grades 7 and 8, and the international mean difference in level of performance between upper and lower grades amounted to 30 points. All

countries did not have the same mean difference, however. The largest difference was observed for Lithuania (47 points) and the smallest difference for South Africa (7 points), with the other countries in between these extremes.

These results indicate that the combined age/grade effect is almost twice as large for the younger students as it is for the older students. However, as was observed by Beaton et al. (1996, p. 28) the differences in level of achievement in grades 7 and 8 was in some cases affected by policies regarding promotion of students from one grade to another, which caused the performance difference to be underestimated. There may, thus, be some downward bias in the between-grade estimates, but the results for both populations nevertheless indicate a sizeable age effect.

These estimates also agree quite well with results presented by Cahan and Cohen (1989), who furthermore showed that with a regression discontinuity design the effects on achievement of one year of schooling and becoming one chronological year older could be separated. The results indicated that approximately one third of the total effect was due to chronological age and two thirds to schooling.

Results for grade 8

In order to estimate the age/grade effect from trend data for grade 8 a database has been constructed comprising 22 countries that participated both in TIMSS 1995 and in TIMSS 2003. Table 1 presents the 22 included countries. This list includes one country less than was included in the analysis of trend between 1995 and 2003 presented by Mullis et al. (2004), namely Bulgaria. The reason for this is that the data collected in 1995 for Bulgaria lacks most of the background questionnaires, which makes this country of limited interest for the current analytical purposes. The TIMSS 1995 study comprised both grades 7 and 8, while the TIMSS 2003 study only included grade 8. Therefore, only grade 8 students were selected from TIMSS 1995 to be included in the analysis, in the same manner as students were selected for the trend analyses reported by Mullis et al. (2004).

Table 1 also presents the mean country level mathematics achievement scores and student age. Included are also change scores for achievement and age between the two surveys. The achievement scores were computed from a mean of the five plausible values assigned to each student, and the results presented in Table 1 agree perfectly with those presented by Mullis et al. (2004).

Insert Table 1 about here

For a couple of countries there was quite a substantial difference in the mean age of the participating students in 1995 and 2003. This was true for Latvia and Lithuania, and, to a somewhat smaller extent, for Korea and Romania. For Latvia and Lithuania the difference amounted to about 9 months. Differences of this magnitude must have been caused by some systematic factor, even though it is not clear why these age differences appeared. However, for most of the countries the differences in mean age of the students were quite small, and are likely to have been caused by random differences when exactly the assessments were conducted. The countries were allowed freedom to conduct the field work in a time interval ranging from late April to early June, and it is likely that the decisions involved in when to conduct the survey have caused the observed variation in mean student age.

Table 2 presents correlations between the mathematics achievement variables and the age variables at the country level. Mathematics achievement in TIMSS 1995 correlated .93 with math achievement in TIMSS 2003, and there was also a substantial correlation of .75 of mean student age between the two surveys. Within each survey there was no correlation between mean age and mean achievement. For TIMSS 1995 achievement correlated .19 with age, and for TIMSS 2003 the corresponding correlation was .16. Even though these correlations are marginally positive they are non-significant. However, there was a highly significant correlation of .58 ($p < .005$) between the Mathematics change variable and the Age change variable.

Insert Table 2 about here

This simple analysis thus yields the somewhat paradoxical result that within each survey there was no association between student age and student achievement at the country level, but there was a strong correlation between change in age and change in mathematics performance between 1995 and 2003. One explanation for this pattern of results is that at each of the two occasions the correlation between age and achievement is influenced by other factors, such as school starting age in different countries, and cultural and economic factors. These factors, which are omitted variables in the analysis of cross-sectional data, may conceal a true correlation between student age and achievement. However, the correlation between change in achievement and change in age within countries keeps these factors associated with countries constant, thereby allowing the correlation between age and achievement to appear.

A regression analysis of the Math change variable on the Age change variable gives an unstandardized regression coefficient of 38, which implies that an age change of one year is associated with an increase in achievement of 38 points. This estimate reflects the combined effect of becoming one chronological year older and going to school one more year, so it agrees reasonably well with the grade 7 – grade 8 achievement differences found in TIMSS 1995 (30 points). However, it should be noted that there is a subtle difference between the meanings of these estimates. The difference in level of achievement between grades 7 and 8 is due to the combined effect of a 12 month chronological year and a school year which is typically around 9 months. However, the unstandardized regression coefficient is based on the observed age variation within a school year so when this is expressed in terms of the expected effect of a one-year difference, it captures the combined effect of a 12 month chronological year and a 12 month year of schooling. Assuming that the effect of another month in school is twice as large as just becoming one month older, the regression estimate of 38 points may be rescaled into an estimate that is comparable with what has been obtained when

comparing adjacent grades. The adjusted estimate is 32 points, which makes it very close to the estimate of 30 points derived from the comparison of adjacent grades.

Figure 1 presents a plot of the Math change variable against the Age change variable. From the figure it may be seen that for Lithuania and Latvia the mean age of the students was 8 to 9 months higher in TIMSS 2003 than in TIMSS 1995, and for Lithuania achievement was 30 points higher in 2003, while for Latvia it was 20 points higher. For most other countries the mean age difference only amounted to a few months, and in most cases the achievement change was smaller than in Lithuania and Latvia. However, if the results for Lithuania and Latvia are excluded from the regression, the estimated coefficient is 28, which is still a substantial relation.

Insert Figure 1 about here

Results for grade 4

The TIMSS 1995 and 2003 studies also included samples of students from grade 4, and there were 15 countries which participated in both studies. It is of great interest to see to what extent the results presented above for grade 8 may be replicated with the data for the grade 4 students. Table 3 presents the list of countries included, along with their mathematics achievement score and the mean age of the students in 1995 and 2003.

Insert Table 3 about here

Most of the countries had improved their level of mathematics achievement between 1995 and 2003 and the mean change was 10 points. There was, however, a considerable variation in the amount of change, with the highest improvement (47 points) being observed for Latvia and largest drop (-25 points) for Norway. The mean age was somewhat higher in 2003 than in 1995, the difference amounting to .07 years. Here too, however, there was variation among the countries. In Latvia the average age of the

students was .60 years higher in TIMSS 2003 than in TIMSS 1995, while in Iran students were .10 years younger in 2003 than in 1995. With the exception of Latvia these age differences were not larger than may be expected from the fact that the countries may conduct the assessments within a time frame of about two to three months. Table 4 presents the correlations among the variables.

Insert Table 4 about here

The pattern of correlations for grade 4 was highly similar to the pattern observed for grade 8. The correlation between age and achievement was low within each of the two assessments (.21 and .33 for 1995 and 2003, respectively), but the correlation between change in mathematics achievement and change in age was substantial and highly significant ($r=.63$, $p < .013$). While the correlation between age change and change in achievement was highly similar for grade 4 and grade 8, the estimate of the regression coefficient was higher in grade 4 ($b = 71$). However, this estimate too expresses the combined effect of a 12 month chronological year and a 12 month school year, so to be comparable to the estimate obtained from the comparison of adjacent grades (57 points) it should be adjusted in the same manner as was done for grade 8. The adjusted estimate is 59, so for grade 4 too the estimates agree excellently.

Figure 2 presents the plot of change in achievement against change in age.

Insert Figure 2 about here

The plot clearly indicates the high level of correlation between the two change variables. It also indicates that Latvia is somewhat of an outlier, with a very high improvement in achievement and a high change in mean age. It may be asked, therefore, if the high correlation between the two variables is due to the results for this particular country. However, if the correlations are recomputed with Latvia excluded the same result is obtained ($r = .63$, $p < .015$).

Conclusions

In theory, the mean student age should be the same for countries participating in TIMSS 1995 and TIMSS 2005, because samples are drawn from the same grade and the testing should be conducted at the same time of the year, which is specified to be at the end of the school year (April, May or June for countries in the Northern Hemisphere). However, because the testing may be conducted during a time interval, the student age at testing need not be quite the same during the two assessments. Even though this variation in testing time may be random, the existing variation is related to achievement. And if the variation in student age between the different cohorts is random it is not correlated with other variables, which makes inferences about the effect of age on achievement unaffected by bias due to omitted variables.

Analyses of the relation between student age and achievement within each of the two cross-sectional data sets did not show any correlation between age and achievement. This indicates that these analyses are influenced by omitted variables which disturb the positive relation between age and achievement. The school start age varies between countries and this is an important determiner of the age variability at the time of testing. However, the school start age varies over groups of countries, the Nordic countries for example having a high school start age, which makes it 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.

The analysis of the relation between change in achievement and change in mean student age between 1995 and 2003 gives an estimate of the effect of student age which agrees with results from other approaches. This indicates that this is a simple method of controlling for the influence of omitted variables, and it is reasonable to assume that this is due to the fact that the differences in mean student age between the two assessments are not systematically related to any other variable.

Example 2: Effects of class size on student achievement

Let me now turn to a quite different example. Most parents, teachers and students expect learning to be more efficient in a smaller class than in a larger class. Furthermore, class size is one of the most important variables to determine the level of resources needed for an educational system. Many studies have been conducted on the effects of class size on achievement, but the results have tended to be inconsistent, and at least up to the early 1990s the general conclusion was that neither resources in general, nor class-size in particular had any relation to achievement.

However, during the 1990's influential studies were published which demonstrated that class size does indeed matter. In particular results reported from a large-scale experiment on the effects of class size, namely the so called STAR-experiment (Student/Teacher Achievement Ratio), were influential. The experiment, which started in 1985, had three treatment groups: small classes with 13-17 students; regular classes with 22-26 students; and regular classes with an assistant teacher. For a school to be included in the study it had to be large enough to have at least one class of each type. Some 80 schools participated, with more than 100 classes of each type. During the first year of the study about 6 000 students were included, and throughout the four years of experimentation almost 12 000 students were involved, because of addition of new students. Within schools both students and teachers were randomly assigned to the three treatments. Most of the students entered the study either in kindergarten or grade 1, while a few entered in grades 2 or 3. In the first phase of the study the students were followed till the end of grade 3, measures of achievement being made at the end of each grade. The great strength of this design is, of course, that the randomization of both students and teachers over the three treatment conditions implied that there were no problems of causal inference due to selection bias or omitted variables.

At the end of grade 3 the results showed quite a striking advantage for the students who had been assigned to the small classes, while there was no clear difference between the results achieved in regular classes with one teacher, and regular classes which had an assistant teacher as well (Finn & Achilles, 1990, 1999). The results were particularly

strong for reading with an effect size of around .25, and it was found that the small class advantage was larger for students who came from socio-economically and ethnically disadvantaged groups (Finn & Achilles, 1990; Kreuger, 1999). The results showed the effect of class-type to be strongest for grade 1, while the difference kept more or less constant over grades 2 and 3.

The results from the STAR experiment have been replicated in other studies as well (see Gustafsson, 2003, for a review). Robinson (1990) reported a large meta-analysis, which comprised more than 100 studies of class size. The results indicated that the effect of class size interacts with the age of the students. For grades K-3 Robinson found a positive effect of small classes. For grades 4-8 a weak positive effect was found "... but the evidence is not nearly as strong as in grades K-3" (Robinson, 1990, p 84). Studies conducted on students from grades 9-12 provide no support for the hypothesis that class size has an effect on achievement.

Other studies have relied on natural variation in class size, using sophisticated statistical techniques to try to sort out causal effects of class size from selection effects. Angrist and Lavy (1999) conducted a study in Israel, where there is a quite strict rule for the maximum number of students in a class, and which implies splitting one large class into two smaller classes. The variation in class size close to the maximum class size is essentially unrelated to factors such as the socio-economic status of the area that the school is recruiting from, which implies that the variation in class size caused by the splitting rule can be used to estimate effects of class size. Using so called instrumental variable estimation on data from some 2000 fourth and fifth grade classes, class size was found to significantly affect achievement in reading and mathematics in grade 5, and in reading in grade 4, smaller classes producing the better results. The effect sizes were somewhat lower than those found in the STAR experiment (for example, .18 for grade 5, assuming a reduction of class size with 8 students), but were still judged large enough to be practically important. Just like in the STAR experiment, Angrist and Lavy (1999) also found that there was an interaction between socio-economic background and class size,

the benefits of small classes being larger in schools with a large proportion of students from a disadvantaged background.

Hoxby (2000) took advantage of the fact that natural variation in population size influences class size, and this variation causes random variation in class size which is not associated with any other variation, except perhaps achievement. She also used a similar approach as did Angrist and Lavy (1999), investigating the abrupt changes in class size caused by rules about maximum class size. Using data from 649 elementary schools covering a period of 12 years this approach to estimation of effects of class size was used. In no case a significant effect of class size was found, in spite of the fact that Hoxby demonstrated that power was sufficient to detect class size effects as small as those found in the experimental research. Hoxby (2000) suggests that the differences between the results obtained in the experimental studies and her studies of natural variation may be interpreted as being due to the fact that the teachers in the experimental studies tried to make good use of small classes, because an outcome showing an advantage for small classes would be favored by the teachers.

Several studies have used the IEA data, and TIMSS in particular, to investigate effects of class-size and other resource factors. Hanushek and Luque (2003) used data from the TIMSS 1995 study to investigate effects of resources on educational achievement, focussing on possible differences in effects between different countries. They analyzed data for both 9-year olds (Population A) and 13-year olds (Population B), modelling relations at the school level within each country between resource and background factors on the one hand and achievement on the other hand.

One general finding that emerged from the analyses was that effects of resources seemed to be stronger in other countries than the USA. However, there was little consistency over countries and age groups in the pattern of results. For class-size it was for the age 9 samples found that 14 out of 17 estimated relations with achievement had the expected negative sign, while for the age 13 samples 23 out of 33 estimated relations with class size had a positive sign. This pattern of results thus suggests that smaller classes are

beneficial for the achievement of young students, but not for older students, which agrees quite well with the findings in the research literature that any positive effect of a smaller class size is restricted to the first years of schooling.

Hanushek and Luque (2003) were suspicious, however, concerning the result for Population B that in the majority of countries that there was a positive effect of being in a larger class. They hypothesized that this may be an effect of selection bias caused by a tendency to place weaker students in smaller classes, and they tested this hypothesis in two different ways. In one approach they performed a separate analysis of schools in rural areas, arguing that such schools typically only have one class-room, leaving no room for any mechanism of selection bias to operate. However, this analysis did not provide any other pattern of results than the analysis of the complete set of schools. In the other approach information provided by the principal whether the particular sampled class-room in a school had a class size below average for the grade in the school was used. The results showed that five of 32 countries for the 13 year-olds showed lower achievement in the classrooms with smaller size than the grade average for the school, which supports the hypothesis that weaker students tend to be placed in smaller classes. However, according to Hanushek and Luque (2003) taking such within-school compensatory placements into account did not change the estimated class size effects.

These results thus indicate that for the older age group class-size does not seem to be related to student achievement, while for the younger age group the expected positive effect of being in a smaller class was found in quite a few countries.

Wössman and West (2005) made another analysis of the TIMSS 1995 population B data in order to determine effects of class size. In order to come to grips with the problem of selection bias within and between schools they used a sophisticated instrumental variable estimation approach, which relied upon differences in class size between adjacent grades, and differences between mean class size of the school and the actual class size. The analysis also took into account country fixed effects. Adequate data to perform the estimation was available for 11 countries. The estimation results showed that there were

statistically significant negative effects of class size on achievement in a few cases (France and Iceland in mathematics, and Greece and Spain in science), while in the vast majority of cases no effect of class size was found. Thus, this study too indicates that for students in grades 7 and 8, class size is not an important determinant of achievement.

Summarizing the results from the studies of effects of class size, it seems that there is some support for the conclusion that there is a positive effect of smaller classes during the first few years of schooling, but little or no effect after that. There are, however, many studies in which the results run against this generalization, so further research is necessary. Below results are reported from analyses in which class size change has been related to mathematics achievement change between 1995 and 2003 for grades 8 and grade 4.

Results for grade 8

The analyses of class size effects rely on the same set of grade 4 and grade 8 countries in TIMSS 1995 and 2003 that were analyzed in the previous example. However, unless there have been changes in the class sizes of the participating countries there is no reason to investigate correlates of this variable. Class size at the country level has been estimated from the information about the size of mathematics classes asked about in the teacher questionnaire. It should be noted that for the 1995 data there were only separate variables for the number of boys and girls in the class, while for the 2003 data there was a variable providing the total number of students in each class. This difference, which only applied to grade 8, may have introduced some extra variability in the estimates of change of class size, given that there were some peculiarities in the distributions of class size for boys and girls.

In 1995 the international mean class size for grade 8 was 27.3 and in 2003 the international mean was 26.1, so there is a tendency towards decreasing class size. The standard deviation of the class size change was 3.7, indicating considerable variability in the amount of class size change over the countries.

The correlations among the variables are presented in Table 3.

Insert Table 3 about here

It can immediately be observed that there is no correlation ($r = .08, p < .71$) between change in class size and change in achievement. This result thus agrees with the conclusion drawn in several studies that there is no causal effect of class size on achievement for older students.

But it is also interesting to note that according to the cross-sectional data there is a significant positive relationship between class-size and achievement in mathematics both in 1995 ($r = .42, p < .050$) and in 2003 ($r = .60, p < .003$). These results indicate that larger class-rooms produce a higher level of achievement than smaller class-rooms. However, the class size variable may be correlated with a large number of other variables that may be instrumental in causing a higher level of achievement, so the analysis may be influenced by omitted variable problems. Figure 2 presents a plot for the 22 countries of mean mathematics achievement and mean class size.

Insert Figure 2 about here

This plot shows that the positive correlation is caused by the high level of performance and the large class size of a group of four countries, namely Hong Kong, Japan, Korea, and Singapore. These countries are all Confucian Asian countries, and there is no other country which belongs to this group. However, the Confucian Asian countries differ in many other ways than just with respect to having a larger class size than other countries in the world. For example, there is a strong emphasis on education and math not only in school but also as a strong cultural value. There is also a high level of quality in the teaching, the lessons being carefully planned and the teachers taking advantage of the large groups of students (Biggs, 19xx). This suggests that the positive relation between country level class size and math achievement may be accounted for in terms of country

differences in a complex set of cultural and educational factors, rather than in terms of country differences in class size per se.

The absence of a zero-order correlation between change in class size and change in mathematics performance makes it unlikely that bringing in further variables into the analysis will yield any other results. However, the fact that there was a substantial correlation between change in age and change in mathematics achievement may make it worthwhile to include both age change and class size change as independent variables in a regression analysis. According to this analysis the partial regression coefficient for class size change was almost exactly 0, and the regression coefficient for age was 38, which was also the estimate that was obtained when the variable was entered alone into the model.

As has been the case in many other studies, this analysis thus provides no ground for claiming that there is an effect of class size on mathematics achievement for grade 8 students.

Results for grade 4

For the 15 countries that participated with grade 4 students in TIMSS 1995 and TIMSS 2003 the international mean class size was 26.3 in 1995 and it was 25.7 in 2003. For this group of students too there is thus a slight decrease in the class size. The standard deviation of the class size change was 2.6, which indicates that there is variability over the countries with respect to the amount of change. The correlations between the variables are presented in Table 6.

Insert Table 6 about here

For grade 4 too there were positive correlations between mathematics achievement and class size for the cross sectional data for 1995 and 2003 (.34 and .54, respectively). The class size change variable correlated -.22 with change in mathematics achievement,

which was not significant ($p < .44$). However, adding also the age change variable caused both the age change variable and the class size change variable to be highly significant ($t = 4.91$ and $t = -3.32$, respectively), these two variables accounting for 68 % of the variance in the mathematics change variable.

The regression coefficient was -4.44 for the regression of change in mathematics achievement on class size change, which implies that a class size reduction of 7 students, as in the STAR experiment, would yield an improvement of 31 points. This improvement roughly translates into an effect size of .31. According to Finn and Achilles (1999) the effect size of being in a small class at the end of grade 3 was .26 for reading and .23 for mathematics. However, the effect sizes were quite different for different groups of students. For mathematics the effect size was .16 for white children, while it was .30 for minority children. Thus, even though the effect size found in the present analysis seems somewhat higher than what was found in the STAR experiment the results do not seem to be contradictory.

Conclusions

These analyses suggest that for the TIMSS 1995 and 2003 grade 8 data there was no effect of class size on academic achievement, because change in class size at country level was not related to change in achievement over time. It also could be concluded that the strong correlation between achievement and class size at the country level at both occasions was due to the fact that both these variables had high values for a group of four Confucian Asian countries. It must be emphasized, of course, that the negative results for grade 8 should not be interpreted as proof that there is no effect of class size on educational achievement. We cannot prove the correctness of the null hypothesis, and there may be alternative explanations for the lack of relationship between class size change and achievement change. It was, for example, noted that there was a slight change in the definition of the class size variable between the grade 8 1995 data and the 2003 data.

For the grade 4 data the trend analysis between 1995 and 2003 showed a significant effect of class size change on change in mathematics achievement, with the expected negative relationship between class size and achievement. The size of the estimated effect agreed reasonably well with what has been found in previous research, even though it was somewhat higher than what has been found in most studies.

The general pattern of findings emerging from these two sets of analyses for populations A and B agrees quite well with what has emerged from the cumulated research on the effects of class size, namely that class size does not affect achievement for older students in grade 6 and upwards, but that it is of importance for primary school students. This agreement should not be interpreted as proof that the analytical procedure focusing on change over time at the country level provides unbiased estimates of causal relations, but the results are encouraging, and do support further work along these lines.

Discussion and Conclusions

The main aim of the present chapter was to identify possible threats to the correctness of causal inference from the international studies of educational achievement, and to suggest approaches which would make it possible to avoid negative effects of these threats. It was concluded that selection bias (or reverse causality, or endogeneity problems) and omitted variables were two major sources of problems. These problems seem almost impossible to deal with through analyses at low levels of aggregation of cross sectional data. It was therefore suggested that data may be analyzed at higher levels of aggregation focussing on change over time for the same units of observation. Given that the latest generation of international studies have been designed to yield information about trends in the development of knowledge and skills within countries, this suggests that aggregation should be made to the country level, and that an analytical approach taking advantage of the longitudinal design should be adopted. There should be no mechanisms generating selection bias at the country level, and the fact that change over fixed countries is analyzed turns many of those variables which vary over countries into constants so that they cannot correlate with the independent variables under study.

In order to investigate the tenability of these ideas two examples were studied using data from grades 4 and 8 in TIMSS 1995 and 2003, namely the effects of age on achievement and the effects of class size on achievement. Both these examples demonstrated that the analyses based on cross-sectional data yielded biased results, while the results from the longitudinal country-level analyses were reasonable and in good agreement with results obtained in studies using other methodological approaches. This suggests that the country-level trend analyses may be a useful addition to the set of tools available for secondary analysis of the data from the comparative international studies.

It must be emphasized, though, that this analytical approach is not a panacea, and that there are many problems which cannot be studied with this approach. It also is easy to envision threats to the validity of causal inferences from country-level longitudinal analyses.

Problems which focus upon explanatory variables where there is no change over time obviously cannot be studied with the longitudinal approach proposed here. For example, the institutional factors investigated by Wössman (2003), such as the use of centralized examinations, are not likely to change over shorter periods of time, or at least not to such an extent that any effects can be determined. To investigate the effects of institutional factors at the country level other approaches are needed, even though it does not seem that the micro-econometric modelling approach applied by Wössman (2003) can avoid the omitted variable problem.

There is, of course, no guarantee that change measures of explanatory variables are uncorrelated with other independent variables, and that they therefore are not afflicted by the omitted variable problem. It may, perhaps, be argued that the age change variable investigated in Example 1 at least for grade 4 is a more or less random variable. This makes it optimal as an explanatory variable, because if it is random it is not correlated with other independent variables, even though it still exerts an influence on change in achievement. But for most other independent variables of any educational interest we

cannot assume that change is uncorrelated with other variables. Suppose, for example, that a set of countries who have achieved a poor result in a comparative study take different measures to improve achievement, such as lowering class size and improving teacher competence. This implies that there will be correlations among the different resource factors, which implies that it will be difficult to sort which, if any, is having any effect. To add to the analytical complexity, there is in this scenario likely to be some regression toward the mean of the country results when the assessment is repeated. As always, great care must thus be taken in analyzing and interpreting the results.

In this paper the simplest possible analytical techniques have been applied, in order to keep focus on the basic methodological issues. However, correlation and regression analysis of difference scores can only be applied in a meaningful manner when there are two waves of measurement. For TIMSS, which is conducted with a four-year cycle, three waves have already been completed, and a fourth wave of data collection is conducted in 2007. For PIRLS the second wave of measurement was completed in 2006, and the third will be completed in 2011. To take full advantage of such multi-wave data other techniques should be employed. One technique which would seem to be a natural choice when a systematic trend over a longer period of time is expected is growth modeling at the country level.

There are many other interesting ways in which the analysis of trend data can be extended. There is no reason to restrict the outcome to apply to the total sample of students, but it may be broken down into results for different subgroups, such as by gender, language spoken at home, and socio-economic background. It also might be interesting to investigate variability of scores as an outcome when questions of equity are investigated.

While the current paper primarily has its focus on methodological issues, the results concerning class size are of substantive interest, even though that will not be further discussed here. The results concerning effects of age differences at time of testing for the different waves of measurement are of limited theoretical interest, but they do seem to be

of practical interest when designing the implementation of the assessments. It has been demonstrated here that these age differences accounted for 30 - 40 % of the variance in the change of level of achievement between 1995 and 2003. This is a substantial share of the total variance and there may be reason to try to reduce that, perhaps through allowing a more narrow time frame when the assessments can be carried out. Alternatively, it is, of course, also possible to take the age differences into account through statistical adjustment.

In the introduction it was observed that IEA was founded with the intention to use the world as an educational laboratory. While the founding fathers might have been overly optimistic about the possibility to overcome the technical and methodological problems encountered when setting up this laboratory, many problems, such as those associated with sampling and measurement, have successfully been mastered during the decades of work within the IEA. The laboratory is still somewhat messy, however, and in particular the problem of how to sort out the multitude of factors influencing achievement remains to be solved. It has in this chapter been argued that systematic analysis of trends of achievement in different countries may provide at least a partial solution to this problem. Further work will be needed to determine to what extent this is true.

References

- Allard, E. (1990). Challenges for comparative social research. *Acta Sociologica*, 33(3), 183-193.
- Angrist, J. D., & Lavy, V. (1999a). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114, 533-575.
- Biggs, J. (1998). Learning from the Confucian heritage: so size doesn't matter? *International Journal of Educational Research*, 29(8), 723-738.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: Sage Publications.
- Cahan, S., & Cohen, N. (1989). Age versus schooling effects on intelligence development. *Child Development*, 60, 1239-1249.
- Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, 28, 557-577.
- Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications and misconceptions. *Educational Evaluation and Policy Analysis*, 21, 97-110.
- Gustafsson, J.-E. (2003). What do we know about effects of school resources on educational results? *Swedish Economic Policy Review*, 10(2), 77-110.
- Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics*, 115(4), 1239-1285.
- Husén, T. (Ed.) (1967). *International study of achievement in mathematics: a comparison of twelve countries. Vol. 1 & 2*. New York: Wiley.
- Jones, L. V., & Olkin, I. (Eds.) (1994). *The Nations Report Card. Evolution and Perspectives*. Bloomington, Indiana: Phi Delta Kappan.
- Krueger, A. B. (1999). Experimental estimates of educational production functions. *Quarterly Journal of Economics*, 114(2), 497-532.
- Lazear, E. P. (2001). Educational production. *Quarterly Journal of Economics*, 116(3), 777-803.
- Mullis, I. V., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). TIMSS 2003 international mathematics report. Findings from IEAs Trends in International Mathematics and Science Study at the fourth and eighth grades. Lynch School of

- Education, Boston College: TIMSS and PIRLS International Study Center.
- Mullis, I. W., Campbell, J. R., & Farstrup, A. E. (1993). *NAEP 1992 Reading Report Card for the Nation and the States: Data from the National and Trial State Assessments*. NCES Report No. 23-ST06; September. Washington, D. C.: U. S. Department of Education.
- Muthén, L. & Muthén, B. O. (2004). *Mplus User's Guide*. Third edition. Los Angeles, CA: Muthén & Muthén.
- Nóvoa, A., & Yariv-Mashal, T. (2003). Comparative research in education: a mode of governance or a historical journey? *Comparative Education*, 39(4), 423-438.
- OECD (2001). *Knowledge and skills for life: first results from PISA 2000*. Paris: OECD.
- Robinson, G. E. (1990). Synthesis of research on the effects of class size. *Educational Leadership*, 47 (7), 80-90.
- Williams, T., Williams, K., Kastberg, D., & Jocelyn, L. (2005). Achievement and affect in OECD countries. *Oxford Review of Education*, 31(4), 517-545.
- Wössman, L. (2003). Schooling resources, educational institutions and student performance: the international evidence. *Oxford Bulletin of Economics and Statistics*, 65(2), 117-171.
- Wössman, L., & West, M. (2005). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50, 695-736.

Table 1. Mathematics achievement and student age for countries in TIMSS 1995 and 2003.
Grade 8 (N=22).

Country	Math 1995	Math 2003	Math change	Age 1995	Age 2003	Age change
Australia	507	505	-2	14.04	13.88	-0.16
Belgium (Flemish)	550	537	-13	14.14	14.12	-0.02
Cyprus	468	459	-8	13.74	13.77	0.03
England	498	498	1	14.05	14.29	0.25
Hong Kong SAR	569	586	17	14.18	14.39	0.21
Hungary	527	529	3	14.28	14.51	0.23
Iran	418	411	-7	14.63	14.44	-0.20
Japan	581	570	-11	14.38	14.40	0.01
Korea	581	589	8	14.20	14.60	0.40
Latvia	488	508	20	14.27	15.05	0.78
Lithuania	472	502	30	14.26	14.94	0.68
Netherlands	529	536	7	14.35	14.26	-0.09
New Zealand	501	494	-7	14.00	14.05	0.05
Norway	498	461	-37	13.89	13.80	-0.08
Romania	474	475	2	14.58	14.97	0.39
Russian Federation	524	508	-16	14.03	14.19	0.16
Scotland	493	498	4	13.70	13.68	-0.02
Singapore	609	605	-3	14.55	14.33	-0.22
Slovak Republic	534	508	-26	14.26	14.32	0.05
Slovenia	494	493	-2	13.82	13.90	0.08
Sweden	540	499	-41	14.93	14.89	-0.04
United States	492	504	12	14.23	14.23	0.01

Table 2. Correlations between achievement and age, grade 8 (N=22).

	Math 1995	Math 2003	Math change	Age 1995	Age 2003	Age change
Math 1995	1.00					
Math 2003	0.93	1.00				
Math change	-0.12	0.26	1.00			
Age 1995	0.19	0.14	-0.12	1.00		
Age 2003	0.05	0.16	0.29	0.75	1.00	
Age change	-0.14	0.08	0.58	-0.01	0.65	1.00

Table 3. Mathematics achievement and student age for countries in TIMSS 1995 and 2003, grade 4 (N=15)

Country	Math 1995	Math 2003	Math change	Age 1995	Age 2003	Age change
Australia	495	499	4	9.86	9.89	0.03
Cyprus	475	510	35	9.84	9.90	0.06
England	484	531	47	10.04	10.27	0.23
Hong Kong	557	575	18	10.14	10.24	0.10
Hungary	521	529	7	10.41	10.55	0.14
Iran	387	389	2	10.50	10.40	-0.10
Japan	567	565	-3	10.39	10.41	0.02
Latvia	499	536	37	10.46	11.05	0.59
Netherlands	549	540	-9	10.26	10.23	-0.03
New Zealand	469	493	24	9.98	10.03	0.05
Norway	476	451	-25	9.87	9.81	-0.06
Scotland	493	490	-3	9.71	9.70	-0.01
Singapore	590	594	4	10.31	10.33	0.02
Slovenia	462	479	17	9.87	9.78	-0.09
United States	518	518	0	10.19	10.24	0.06

Table 4. Correlations between achievement and age. Grade 4 (N=15).

	Math 1995	Math 2003	Math change	Age 1995	Age 2003	Age change
Math 1995	1.00					
Math 2003	0.93	1.00				
Math change	-0.16	0.22	1.00			
Age 1995	0.21	0.21	-0.01	1.00		
Age 2003	0.22	0.33	0.29	0.89	1.00	
Age change	0.14	0.38	0.63	0.34	0.72	1.00

Table 5. Correlations between achievement and class size, grade 8 (N=22)

	Math 1995	Math 2003	Math change	Class size 1995	Class size 2003	Class size change
Math 1995	1.00					
Math 2003	0.93	1.00				
Math change	-0.12	0.26	1.00			
Class size 1995	0.42	0.47	0.16	1.00		
Class size 2003	0.52	0.60	0.25	0.87	1.00	
Class size change	-0.03	0.00	0.08	-0.64	-0.17	1.00

Table 6. Correlations between achievement and class size, grade 4 (N = 15)

	Math	Math	Math	Class size	Class size	Class size
--	------	------	------	------------	------------	------------

	1995	2003	change	1995	2003	change
Math 1995	1.00					
Math 2003	0.93	1.00				
Math change	-0.16	0.22	1.00			
Class size 1995	0.34	0.39	0.13	1.00		
Class size 2003	0.53	0.54	0.05	0.90	1.00	
Class size change	0.25	0.16	-0.22	-0.52	-0.10	1.00

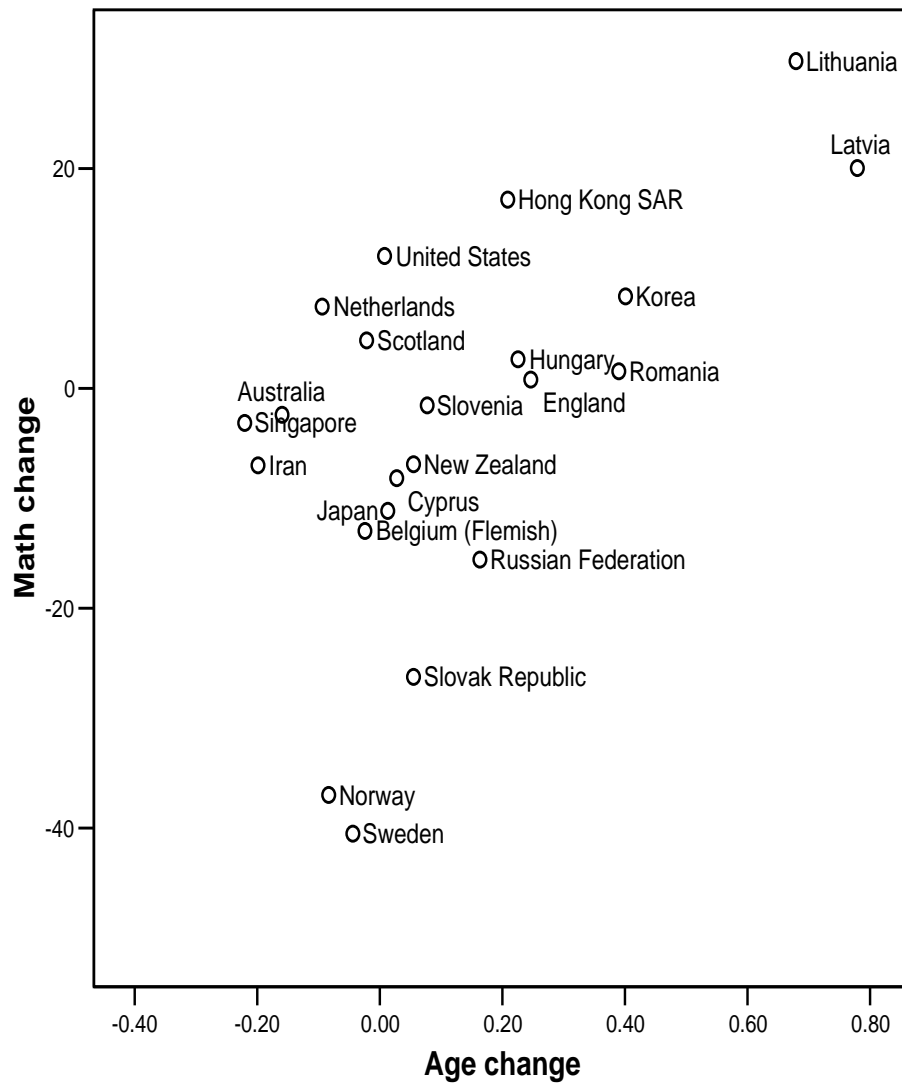


Figure 1. Change in mathematics achievement as a function of change in mean student age between TIMSS 1995 and TIMSS 2003. Grade 8 (N=22).

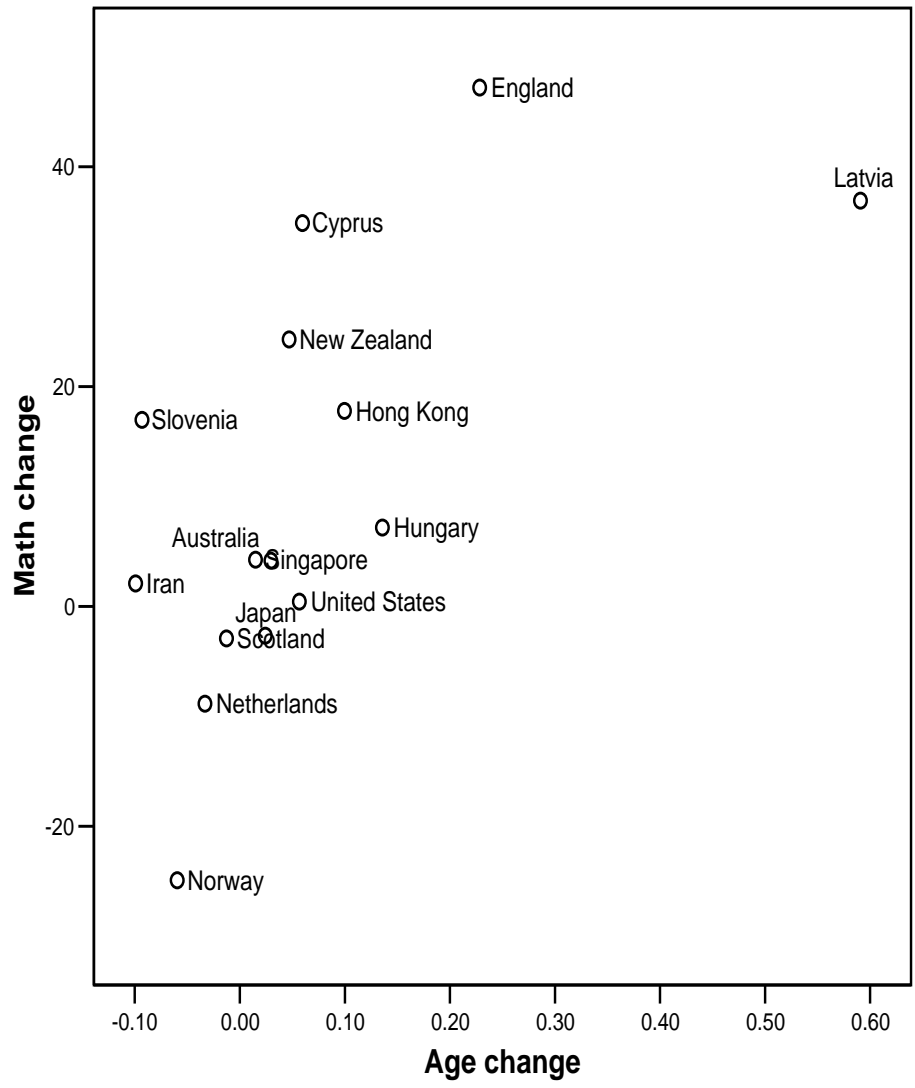


Figure 2. Change in mathematics achievement as a function of change in mean student age between TIMSS 1995 and 2003. Grade 4 (N=15).

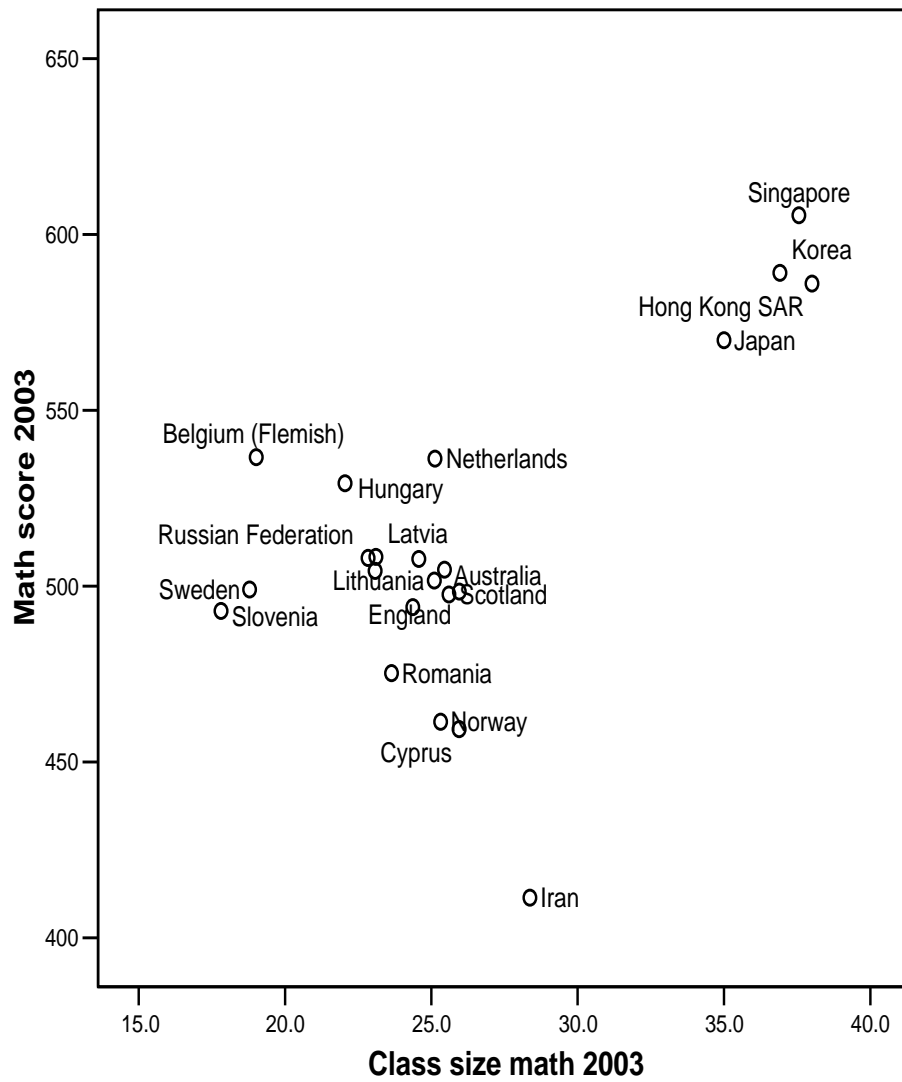


Figure 3. Plot of mean mathematics achievement as a function of mean country class size in TIMSS 2003. Grade 8 (N = 22).