Peer effects in European primary schools: Evidence from PIRLS

Andreas Ammermueller
German Federal Ministry of Labour and Social Affairs (BMAS)

Jörn-Steffen Pischke[*]
London School of Economics

May 2009

Abstract

We estimate peer effects for fourth graders in six European countries. The identification relies on variation across classes within schools, which we argue are formed roughly randomly. The estimates are much reduced within schools compared to the standard OLS results. This could be explained either by selection into schools or by measurement error in the peer variable. Correcting for measurement error we find within school estimates close to the original OLS estimates. Our results suggest that the peer effect is modestly large, measurement error is important in our survey data, and selection plays little role in biasing peer effects estimates.

We are grateful to Aaron Sojourner for extensive discussions on the relationship of his work and ours. We also thank Joshua Angrist, David Autor, Per-Anders Edin, Edwin Leuven, Eric Maurin, Sandra McNally, Matt Pritsker, Kjell Salvanes, Jacob Vigdor, and Gylfi Zoega for helpful comments. Ammermueller gratefully acknowledges financial support from the Fritz Thyssen Foundation under the project "Bildungschancen zwischen Grundschule und Sekundarstufe" and would like to thank the Center for Economic Performance for their hospitality during his stay in London, during which the main work on this paper was completed.

[*] Corresponding author. CEP, LSE, Houghton Street, London WC2A 2AE, UK, s.pischke@lse.ac.uk

1 Introduction

Learning in schools takes place in a group setting, and the composition of the group possibly affects individual outcomes. There has been a lot of interest in these types of social interactions in economics recently, and in peer effects in school in particular. We revisit this issue in this paper, drawing on a previously unexploited data set in this context, the Progress in International Reading Literacy Study (PIRLS) for fourth graders. Our analysis covers six European countries, Germany, France, Iceland, the Netherlands, Norway, and Sweden.

One of the main challenges in the literature on peer effects is the feature that schools and class rooms are not formed randomly. School and class composition typically reflects neighborhood characteristics, and therefore the family background of students. The estimated peer effect may capture unobserved aspects of an individual student's performance if this problem is ignored. We exploit the fact that the PIRLS survey samples multiple class rooms within a single school. This allows us to estimate peer effects within schools. Since we study students in primary schools, there is no explicit tracking in any of the countries in our sample. In section 4 we argue that classes within schools are in fact formed more or less randomly with respect to family background characteristics (other than immigrant status). The variation in our peer variable therefore most likely reflects that there will be small differences in composition when multiple groups are formed out of a small population (in essence the absence of the law of large numbers). Hence, our research design allows for a relatively credible identification of peer effects on student test scores.

The existing literature has used a wide variety of approaches to identify peer effects. The papers closest in spirit to ours are the ones by Vigdor and Nechyba (2004) and Hoxby (2000) for the US, and Gould, Lavy, and Paserman (forthcoming) for Israel. These papers similarly rely on differences in the compositions of individual classes within a school, which come about by chance. Hanushek et al. (2003) and McEwan (2003) also use within school variation to identify peer effects. However, it is more difficult to believe that differences in class composition are random in their cases. We will compare our methodology in detail to the existing literature in the next section. A number of recent studies have also used explicit random assignment to classes or schools, or other natural experiments. However, none of these studies is for European countries.

An important issue in our context is the fact that our peer measure is self-reported by the students' parents and that the sample does not include all students in a class room. These features will lead to measurement error in both the individual and peer level variables in the regression, and hence to biased estimates. Moreover, the size of the bias will differ in the OLS and within school estimates. We address these issues in section 6 in two ways. The first is to use an alternative variable for family background as instrument. The second is to look at the ratio of the peer and individual level family background effect. We show that the within school estimator for this ratio is not affected by measurement error under a simple measurement error model. Nevertheless, even this estimator is biased when not all students in a class are sampled. We adjust our estimates for the sampling error in the peer variable arising from missing students. Clarifying the impact of measurement error on estimates of peer effects is an important contribution of our paper. The only other paper in the literature dealing with missing observations in peers' models is concurrent work by Sojourner (2008). His analysis complements ours since he allows for more general processes generating missing students but he does not analyze measurement error explicitly.

On average across countries, we find that a one standard deviation change in our measure of peer composition leads to a 0.17 standard deviation change in reading test scores, and this estimate is marginally significant. The size of this effect is slightly larger than most estimates reported in the literature. However, the confidence interval for our measurement error corrected estimates is fairly large. Moreover, the pattern of our results is largely consistent with a story based purely on measurement error, while systematic selection into schools by family background seems to play little role.

2 Empirical framework and existing literature

Like many peer effects studies, we start from a reduced form specification of an education production function of the form

$$y_{ics} = \alpha_s + \beta X_{ics} + \gamma S_{cs} + \lambda \overline{X}_{(-i)cs} + \mu_{cs} + \varepsilon_{ics} \qquad (1)$$

where y_{ics} is a student outcome, like a test score, for student *i* in class room *c* and school *s*, X_{ics} are student or family characteristics, like sex, family background, etc., S_{cs} are school or class level characteristics, like class size, teacher experience, characteristics of the municipality, etc., and $\overline{X}_{(-i)cs}$ are the average characteristics of the peers of student *i*. In addition, μ_{cs} and ε_{ics} are a class level and an individual level error term. The reduced form is silent on how the peer effect arises. In the language of Manski (1993), λ could either capture exogenous (or contextual) or endogenous effects. Exogenous effects arise when individuals learn more because the group of peers is more favorable in terms of their predetermined characteristics. Endogenous effects arise when individuals learn more because peers are learning more. We make no attempt at separating these.

μ_{cs} reflects correlated effects. Correlated effects arise when the group of peers is subject to a common influence, which is not modeled directly. These effects will give rise to a bias if they are correlated with peer group composition. For example, consider a remedial class room with relatively poorly performing children. This class room may be assigned a particularly able teacher but the exceptional characteristics of this teacher are not observable. Removing the potential bias from correlated effects is one of the main challenges in the peer effects literature.

If some relevant school or class room characteristics are not controlled, the estimated peer effect $\hat{\lambda}$ will be biased. Random assignment of students and teachers to class rooms solves this problem, because random assignment breaks the link between peer characteristics and extraneous effects on the class, like unobserved teacher quality. Boozer and Cacciola (2001) and Graham (2008) exploit the random assignment in the Tennessee STAR experiment on class size. Cullen, Jacob, and Levitt (2006) use lotteries at oversubscribed Chicago public schools. However, their paper does not focus on the issue of peer effects.

True random assignment variation is rare in an education context, and unavailable in many countries. Hence, researchers have to resort to other strategies utilizing the existing data. In this paper, as in a variety of related studies, we use variation within schools in order to identify the peer effect. This means that we include school fixed effects α_s in our regression (1).[1]

The idea behind this strategy is the observation that different schools draw students from different neighborhoods, and hence family backgrounds. Hence, the unobserved characteristics μ_{cs} will be systematically related to $\overline{X}_{(-i)cs}$ at the school level. However, students are not generally grouped into classes on the basis of ability or family background in primary school. Although some countries, like Germany, track students into a rigid system of separate schools at the secondary level, there is no system wide tracking at the primary level. In fact, classes in primary schools with multiple class rooms at the same grade level are typically formed more or less on a random basis. In this case, $\overline{X}_{(-i)cs}$ will be uncorrelated with the class level shocks μ_{cs} conditional on a set of school fixed effects, or the characteristics of school peers. The bias from correlated effects is thus removed and λ can be estimated consistently.

In order to make this argument more precise, consider the following simple model generating student characteristics:

$$X_{ics} = \eta_{cs} + v_{ics} \qquad (2)$$

i.e. student characteristics consist of a common class room level mean η_{cs} and an idiosyncratic, mean zero student level component v_{ics}, which is uncorrelated with η_{cs} and ε_{ics}. The peer mean is

$$\overline{X}_{(-i)cs} = \eta_{cs} + \overline{v}_{(-i)cs} \qquad (3)$$

Correlated effects arise whenever cov(η_{cs}, μ_{cs}) $\neq 0$. Estimates of both β and λ will therefore be biased in the estimation of equation (1). Our basic identifying assumption is $\eta_{cs} = \eta_s$, i.e. the systematic component of the student background characteristic arises only at the school level but not at the class level. Random

[1] Alternatively, we could introduce peer variables at the school level directly into the estimating equation. Both approaches lead to very similar results.

assignment of students and resources to classes within schools would ensure that this condition is met. Let the operator $\tilde{\Delta}$ perform the within transformation, so that $\tilde{\Delta} a_{ics} = a_{ics} - \overline{a}_s$. Hence, peer characteristics within schools are $\tilde{\Delta}\overline{X}_{(-i)cs} = \tilde{\Delta}\overline{v}_{(-i)cs}$, i.e. variation in the peer measure comes only from the fact that $\overline{v}_{(-i)cs} \neq 0$ in small groups. A necessary condition for the within school estimation to work is, of course, that there is sufficient variance in peer composition of a class room within a school, which is the case in our data, see section 3 below.

Our identification strategy is most closely related to that of McEwan (2003) and Vigdor and Nechyba (2004, 2006). McEwan (2003) studies peer effects for 8th graders in Chile. However, random assignment to classes within schools is much less likely to happen at the secondary level because schools in many countries, including Chile, track students to at least some degree. If there is tracking on the basis of (unobserved) ability, estimates of λ are still confounded by correlated effects. Vigdor and Nechyba (2004) also rely on school fixed effects for identification. Their results are for 5th graders in North Carolina, an age group where tracking is less of an issue. However, they report evidence that class room assignment does not look random in most schools. Hence, in their preferred estimates, they restrict themselves to a subsample of schools where class room assignment looks random based on preliminary tests. However, this pre-testing approach is not completely satisfying. In this paper, in contrast, we argue that class room assignment is random in European schools for institutional reasons, and we successfully verify this claim with similar tests to those employed by Vigdor and Nechyba (2004).

The papers by Clotfelter, Ladd, and Vigdor (2006) and Vigdor and Nechyba (2006) cast some doubt on their main identifying assumption of random class room assignment within schools. Using the same data for North Carolina elementary schools, Clotfelter et al. (2006) present some evidence that better teachers are assigned to classrooms with better students, even within schools. This may be due to “teacher shopping” by parents or to the ability of better teachers to avoid assignment to classes with more poorly performing students.

In an alternative approach, Vigdor and Nechyba (2006) find positive and significant peer effects in models with school fixed effects. They then go and introduce teacher-fixed effects, hence comparing successive cohorts of students assigned to the same teacher. The introduction of teacher-fixed effects in addition to school-fixed effects leads to significantly negative estimates of peer effects. This suggests that random assignment of students to class rooms does not seem to be satisfied in the North Carolina context, and their results casts some doubt on their own earlier findings (Vigdor and Nechyba, 2004). We are less worried about their findings for the case of European primary schools because we believe that the practice of "teacher shopping" or the purposeful assignment of good teachers to better classes is absent or comparatively unimportant in the countries we analyze.

Gould et al. (forthcoming), Hanushek et al. (2003), and Hoxby (2000) also use within school variation to identify peer effects. The Gould et al. and Hoxby studies are similar in spirit to ours. We use comparisons across class rooms within the same grade for the same cohort of students. Hoxby uses comparisons between classes in the same grade across adjacent cohorts and years. Hence, she identifies peer effects from variation arising from the composition of subsequent cohorts. For example, one cohort may have more girls and the next cohort fewer for purely random reasons. Gould et al. also use data on multiple cohorts in the same grade. They condition on the student composition of the grade across multiple cohorts of students. Effectively, like Hoxby, they therefore exploit year to year variation in the composition of successive cohorts of students. However, these studies tend to focus on different peer group measures than ours. Hoxby looks at gender and race composition of the class room and performance by opposite gender and race groups, while Gould et al. look at the share of immigrants.

Hanushek et al. (2003) focus on a peer measure more similar to ours. They also control for school by grade effects like Hoxby and Gould et al. However, they track the same cohort of students over time, rather than different cohorts, and they also control for student fixed effects. This means that they effectively only consider changes in the peer group which come about through changes in a student's class room assignment over time and not changes in cohort composition, as in Hoxby and Gould et al. Including student fixed effects should exacerbate any problems from the non-random assignment of teachers to class rooms highlighted by Clotfelter et al. (2006). Hence, controlling for student fixed effects may lead to a larger upward bias in the estimates when there is "teacher shopping." In fact, Hanushek et al. find an increase in the peer coefficient when they control for individual student effects compared to a similar specification without individual effects.

The previous literature finds peer effects which range from close to zero (Cullen, Jacob and Levitt, 2006) to effects of a one standard deviation change in the peer measure of about 0.5 (Hoxby, 2000, Boozer and Cacciola, 2001). The results of many other studies fall within this range but are clustered around the bottom end.

3 Data and descriptive statistics

Thirty-five countries participated in the Progress in International Reading Literacy Study (PIRLS). This study was conducted by the International Association for the Evaluation of Educational Achievement (IEA) in 2001 by testing nine- and ten-year-olds in reading literacy. Extensive information on home and school environment is available from student, parent, teacher, and school questionnaires. With 150,000 students tested, PIRLS 2001 is the first in a planned 5-year cycle of international trend studies in reading literacy (Mullis et al., 2003).

The data are collected in a two-stage stratified sampling design. First, participating schools were chosen. Therefore, the schools are the primary sampling units and not the classes or students. Within each school, a sample of classes from the targeted grade was drawn. The targeted grade is the upper of the two grades with the most 9 year-olds at the time of testing. This is always the fourth grade in our sample of countries. Within each class, in principle, all students are sampled. In practice, the number of sampled students can be smaller than the actual class size because of student non-participation. We use all European countries with a sufficient number of schools with at least two classes. These are France, Germany, Iceland, the Netherlands, Norway and Sweden.

Student performance is measured by test scores in reading literacy, which Campbell et al. (2001) describe as “one of the most important abilities students acquire as they progress through their early school years. It is the foundation for learning across all subjects.” The test scores are plausible values that are drawn from an estimated proficiency distribution. Plausible values are imputed scores based on the students’ answers to the test items (cf. Mislevy, 1991). The scores have then been standardized to an international mean of 500 and a standard deviation of 100, which facilitates the comparison across countries. The reliability of the PIRLS testing instrument across 10 versions of the test ranges from 0.83 to 0.89 in our sample countries (Mullis et al., 2003).

Table 1 provides information on mean reading scores and sample sizes in PIRLS at the student, class, and school level. Students, classes, and schools can be directly identified. Missing values of student background, class, and school variables are a serious problem in the data set. For parents' education, 36 percent of all values are missing. Instead of parents' education, we use the number of books at home as our indicator of family background. Among the variables reflecting family background, this is the one with best item response rate. In addition, this is an appealing variable in its own right. It is highly correlated with parental income, education, and origin. The variable also reflects whether the parents value literary skills. Parents who own many books most likely will also promote reading among their children. In fact, Wößmann (2008) found the number of books to be the single most important predictor of reading skills among various family background variables in the Third International Math and Science Study (TIMSS) and Ammermueller (2005) in PIRLS and the Programme for International Student Assessment (PISA) data. Another advantage of the of the books at home variable is that it is asked of both parents and students, allowing us to use IV estimation in order to address potential measurement error in the variable.

Table 1 demonstrates that the sample size, conditioning on non-missing student background and school variables, shrinks to about 40 to 75 percent of the original. The row labeled "no. of students (sample)" gives the sizes of the samples we actually use. All figures in this row and below refer to the sample with no missing values. Reading scores in the selected samples are slightly higher than in the overall sample, as can be seen in the first two rows in the table. Some sample schools have only one class. Our within school estimates will only be utilizing the schools with two or more classes. Information on the students, classes, and schools with more than one class can be found in the bottom rows of Table 1. The peer effects estimations have also been performed including all observations for which test scores are reported. Missing values have been replaced by zeros and dummy variables for missing values for each variable have been added to the regressions. The estimated peer effects are comparable to the results presented below.

The home questionnaire asked parents to report the number of books in their home in five categories: none or few books (0 – 10), enough to fill one shelf (11 – 25), enough to fill one bookcase (26 – 100), enough to fill two bookcases (101 – 200), enough to fill three or more bookcases (more than 200).[2] In order to form a single measure of students' background, after some experimentation, we chose a simple index which assigns 1 to the lowest category (0 – 10), and 5 to the highest category (more than 200). The median parent reports 26 – 100 or 101 – 200 books, and the mean of the indices range from about 3.3 to 4, depending on the country (see Table 2 below).

We generated peer variables as the class average of five student background variables: number of books at home, student's sex and age, whether at least one parent was born abroad, and whether a foreign language is spoken at home. There is a discussion in the literature on peer effects whether class rooms or schools (or possibly even neighborhoods) are the more appropriate unit of peer interactions. Of course, peer interactions may occur at each of these levels, and it is an open question which is the most important. We focus on the class room level for the pragmatic reason that we want to analyze differences within schools. In the within school estimates, all peer interactions with students from other classes in the school will be absorbed into the school fixed effects. However, peer effects in the class room are clearly of interest for academic outcomes, since classes are the basic unit where learning takes place. It is therefore natural to expect that a large fraction of total peer effects should arise at the class room level.

The peer averages are formed using information for all students who report a value for this specific variable in the data set, not just the students in the final sample. In Table 2, we decompose the total variance in these class averages into the parts of the variance within and between schools using the relationship

$$\frac{1}{C}\sum_{s=1}^{S}\sum_{c=1}^{C_s}\left(x_{cs}-\overline{x}\right)^2 = \frac{1}{C}\sum_{s=1}^{S}\sum_{c=1}^{C_s}\left(x_{cs}-\overline{x}_s\right)^2 + \frac{1}{C}\sum_{s=1}^{S}C_s\left(\overline{x}_s-\overline{x}\right)^2 \qquad (4)$$

[2] Using instead the number of books at home reported by students yields comparable results.

where x is the specific variable we are interested in, $s = 1, 2, \ldots, S$ is a school indicator, $c = 1, 2, \ldots, C_S$ is a class indicator, and there are C_S classes in school s. C is the total number of classes across all schools in our sample.[3]

Table 2 presents the total, between, and within school variance of the peer variables. The variation for the average reading test score is shown as well. It is obvious that most of the variance in all of these measures is between schools.[4] Between 7 and 18 percent of the variance in the index for the number of books at home is within schools. The fraction is higher for the reading test scores. However, 70 percent or more of the peer variation in test scores is also between schools. This suggests that a large part of the variation in all these measures is accounted for by school effects. Nevertheless, there is also some non-negligible amount of variance left within schools.

4 Selection in class room formation

In this section we will discuss the assignment of students both between and within schools. We start by presenting some basic information on primary schooling in the countries we study. We then go on to present some evidence from the PIRLS data to shed light on the question whether classes are formed (more or less) randomly, and whether different class rooms systematically get different resources.

In all six countries in our sample, students attend a single track primary school from school enrolment to at least grade four, in which students have been tested in PIRLS.[5] While students are assigned to various school types after grade four in Germany, they stay on for at least two more years in primary school in most other countries (France, Iceland, the Netherlands, and Sweden) or go on to a single tracked secondary school (Norway). School choice at the primary level is unrestricted in some countries (Germany and the Netherlands) while school assignment depends on the place

[3] For the variance decomposition to add to the total variance in an unbalanced panel, it is necessary to weight the between component by the number of classes in the sample. This is not what, for example, the Stata `xtsum` command calculates.

[4] The reader thinking of individual level variation in student performance may be surprised by this. Most student level variation is within schools. However, most of this variation is also within classrooms, and we consider the variation in classroom means here.

[5] The information on the schooling systems is taken from Eurybase, the database in the information network on education in Europe, http://www.eurydice.org.

of residence in the other countries. However, parents have some means to influence the choice of schools also in these countries. In practice, most parents choose the nearest school for convenience in all countries (or live near the school of their choice). The heads of the school are responsible for the assignment of students to classes within schools. Most countries have legal rules on maximum class size and some school systems provide extra resources for schools with a high share of immigrant students. The final responsibility in assigning students to classrooms lies with the heads of the school, however. Grouping of students seems to happen in some cases based on the migration background of students. Most of our sample countries do not use any explicit grouping of students by ability in primary school. The Netherlands and Sweden have the most decentralized systems, and schools are relatively free to decide how to form classes. In practice, students are mostly grouped by age in the Netherlands, although classes are sometimes formed by proficiency across age groups. In Sweden, class groups may not be fixed and ability grouping might happen for short periods of time (Mullis et al., 2002). For Iceland, the Compulsory Schools Act of 1995 states explicitly that there is no selection or streaming by ability of students.[6]

In order to corroborate that these institutional descriptions translate into more or less random assignment of students to classes, we conducted a small survey by email among heads of primary schools in Germany. The results in Table 3 confirm that heads of schools are primarily responsible for forming class rooms, often together with a teacher. The composition of classes does not usually change during the first four years of school for three quarters of all respondents. When classes are rearranged, this is mostly due to a large number of newly arriving students. Individual students who are disruptive in their current class may be allocated to other classes by the head of the school. Additional information from open ended responses provides no indication that students change classes on their behalf or for other forms of "teacher shopping." Classes are actually mostly formed so that they are well balanced (94 percent). Students from the same neighborhood or kindergarten are put in the same class in a third of all school. Only six percent of schools mention grouping students of similar abilities.

The PIRLS data also asked in the schools questionnaire whether the school forms sample classes on the basis of ability. The last row in Table 1 reports the fraction

[6] We consulted researchers in each of the sample countries and they also confirmed the impression that

of students in schools that report some ability grouping at the class level. This fraction is very low except in France and the Netherlands, where it reaches in the order of 30 percent. While we do not find much evidence that the classes in these tracked schools look very different from classes in other schools, we also show results excluding these schools which report some tracking.

We investigate two separate and distinct questions about class room formation with the PIRLS data. The first, and most important, question is whether classes within schools are being formed randomly. The second question is whether class rooms which differ in composition, for random or non-random reasons, receive different resources.

In order to test whether class rooms are formed randomly with respect to a particular student characteristic, we perform a series of Pearson χ^2 tests. If classes are formed randomly, the student characteristic under study and the class the student is assigned to should be statistically independent. Consider student sex, for example. The Pearson χ^2 test asks whether there are more females in a particular class than is consistent with independence, given the number of students in the school. Formally, for each school the test statistic is given by

$$P = \sum_c \sum_j \frac{\left(n_{cj} - \hat{n}_{cj}\right)^2}{\hat{n}_{cj}} \tag{5}$$

where n_{cj} is the number of students with characteristic j = 1, …, J in class room c = 1, …, C_S. Define

$$n_{c\bullet} = \sum_j n_{cj} \qquad n_{\bullet j} = \sum_c n_{cj} \qquad \hat{n}_{cj} = \frac{n_{c\bullet} n_{\bullet j}}{\sum_c \sum_j n_{cj}}$$

where $\hat{n}_{cj}$ is the predicted number of students with characteristic j in class room c when characteristic and class room are independent. Then, under the null hypothesis of independence, $P \sim \chi^2$ with $(C_S - 1)(J - 1)$ degrees of freedom.

ability grouping would be rare.

We further assume that the *S* schools in a country are independent. In this case, we can simply add up the *S* test statistics to get an aggregate test statistic with $\left[\sum(C_s-1)\right](J-1)$ degrees of freedom (see, e.g. DeGroot, 1984, p.384). Obviously, the test can only be carried out on the sub-sample of schools with two or more class rooms. We found in a small Monte Carlo experiment that the test generally performs well although it rejects somewhat too often under the null. The empirical rejection rate for a 5 percent nominal size is about 0.13. On the other hand, the test seems to have good power.[7]

Table 4 presents the *p*-values for these tests in the first row of each panel for various different student characteristics. The *p*-values for books at home are well above the 5 percent level except for Sweden, where the *p*-value is 0.036 (we find such a *p*-value about 10 percent of the time in the simulations under the null). We also find evidence of non-random assignment of non-native language children for Sweden and possibly Germany. Recall that principals in a significant number of schools in France and the Netherlands report ability grouping in their schools. The *p*-values differ only slightly when we split the sample between the schools reporting tracking in France and the Netherlands and those which don't. One exception is the evidence for sorting by age within the 19 schools that may be tracked in France. Overall, we conclude that there is little evidence for systematic formation of class rooms, particularly with respect to our family background measure books at home. Sweden might be the only exception.

Even if class rooms are formed randomly, they may still differ in systematic ways because school resources also have to be distributed to classes. The assignment of class room resources may not be random, even if classes are formed randomly. For example, a class may end up with more children from less advantaged family backgrounds purely by chance, and the school might assign this class a better (or a worse) teacher. Our estimates of the peer effects would be biased if this happened systematically across schools.

In order to shed light on this question, we ran a set of regressions of the peer variables described in the previous section on class room and teacher characteristics. The observable characteristics of class rooms are class size and its square, teacher

[7] Details on the simulation study are available from the authors upon request.

gender, education, experience, and its square. Table 4 shows *p*-values for the corresponding *F*-tests on the joint significance of these variables from a regression including school fixed effects in the second row of each panel. For our family background variable of interest, the number of books at home, the class variables are insignificant, except in Iceland and in Sweden. In the case of Iceland, it turns out that this correlation is solely driven by a single class room with a teacher with 20 years of experience (while all other teachers in Iceland have 10 or fewer years of experience). We discount this result as spurious. In the case of Sweden this seems indeed to indicate a non-random allocation of class room resources to classes with students from different backgrounds, even within schools. In particular, there is evidence that class size increases with average background of students in a class. The coefficients for the other class and school variables are not significant.

We also find some evidence that class rooms differ for students by age (in Germany, Iceland, and Norway) and by student sex (in Iceland, the Netherlands, and Norway). It also seems fairly clear that classes are different for students not speaking the native language at home in most our sample countries. The higher the share of immigrant students in a class, the lower is teacher's education in Germany and Norway. In Sweden, there is weak evidence for an allocation of immigrant students to larger classes.

Our results largely confirm that classes in the sample countries seem to be formed roughly randomly within schools. There is little evidence that students of different family backgrounds are more likely to be grouped in certain classes conditional on the school they attend, or that classes with different compositions receive different (observable) resources. This is comforting for our analysis. The only country, where this does not seem to be the case, is Sweden. Hence, the Swedish results may have to be taken with a grain more of salt. But the Swedish results turn out to be very close to the average of the other countries so that this does not seem to matter for our findings in practice. In addition, immigrant children, which are an important group in all of the sample countries, also seem to be non-randomly assigned and given different teaching resources. Nevertheless, we do not find any evidence that the non-random sorting of immigrant children to classes affects our results on the books at home variable.

5 Basic results on peer effects

We now turn to our results on peer effects. Table 5 summarizes the results for the six countries. Our family background and peers variable, books at home, takes on five values. The most flexible way to use this variable is to create a set of four dummy variables, and correspondingly the fraction of peers in these four categories. Since this leads to a large number of coefficients, and given that the coefficients estimates are roughly monotonically increasing in the categories, we have chosen to simply create an index taking on the values 1 – 5 which we created from these five categories.[8] In each case, the peer variable for student *i* used in the regressions is the leave-out mean for the classroom, omitting the value of the variable for student *i* from the calculation of the mean.

We find a relatively consistent pattern of results for all six countries in our sample in Table 5. The size of the estimated peer effect is similar across the specifications with and without school and class level variables, and is in the order of 15 to 22 for moving peer quality to the next higher category. Only in Norway does the peer coefficient fall when school level covariates are added to the regression. Once we include school fixed effects in column (4), the effect always falls, although the amount of the change is different across countries. In Germany, the Netherlands, and Norway the peer effect weakens the most in these specifications, while there is little change in France. Excluding schools that form classes based on student ability predictably only changes the results in France and the Netherlands, the two countries with moderate shares of students in schools which form classes based on ability (see Table 1). Curiously, estimated peer effects are larger when the schools which report tracking are excluded for the Netherlands.

One reason for the high variation in the coefficients from the fixed effects models is that the standard errors of these estimates are reasonably large, so that the effects for each individual country cannot be estimated very precisely. If we believe that the peer effects are the same in each country then it makes sense to combine the estimates into a single estimate. The average of the six coefficients in the fixed effects specification in column (4), weighted by the inverse of the sampling variance, is 7.6. If the variation in country level estimates around this overall mean is only due to sampling variation, then the standard error for the meta-estimate is 3.2.[9] This estimate is much more precise than the country level estimates, and it is significant at the 5 percent level. One concern is with the results for Sweden, because we found some evidence for non-random assignment and targeted class room resources for Sweden above. The meta-estimate for the countries without Sweden is only slightly lower.

Our results show that standard OLS estimates of the peer effect may be biased upward substantially if the within school results are indeed reliable estimates of the true peer effect. One reason why even the fixed effects estimates may be biased is the presence of immigrant children. We showed above that immigrant children are often not randomly assigned to classes within schools, and the classes with many immigrant children may get different resources. Since immigrants in these countries tend to be of lower SES (the index for books at home is on average 3.15 for immigrant families in the six countries but 3.56 for non-immigrant families), part of the peer effect may be explained by the non-random allocation of immigrants.

In order to probe this, we reran the regressions in Table 5 including the fraction of foreign born children in the class, and the fraction of children speaking a foreign language at home. This attenuates the estimated peer effects at most very slightly.[10] We also experimented with regressions on the sub-sample of schools with few immigrant children. However, most sample countries have enough immigrants that there are relatively few such schools leading to small samples, and hence imprecise estimates. These results indicate that the effect of immigrant children in a class seems to be relatively well captured by our family background variable.

A further question is whether peer effects vary across students. This could give insights into the optimal assignment of students to classes. When students from a lower

[8] We also experimented with assigning midpoints to the categories to form an alternative cardinal variable with roughly similar results.

[9] The sampling variance of the mean is obtained as $v = \left[\sum v_c^{-1}\right]^{-1}$, where v_c is the sampling variance of the estimate for country *c*. One interpretation of this calculation is that the country average is the minimum distance estimate of the common peer effect across countries.

social background profit more from their peers' background than students from a high social background, more heterogeneous classes would benefit overall performance (Glewwe, 1997). To investigate this, we add interaction effects between the peer variable and the individual variable books at home to the regressions presented in Table 5. Since about half the students have more than 100 books at home, we interact the peer average with a dummy indicating whether the individual reports more than 100 books at home. The results are presented in Table 6. Peer effects seem to be stronger for students with a higher social background in France and the Netherlands, while they are stronger for students with a lower social background in Sweden. The meta-estimate of the interaction is small and insignificant.

6 Measurement error and missing students

Survey reports are subject to a lot of measurement error. In our case, measurement error in the books at home variable implies that there is measurement error in both the individual and the peer level regressor. In addition, the peer measure is not based on all students in a class because some students have not been sampled and others have not responded to the respective question. This problem will also arise in many studies based on administrative data, which frequently use lagged test scores as peer measure, since test taking may be incomplete or lagged scores cannot be matched to all students. Both these measurement problems will interact in leading to biased estimates of the peer effect in a non-standard way.

In order to investigate the impact of measurement error in our setup we will return to the model we outlined in equations (1) to (3) above. In order to focus on the variables of interest, consider a simplified version of equation (1) with only the individual level and the peer group regressor but no other covariates:

$$y_{ics} = \beta X_{ics} + \lambda \overline{X}_{(-i)cs} + \varepsilon_{ics} \,. \tag{6}$$

[10] Including measures from the school questionnaire on the share of economically disadvantaged students or the share of students leaving before the end of the academic year as alternative school level variables

Moreover, to focus on the role of measurement error we set $\mu_{cs} = 0$, i.e. the error term has no class or school level component. Hence we abstract from the biases arising from correlated effects. In practice, these might of course exist on top of any biases from measurement error.

The student background variable X_{ics} is still given by equation (2) but this variable is not directly observed. Instead we observe

$$\tilde{X}_{ics} = X_{ics} + u_{ics} = \eta_{cs} + v_{ics} + u_{ics} \tag{7}$$

where u_{ics} is a classical measurement error. Moreover, the observed peer variable is only computed from the subset of observed peers, while students are actually affected by all peers in (6).

We do not assume that students are missing at random. Instead our derivations assume that the v_{ics} for missing students are drawn from a distribution that may differ from the distribution for the observed students but this distribution is independent of class room assignment or of ε_{ics}. This allows for the possibility, for example, that the probability of a student being missing depends on the student's background characteristics.[11]

Our argument above has been that the common component of student background η_{cs} only arises at the school level. Hence, we can think of our standard OLS results corresponding to those with $\sigma_\eta^2 > 0$ and the within school results to $\sigma_\eta^2 = 0$, because this component has been absorbed by the fixed effects.

In this setup, the OLS estimate of λ will converge to

$$\operatorname{plim} \hat{\lambda}_{OLS} = \beta \frac{(\bar{n}-1)\sigma_u^2 \sigma_\eta^2}{(\sigma_v^2 + \sigma_u^2)(\bar{n}\sigma_\eta^2 + \sigma_v^2 + \sigma_u^2)} + \lambda \frac{\frac{\bar{n}-1}{\bar{N}-1}\sigma_v^2(\sigma_\eta^2 + \sigma_v^2 + \sigma_u^2) + (\bar{n}-1)\sigma_\eta^2(\sigma_v^2 + \sigma_u^2)}{(\sigma_v^2 + \sigma_u^2)(\bar{n}\sigma_\eta^2 + \sigma_v^2 + \sigma_u^2)} \tag{8}$$

did not change the results.

[11] Sojourner's (2008) work and discussions with the author first alerted us to the possibility that assumptions weaker than missing at random are feasible when students are (quasi-) randomly assigned to classrooms.

as we show in the appendix. $\overline{N}$ is the average number of students in a classroom and $\overline{n}$ is the average number of students sampled in each class, and all the variances refer to the distributions of the relevant variables in the sub-population of observed students.

In order to understand the different sources of measurement error and the sign of the bias, it is instructive to look at some special cases. First, consider the case where all students in each class are sampled, so the only problem is classical measurement error. In this case

$$\operatorname{plim}\hat{\lambda}_{OLS} = \beta\frac{\left(\overline{N}-1\right)\sigma_u^2\sigma_\eta^2}{\left(\sigma_v^2+\sigma_u^2\right)\left(\overline{N}\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right)}+\lambda\frac{\sigma_v^2\left(\overline{N}\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right)+\left(\overline{N}-1\right)\sigma_\eta^2\sigma_u^2}{\sigma_v^2\left(\overline{N}\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right)+\sigma_u^2\left(\overline{N}\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right)} \tag{9}$$

It is easy to see in this formulation that the second term implies an attenuation bias of λ if there is classical measurement error in X_{ics}. This measurement error will carry over to $\overline{X}_{(-i)cs}$, and lead to the standard attenuation. Since λ is likely positive, this will imply an underestimate of λ. Returning to equation (8), it becomes clear that the attenuation is greater, when some students in the class are not sampled. If $\sigma_\eta^2 > 0$ a second component of the bias arises, and this is captured by the first term in equations (8) or (9). The individual level regressor X_{ics} is also subject to error, which will lead to an attenuation of the estimated $\hat{\beta}$. Since the peer variable $\overline{X}_{(-i)cs}$ contains information on η_{cs}, part of the signal in the individual level regressor will load on to the peer coefficient. This term is positive, and hence yields an upward bias.

Because of these two conflicting sources of bias it is impossible to tell what the net effect of the bias on $\hat{\lambda}_{OLS}$ is. The first term can dominate when β is sufficiently large compared to λ. Hence measurement error may not lead to an underestimate of the peer effect in the standard OLS specification.

The within school model corresponds to the case where $\sigma_\eta^2 = 0$, the first term in equations (8) and (9) vanishes and we have

$$\operatorname{plim}\hat{\lambda}_W = \lambda\left(\frac{\bar{n}-1}{\bar{N}-1}\right)\frac{\sigma_v^2}{\left(\sigma_v^2+\sigma_u^2\right)}, \tag{10}$$

so that the peer effect is now underestimated. Hence, measurement error alone may explain why we find lower peer effects in the fixed effects estimates in Table 5.

Furthermore consider the within estimator of the individual level covariate

$$\operatorname{plim}\hat{\beta}_W = \beta\frac{\sigma_v^2}{\left(\sigma_v^2+\sigma_u^2\right)}.$$

The bias in this coefficient is just the standard classical attenuation bias. Moreover, the attenuation bias terms $\sigma_v^2/\left(\sigma_v^2+\sigma_u^2\right)$ are the same in the expressions for $\operatorname{plim}\hat{\beta}_W$ and $\operatorname{plim}\hat{\lambda}_W$. Since $\bar{N}$ and $\bar{n}$ are observable in our data, this yields

$$\operatorname{plim}\frac{\hat{\lambda}_W}{\hat{\beta}_W}\left(\frac{\bar{N}-1}{\bar{n}-1}\right)=\frac{\lambda}{\beta} \tag{11}$$

which suggests that the ratio of the coefficient on the peer variable and the individual level background variable can be estimated consistently. It tends to be difficult to interpret the magnitudes of the peer effect estimate in any case. One way to facilitate this interpretation is to look at this ratio.

The more standard way to address the measurement error problem is to rely on instruments for both X_{ics} and $\overline{X}_{(-i)cs}$. Recall that in our case the background variable X_{ics} is the parents' report of the number of books at home. The same question was asked of the students as well, so we use the students' report of the number of books at home as our instrument for the parents' report, and the peer mean of the students' report as instrument for the peer variable. Of course, the errors in parents' and students' reports may well be correlated. Nevertheless, using independent reports by different individuals on the same variable and assuming independent errors is a standard strategy in the literature when such measures are available (see, e.g. Ashenfelter and Krueger, 1994). We therefore pursue this avenue here as well.

In the classical measurement error case with an unbounded support for X_{ics}, the IV estimate of λ will converge to

$$\operatorname{plim}\hat{\lambda}_{IV} = \lambda\left(\frac{\bar{n}-1}{\bar{N}-1}\right)\frac{\bar{N}\sigma_{\eta}^{2}+\sigma_{v}^{2}}{\bar{n}\sigma_{\eta}^{2}+\sigma_{v}^{2}}. \tag{12}$$

This turns out to be the same as the expression in equation (8) with $\sigma_{u}^{2}=0$, so IV solves the standard measurement error problem. It does not resolve the attenuation in the peer effect that arises due to the fact that we do not sample all the students in a class. For the within estimator, equation (12) becomes

$$\operatorname{plim}\hat{\lambda}_{IV,W} = \lambda\left(\frac{\bar{n}-1}{\bar{N}-1}\right). \tag{13}$$

This again suggests that the within school IV estimate is simple to adjust for the sampling bias using the actual means $\bar{N}$ and $\bar{n}$ in our data. Our adjusted IV estimator will therefore be

$$\hat{\lambda}_{IVadj} = \hat{\lambda}_{IV}\left(\frac{\bar{N}-1}{\bar{n}-1}\right). \tag{14}$$

The first stages corresponding to our IV regressions indicate that both the relevant instruments for the individual level regressor and for the peer variable are always highly significant. The *t*-statistics on the students' report of books at home are above 7 and typically above 10, and the corresponding *F*-statistics are also large.[12] This indicates that our IV models are not likely to suffer from any small sample bias.

One important caveat to these derivations is of course that our background and peer variable, books at home, is categorical, and hence has bounded support. In this case, measurement error will by necessity be non-classical. Moreover, Kane, Rouse, and Staiger (1999) point out that the IV estimator is biased upwards when the mismeasured

[12] The only exception is the Netherlands, where the instrument for the peer variable has a *t*-statistic of 3.15.

regressor is binary. The same will be true if the regressor is multivalued but bounded. For our application this implies that the IV estimates may actually be biased upwards. In this case, once we control for school fixed effects, OLS and IV would bracket the true result. On the other hand, as we discussed above, mistakes in parents' and children's reports of books at home may be correlated. This would bias the IV estimates towards OLS, and the true peer effect could therefore be larger than the IV result.

Before turning to our results, it is important to point out that Sojourner (2008), in an independent and complementary analysis, also considers the estimation of peer effects with missing students.[13] His setup allows for more general processes which generate missing students. In particular, Sojourner's results are valid under our assumptions and random assignment of students to class-rooms but not vice versa. Sojourner suggests an alternative peer effects estimator for his conditions. On the other hand, we explicitly consider measurement error in the background variable. This is not part of Sojourner's analysis.[14]

Table 7 presents the results from OLS regressions similar to the earlier ones in the top panel, and IV results in the lower panel. Both the individual and peers' index of the number of books at home from the home questionnaire are instrumented by the individual and peers' index of books at home from the student questionnaire. We also present estimates for the ratio of peer effect and the individual effect. The table only displays averages over all our six countries.[15]

Instrumenting the individual level index of books at home more than doubles the coefficient in all specifications. This may suggest a large amount of measurement error

[13] This problem has also been recognized by Altonji (1988) although his approach does not solve it completely.

[14] We suspect that our procedure of applying the standard peer effects estimator to the sample of observed students and correcting the estimates as ex-post for missing students as in (11) or (14) should be more efficient than the Sojourner (2008) p-weight estimator under the conditions where our analysis is valid. This is because the p-weight estimator involves a large number of additional covariates which will not affect the residual variance under our scenario. However, we do not have a formal proof for this conjecture.

[15] The averages for the peer and individual effects are obtained as before. The ratio is estimated as the ratio of the country averages (rather than the average of the ratios for each country). This is the efficient estimate under the assumption that the underlying coefficients are the same in each country, and we want to recover this common coefficient. The estimate of the ratio will also generally be biased in small samples (due to sampling error and Jensen's inequality). This bias will be minimized by taking averages first and then forming the ratio.

in the books at home variable. It could also imply that the IV estimate is biased upwards because the regressor and instrument have bounded support.

More interestingly, the coefficient on the peer variable does not increase in the IV specifications when only student and class level variables are included (cols. 1 and 2), and, in fact, it falls slightly. This is consistent with our discussion of equations (8), (9), and (12) above. Measurement error in the peer regressor may actually lead to an upward bias in the OLS specifications if $\sigma_\eta^2 > 0$, as can be seen in equation (8). Moreover, the ratio of the coefficient on the peer effect and the individual effect is around 1.5. This is much too large to be believable and further underscores that these estimates are likely subject to bias from measurement error (and/or correlated effects).[16]

Things are very different when we go to the within school specification in col. (3). The coefficient on the individual level regressor changes little compared to col. (2), while the coefficient on the peer variable falls to a third in the OLS specification. This is consistent with the comparison of equations (8) and (10). The within specification removes the first (positive) bias term in (8), and it exacerbates the standard attenuation bias by removing the potentially important variance component σ_η^2. In the IV results, on the other hand, the coefficient on the peer variable is fairly similar to that in col. (2). A comparison of equations (12) and (13) suggests that the IV coefficient should fall going to the within estimate. However, our result could easily be due to sampling variation. Overall, we conclude that the relative stability of the IV estimates across columns is more consistent with an explanation based on measurement error than one based on correlated effects.

The ratio of the peer effect to the individual effect is now in the range of 0.6 to 0.7. This is more sensible, since we expect the peer effect to be smaller than the individual effect, although it still reflects a large estimate of the peer effect. Moreover, the OLS and IV estimates of the ratio in col. (3) are now fairly similar. This is what we expect from equations (11) and (13). The IV estimate is slightly higher than the OLS one. This is consistent with the idea that our IV estimates are biased up because the

[16] It may seem curious that the standard error for the ratio of the peer and individual effect is smaller for the IV estimates in cols. (1) and (2) than for the corresponding OLS estimates (although the standard errors on the coefficients for the individual and peer books variables go up in the IV estimation compared to OLS). This results from the fact that the coefficient for the individual level effect goes up in the IV results, and this coefficient enters the denominator of the standard error calculation.

regressor and instrument have bounded support. This reasoning would suggest that the OLS estimate of the ratio might be the more reliable one than the IV estimate. Of course, the estimates in col. (3) are still biased because not all students are sampled.

We therefore implement the correction for the sampling bias as suggested in equations (11) and (14) in col. 4.[17] This affects both the peer effects estimate and the estimate of the ratio. The estimates are about 30 percent higher, indicating potentially substantial peer effects. As before, excluding tracked schools in col. (5) makes little difference to the results. Our best estimate for the ratio of the peer and individual effects is therefore around 0.75, which is substantial.

We have tried to argue that the allocation of students into classrooms within schools is approximately random. Nevertheless, it is not possible to rule out some sorting of students in practice. Could our results have been generated simply by sorting of students while true peer effects are zero? It is impossible to rule out this possibility completely. This results from the fact that a general enough model of student achievement has enough free parameters to generate both the test results for random assignment in table 4 and the regression results in tables 5 to 7. In particular, a very small class room level variance component σ_{η}^{2} (relative to the individual level component σ_{v}^{2}), which is highly correlated with the classroom level shock σ_{μ}^{2}, combined with a commensurate individual level effect on the background variable X_{ics} can generate all of our result. Since this combination of parameter values occupies a small region close to (but not on) the boundary of the feasible parameter space, it strikes us as rather unlikely.[18]

7 Effect sizes

Of course, even if we identify a positive peer effect, one might ask whether we care much about the precise magnitude of the coefficient on the peer variable. Books at home is at best a fairly imperfect proxy for the family background of peers. Hence, we

[17] The adjustment for sampling bias is applied to the individual country estimates of the peer effects before taking country averages.

may care more about statistical significance than the actual magnitudes. But this strikes us an overly pessimistic view. We will therefore proceed to use three different methods to assess the economic magnitude of the effects.

It is common in the literature to report effect sizes of the peer effects estimates, so this helps to facilitate comparisons with other studies using different peer measures. Effect sizes are typically calculated as $\sigma_{\overline{X}}\hat{\lambda}/\sigma_y$ where $\sigma_{\overline{X}}$ is the within country variation in the peer variable, and σ_y is the within country variation in the test scores. This quantifies the peer effect as the impact of a one standard deviation change in peer background in terms of individual level standard deviations of the outcome variable.

One complication with this measure in our context is that the standard deviation of the peer variable is not an unbiased measure of $\sigma_{\overline{X}}$ because of the measurement error. However, since we have both the parents' and the children's reports for books at home, the covariance of the two is a measure of the variance of the true variable if both reports are only subject to uncorrelated classical measurement errors. Both our estimate of $\sigma_{\overline{X}}$ and the IV estimate of $\hat{\lambda}$ therefore rely on the classical measurement error model being a good approximation in our case, and the parents' and children's reports being uncorrelated.

We report the effect size measure and the necessary ingredients in Table 8. As before, the effect sizes vary quite widely across countries. The variation in effect sizes comes almost exclusively from variation in the peer coefficients. The average effect size across countries is 0.17. This is larger than most of the estimates in the literature. The bulk of the reported effect sizes is in the range of 0.05 – 0.10. Our estimate is at the upper end of that range but well below the highest estimates reported in studies by Hoxby (2000), Boozer and Cacciola (2001), and McEwan (2003).

Another way to gauge the size of our estimates is to compare them to the effect of a well known alternative intervention. We picked for this comparison the change in class size in the Tennessee STAR experiment, as reported by Krueger (1999). Krueger reports a class size effect of -0.81 per student in third grade (Table VII), which is closest to the age group in our study. This corresponds to an effect size for a change in class

[18] Detailed derivations of these claims and power calculations from a simulation study are available from the authors upon request.

size by one student of about 0.03. A one standard deviation change in peer composition therefore corresponds to a change in class size by about 5 students. This suggests to us that our estimate is fairly large in comparison.

Of course, the size of the peer effect estimate also depends on how well our family background measure actually captures the relevant characteristics of students. It is therefore useful to compare the peer coefficient to the individual level coefficient as we have done already in Table 7. If books at home are a good predictor of reading success then the coefficient on own books at home will be larger and the peer coefficient will also be larger, and vice versa. Column (6) in Table 8 reports the ratio of the two. The average based on the OLS results is 0.77, indicating that the estimate of the peer effect is large compared to the estimate of the individual level effect, since we would expect peers' background to matter much less than own background. One drawback of this comparison is that it depends on what other variables are controlled for in the regression. For example, some studies in the literature control for multiple family background characteristics at the individual level. This makes a comparison across studies very difficult.

One reason why our estimates seem relatively large might be that we are careful about the measurement error in the peer effects variable. However, adjusting for measurement error lowers the estimate of $\sigma_{\overline{X}}$ and raises the estimate of $\hat{\lambda}$, so this cuts two ways. However, the upward adjustment in $\hat{\lambda}$ is much more important. Calculating the effect size on the basis of the estimates ignoring measurement error yields a value of only 0.06, about a third of the size of our IV results. Hence, the treatment of measurement error may be rather important, particularly in studies based on survey data, like Schindler Rangvid (2007) and Schneeweis and Winter-Ebmer (2007). A further explanation for the large effect sizes could be that we estimate the cumulative impact of peers if class room composition is fixed over the previous four years and not the incremental effect of a value-added specification. We should also point out that our confidence intervals are fairly large because the within school and IV estimates are relatively noisy.

8 Conclusion

Peer effects are potentially a major input into the process of educational production but are difficult to estimate empirically. We estimate peer effects across classes within primary schools and argue that classes within schools are formed randomly with respect to family background. We find that a one standard deviation change in our student background measure of peer composition leads to a 0.17 standard deviation change in reading test scores of fourth graders across our sample of six European countries. This is slightly larger than most previous estimates in the literature. The individual country estimates are relatively noisy so that we feel that most is learned from the country averages. For Sweden, the estimated effects are not different from the average for the other countries, although we found some evidence that students may not be randomly allocated to classes in Sweden.

We have argued that there is little evidence for systematic sorting into class rooms within schools for the other countries, and for different classes receiving different observable instructional resources. Hence, comparing students in different classes within schools should be an effective way of dealing with any selection at the school level. Surprisingly, we find that this selection does not seem to be very important once we take measurement error issues into account. We have argued that the within school estimator solves the measurement error problem when we look at the ratio of the peer effect and the individual effect. The OLS and within school results alone are consistent with an explanation based either on selection of students into schools and correlated effects at the school level or measurement error because the estimated peer effects drop substantially when we go from the across school to the within school results.

As an alternative to the OLS results we also present IV estimates. Unlike the OLS estimator, the IV estimator solves the measurement error problem both in the case of the across school and the within school regressions. The IV results are very similar regardless of whether we introduce school fixed effects. This is consistent with a measurement error explanation but not with a role for correlated effects at the school level. The discussion in this literature seems dominated with solving the selection issues, while little attention is being paid to the measurement error and sampling issues, which we find to be important in our data.

References

Altonji, Joseph. 1988. The effect of family background and school characteristics on education and labor market outcomes. Unpublished manuscript, Department of Economics, Northwestern University, Evanston, IL.

Ammermueller, Andreas. 2005. Educational opportunities and the role of institutions. Discussion Paper no. 05-44, ZEW, Mannheim Germany.

Ashenfelter, Orley, and Alan Krueger. 1994. Estimates of the economic return to schooling from a new sample of twins. *American Economic Review* 84: 1157-1173.

Boozer, Michael, and Stephen E. Cacciola. 2001. Inside the 'Black Box' of Project STAR: Estimation of peer effects using experimental data. Discussion Paper no. 832, Economic Growth Center Center, Yale University, New Haven, CT.

Campbell, Jay R., Dana L. Kelly, Ina V.S. Mullis, Michael O. Martin and Marian Sainsbury. 2001. *Framework and specification for PIRLS assessment 2001.* Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41:778-820.

Cullen, Julie Berry, Brian A. Jacob, and Steven Levitt. 2006. The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica* 74:1191-1230

DeGroot, Morris. 1984. *Probability and statistics*. 2nd ed. Reading, MA: Addison Wesley Publishing Company.

Glewwe, Paul. 1997. Estimating the impact of peer group effects on socioeconomic outcomes: Does the distribution of peer group characteristics matter? *Economics of Education Review* 16:39-43.

Gould, Eric D., Victor Lavy, and M. Daniele Paserman. Forthcoming. Does immigration affect the long-term educational outcomes of natives? Quasi-experimental evidence. *Economic Journal*.

Graham, Bryan S. 2008. Identifying social interactions through conditional variance restrictions. *Econometrica* 76:643–660.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. Does peer ability affect student achievement? *Journal of Applied Econometrics* 18:527-544.

Hoxby, Caroline. 2000. Peer effects in the classroom: Learning from gender and race variation. Working Paper no. 7867, National Bureau of Economic Research, Cambridge, MA.

Kane, Thomas J., Cecilia E. Rouse, and Douglas Staiger. 1999. Estimating returns to schooling when schooling is misreported. Working Paper no. 7235, National Bureau of Economic Research, Cambridge, MA.

Krueger, Alan B. 1999. Experimental estimates of education production functions. *Quarterly Journal of Economics* 114:497-532.

Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies* 60:531-542.

McEwan, Patrick J. 2003. Peer effects on student achievement: Evidence from Chile. *Economics of Education Review* 22:131-141.

Mislevy, Robert. 1991. Randomization-based inference about latent variables from complex samples. *Psychometrika* 56:177-196.

Mullis, Ina V.S., Michael O. Martin, Ann M. Kennedy, and Cheryl L. Flaherty. 2002. *PIRLS 2001 Encyclopedia: A reference guide to reading education in the countries participating in IEA's Progress in International Reading Literacy Study (PIRLS)*. Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Mullis, Ina V.S., Michael O. Martin, Eugenio J. Gonzalez and Ann M. Kennedy. 2003. *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools*. Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Schindler Rangvid, Beatrice. 2007. School composition effects in Denmark: quantile regression evidence from PISA 2000. *Empirical Economics* 33:359-388.

Schneeweis, Nicole, and Rudolf Winter-Ebmer. 2007. Peer effects in Austrian schools. *Empirical Economics* 32:387-409.

Sojourner, Aaron. 2008. Inference on peer effects with missing peer data: Evidence from Project STAR. Unpublished manuscript, Department of Economics, Northwestern University, Evanston, IL.

Vigdor, Jacob L., and Thomas Nechyba. 2004. Peer effects in elementary school: Learning from "apparent" random assignment. Unpublished manuscript, Department of Economics, Duke University, Durham, NC.

Vigdor, Jacob L., and Thomas Nechyba. 2006. Peer effects in North Carolina public schools. In *Schools and the equal opportunities problem*, ed. Ludger Woessmann and Paul E. Peterson. Cambridge, MA: MIT Press.

Wößmann, Ludger. 2008. How equal are educational opportunities? Family background and student achievement in Europe and the United States. *Zeitschrift für Betriebswirtschaft* 78:45-70.

Table 1
Mean reading scores and sample sizes

	Germany	France	Iceland	Netherlands	Norway	Sweden
Reading score (all)	539.1 (63.6)	525.2 (66.6)	512.4 (71.0)	554.2 (51.2)	499.2 (77.5)	561.0 (61.5)
Reading score (sample)	548.6 (59.9)	533.7 (64.2)	518.6 (68.4)	565.2 (51.3)	505.0 (76.0)	563.1 (61.3)
Reading score (excl. tracked schools)	548.6 (59.7)	534.2 (65.4)	518.7 (68.3)	562.8 (53.6)	504.4 (76.4)	562.8 (61.5)
Reading scores (tracked schools)	549.4 (64.2)	532.2 (60.5)	506.8 (81.0)	568.8 (47.5)	529.1 (66.5)	570.0 (56.6)
No. of students (all)	7,633	3,538	3,676	4,112	3,459	6,044
No. of students (sample)	4,577	2,312	1,728	1,857	2,548	3,997
No. of students in schools with > 1 class	3,628	1,612	1,301	805	1,748	3,270
No. of schools	183	115	84	105	117	119
No. of schools with > 1 class	114	55	39	29	54	79
No. of classes	301	172	135	141	171	267
No. of classes in schools with > 1 class	232	112	90	65	108	227
Fraction of students in schools that apply tracking	.067	.278	.006	.328	.035	.046

Note.—Scores are weighted by students' sampling probability, standard deviations are in parentheses. The rows starting from "No. of students (sample)" and below refer to the sample used in the estimations. The last row reports the fraction of students in schools in which principals state that classes are formed by ability out of all students in schools for which principals reply to the question.

Table 2
Decomposition of variance in class level means

	Germany	France	Iceland	Netherlands	Norway	Sweden
Index of the number of books at home:						
Mean	3.49	3.32	3.99	3.36	4.03	3.91
Total	.2401	.3138	.1480	.3922	.1542	.2643
Between	.2098	.2726	.1220	.3629	.1297	.2174
Within	.0303	.0412	.0259	.0293	.0245	.0469
Age:						
Total	.0326	.0313	.0065	.0306	.0082	.0111
Between	.0250	.0183	.0050	.0212	.0060	.0060
Within	.0076	.0130	.0015	.0094	.0022	.0051
Female:						
Total	.0145	.0226	.0212	.0156	.0145	.0158
Between	.0085	.0174	.0170	.0139	.0125	.0091
Within	.0061	.0052	.0043	.0017	.0020	.0067
Foreign parent:						
Total	.0459	.0463	.0095	.0488	.0222	.0485
Between	.0404	.0413	.0069	.0451	.0189	.0386
Within	.0054	.0050	.0026	.0036	.0033	.0099
Foreign language at home:						
Total	.0141	.0151	.0088	.0345	.0069	.0230
Between	.0112	.0128	.0058	.0330	.0052	.0167
Within	.0029	.0023	.0030	.0015	.0017	.0064
Reading test scores:						
Total	1144.71	1223.61	751.93	896.62	1075.93	1123.78
Between	978.47	908.63	569.62	799.28	933.10	791.51
Within	166.24	314.97	182.31	97.34	142.83	332.27

Table 3
Results for survey of principals of German primary schools

Question	Responses (%)
Who is responsible for forming class rooms / allocating students to classes within a grade level at your school?	
Principal	86
Other person	42
Does the composition of classes change during the first four years of school?	
No, usually not	75
Yes, class composition is rearranged in certain years	8
Yes, individual students change classes for reasons other than repeating	22
Only under particular circumstances	39
Which are the rules for forming classes / allocating students to classes in your primary school?	
Classes are formed such that similar students are in the same class:	
Students from the same neighbourhood / kindergarten	33
Students with similar abilities	6
Students with similar socio-economic backgrounds	3
Students with similar migration backgrounds / language abilites	3
Classes are formed such that they are well mixed (e.g. by sex, age, abilities etc.)	94
Classes are formed more or less randomly	0
Classes are formed according to other rules / principles	3

Note.—Percentage of principals who chose the respective answer. Multiple answers were possible. Number of observations is 36. The survey was sent by email to 150 schools in the German cities of Bonn, Leipzig and Mannheim.

Table 4
Tests for independence of peer variable and class assignment
and for assignment of class room resources

	Germany	France	Iceland	Netherlands	Norway	Sweden
Index of the number of books at home:						
Pearson χ^2	.2415	.3813	.7964	.7512	.0893	.0364
F-test	.4595	.2552	.0123	.3370	.5675	.0000
Age:						
Pearson χ^2	.0694	.2402	.1452	.0992	.0467	.6247
F-test	.0017	.2672	.0021	.0046	.0000	.9300
Female:						
Pearson χ^2	.1240	.4615	.9608	.6011	.8827	.9657
F-test	.5677	.2838	.0000	.1467	.0036	.1589
Foreign language at home:						
Pearson χ^2	.0495	.6920	.1861	.4217	.4860	.0009
F-test	.0000	.4776	.0001	.0000	.0029	.0000

Note.—The rows labeled "Pearson χ^2" report the *p*-value for Pearson χ^2 tests of independence between the student characteristic and class room assignment within each school using the individual level data. The rows labeled "*F*-test" report *p*-values of Wald tests for the joint significance of classroom resources in within school regressions. See text for details.

Table 5
Regressions for reading test score on peer composition

Country	(1)	(2)	(3)	(4)	(5)
Germany	17.97 (3.04)	17.66 (3.03)	18.83 (3.83)	6.13 (6.24)	7.22 (6.40)
France	22.23 (3.05)	22.84 (2.91)	25.67 (3.94)	22.20 (9.12)	17.80 (12.58)
Iceland	18.08 (5.77)	19.97 (5.04)	22.75 (5.58)	12.14 (11.17)	8.81 (11.18)
The Netherlands	17.58 (4.30)	19.70 (4.37)	22.72 (6.99)	.71 (8.59)	9.56 (9.93)
Norway	15.46 (7.33)	9.84 (7.42)	12.73 (7.85)	-3.20 (8.13)	-2.77 (8.24)
Sweden	18.98 (3.84)	18.04 (4.10)	21.94 (3.75)	11.51 (6.70)	11.08 (7.22)
Average across countries	19.17 (1.61)	19.40 (1.59)	21.65 (1.92)	7.65 (3.22)	7.59 (3.48)
Student level variables	✓	✓	✓	✓	✓
Class level variables		✓	✓	✓	✓
School level variables		✓	✓		
Only schools with > 1 class			✓	✓	✓
School fixed effects				✓	✓
Exclude tracked schools					✓

Note.—Weighted least squares regressions using students' sampling probability as weight. Each entry is the coefficient on the peers' index of books at home from a separate regression. Standard errors in parentheses are robust to clustering at the school level. Student level variables are student's sex and age, parents' origin, language spoken at home, index of number of books at home and number of persons living in household. Class level variables are class size, class size squared, teacher's sex, education, experience and experience squared. School level variables are community size, average daily instruction hours, shortage of staff, teaching material and buildings. Tracked schools are those for which principals state that fourth-grade classes are formed on the basis of ability.

Table 6
Regressions for reading test score on peer composition and interactions with individual family background

	(1)		(2)		(3)	
Country	Peer effect	Inter-action	Peer effect	Inter-action	Peer effect	Inter-action
Germany	17.58 (3.03)	.82 (1.04)	5.92 (6.21)	.55 (1.08)	7.09 (6.37)	0.34 (1.11)
France	20.94 (3.12)	2.46 (1.43)	21.37 (9.01)	1.36 (1.68)	16.01 (11.98)	3.02 (2.24)
Iceland	17.32 (5.82)	1.43 (1.66)	11.67 (11.31)	.76 (1.87)	8.66 (11.41)	.24 (1.83)
The Netherlands	17.74 (4.49)	-.34 (1.36)	-.50 (8.90)	2.53 (1.53)	8.23 (9.79)	3.21 (1.34)
Norway	15.35 (7.58)	.16 (1.76)	-3.71 (8.13)	.94 (1.69)	-3.43 (8.26)	1.18 (1.76)
Sweden	19.91 (3.94)	-1.35 (1.14)	11.85 (6.99)	-.41 (1.19)	11.59 (7.22)	-.60 (1.22)
Average across countries	18.81 (1.64)	.38 (.54)	7.30 (3.26)	.77 (.58)	7.31 (3.46)	.95 (.59)
Student level variables	✓	✓	✓	✓	✓	✓
Only schools with > 1 class			✓	✓	✓	✓
School fixed effects			✓	✓	✓	✓
Exclude tracked schools					✓	✓

Note.—Weighted least squares regressions using students' sampling probability as weight. Coefficients on the peers' index of books at home (columns "Peer effect") and interaction term of peers' index and individual level dummy variable for > 100 books at home (columns "Interaction") are shown in each pair of columns. Standard errors are robust to clustering at the school level. Student level variables are student's sex and age, parents' origin, language spoken at home, index of the number of books at home and number of persons living in household. Tracked schools are those for which principals state that fourth-grade classes are formed on the basis of ability.

Table 7
OLS and IV regressions for reading test score
on books at home and peer composition

	OLS				
Independent Variable	(1)	(2)	(3)	(4)	(5)
Individual level index	13.47	13.60	12.86	---	12.87
of books at home	(.43)	(.51)	(.54)		(.59)
Peer index	19.33	21.38	7.57	9.96	9.61
of books at home	(1.58)	(1.89)	(3.25)	(4.31)	(4.57)
Peer effect/individual effect	1.43	1.57	.59	.77	.75
(ratio of country averages)	(.36)	(.15)	(.24)	(.33)	(.35)
	IV using student's report as instrument				
Individual level index	27.15	28.26	29.08	---	29.17
of books at home	(1.05)	(1.17)	(1.36)		(1.39)
Peer index	16.79	17.14	20.68	26.97	27.79
of books at home	(2.38)	(2.82)	(8.95)	(11.74)	(11.81)
Peer effect/individual effect	.62	.61	.71	.93	.95
(ratio of country averages)	(.10)	(.11)	(.30)	(.39)	(.39)
Student level variables	✓	✓	✓	✓	✓
Class level variables	✓	✓	✓	✓	✓
Only schools with > 1 class		✓	✓	✓	✓
School fixed effects			✓	✓	✓
Corrected for sampling bias				✓	✓
Exclude tracked schools					✓

Note.—Weighted least squares and instrumental variable regressions using students' sampling probability as weight. Averages across six countries are shown. Standard errors are robust to clustering at the school level. In the second panel the individual's and peers' index of the number of books at home from the home questionnaire are instrumented by the individual's and peers' index of books at home from the student questionnaire. A dummy for missing observations for the books variable from the student questionnaire has been added to not further restrict the sample size. Student level variables are student's sex and age, parents' origin, language spoken at home, index of number of books at home and number of persons living in household. Class level variables are class size, class size squared, teacher's sex, education, experience and experience squared. The correction factor for sampling bias in columns (4) and (5) is (N-1)/(n-1). Tracked schools are those for which principals state that fourth-grade classes are formed on the basis of ability.

Table 8
Effect sizes

Country	S. D. test score σ_y (1)	S. D. peer variable $\sigma_{\tilde{\bar{X}}}$ (2)	S. D. peer var. adjusted $\sigma_{\bar{X}}$ (3)	Peer effect $\hat{\lambda}$ (4)	Effect size $\sigma_{\bar{X}}\hat{\lambda}/\sigma_y$ (5)	Peer effect/ Individual effect (6)
Germany	60.00	.459	.362	23.90	.144	.52
France	65.45	.646	.539	60.62	.499	2.11
Iceland	67.58	.348	.251	-2.17	-.008	1.47
Netherlands	56.46	.678	.532	-25.00	-.236	.15
Norway	77.12	.372	.301	3.05	.012	-.24
Sweden	61.50	.461	.400	30.83	.201	1.34
Average across countries	64.68	.494	.397	26.97	.166	.77

Note.—Column (3) is the square root of the covariance between the peer variables index of books at home taken from the student and the home questionnaire. The estimates of the peer effects in column (4) are taken from column (4) in Table 7. The results in column (5) are calculated as (4)*(3)/(1). The results in column (6) are calculated as the ratio of the peer coefficient to the individual coefficient from the OLS regressions in column (4) in Table 7.

Appendix

We are interested in estimating equation (6) in the text

$$y_{ics} = \beta x_{ics} + \lambda w_{cs} + \epsilon_{ics} \tag{1}$$

where $w_{cs} = \overline{x}_{(-i)cs}$ is the peer effect. The background variable x_{ics} is given by

$$\begin{aligned} x_{ics} &= \eta_{cs} + v_{ics} \\ E(v_{ics}) &= 0 \end{aligned}$$

and v_{ics} is iid across observations. Measurement error is classical so that the measured variable is

$$\begin{aligned} \widetilde{x}_{ics} &= x_{ics} + u_{ics} \\ E(u_{ics}) &= 0 \end{aligned}$$

with u_{ics} also iid across observations. Finally, we assume $E\left(x_{ics}\epsilon_{jcs}\right) = E\left(w_{cs}\epsilon_{ics}\right) = 0 \quad \forall i, j$.

The OLS estimator $\widehat{\beta}_{OLS}$ is

$$\widehat{\beta}_{OLS} = \frac{\sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)^2 \sum\left(y - \overline{y}\right)\left(\widetilde{x} - \overline{\widetilde{x}}\right) - \sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)\left(\widetilde{x} - \overline{\widetilde{x}}\right)\sum\left(y - \overline{y}\right)\left(\widetilde{w} - \overline{\widetilde{w}}\right)}{\sum\left(\widetilde{x} - \overline{\widetilde{x}}\right)^2 \sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)^2 - \left[\sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)\left(\widetilde{x} - \overline{\widetilde{x}}\right)\right]^2} \tag{2}$$

and an anologous expression holds for $\widehat{\lambda}_{OLS}$. In order to derive the plims of the estimators, we will need the plims of the sums of squares and cross-products in this expression. There are N_{cs} students in a class. Even though $E(v_{ic}) = 0$, because N_{cs} is relatively small, $\overline{v}_{(-i)cs}$ will generally be different from zero. It is useful to distinguish $\overline{v}_{(-i)cs}$ from η_{cs}, because η_{cs} will carry information about x_{ics}, while $\overline{v}_{(-i)cs}$ will not.

Not all students are observed. Hence the plims of the variance terms are

$$\begin{aligned} \text{plim}\frac{1}{n}\sum_i \left(x_{ics} - \overline{x}\right)^2 &= \sigma_\eta^2 + \sigma_v^2 \\ \text{plim}\frac{1}{n}\sum_i \left(w_{cs} - \overline{w}\right)^2 &= \sigma_\eta^2 + \frac{\sigma_v^2}{\overline{N} - 1} \end{aligned}$$

where the sum is over observed students, n is the total number of students in the sample, and $\overline{N}$ is average class size. Note that while the sum in the plims above is over sampled students, w_{cs} is the peer mean among all students, i.e.

$$w_{cs} = \overline{x}_{(-i)cs} = \frac{1}{N_{cs}} \sum_{j=1, j\neq i}^{N_{cs}} x_{jcs}.$$

In order to interpret the plims it is necessary to consider the process which generates missing students. If students are missing at random, the distribution of v_{ics} among observed and missing students will be the same. As a result, σ_v^2 in the expressions above is the population variance of v_{ics}. However, our derivation holds for weaker conditions than missing at random. Instead suppose that distribution of v_{ics} among missing students is different from that among observed students. Our derivations hold as long as these distributions are independent of class room assignment. In the case where the distribution of missing students is different, the interpretation of σ_v^2 in the expressions above is that of the variance of v_{ics} in the sub-population of observed students. The key to our results is that all the plims of all the variance and covariance terms below will only involve terms σ_v^2 for this particular sub-population. This comes from the fact that all the relevant variance and covariance terms will always involve at least one argument pertaining to observed students. As a result, all the variances in the plims always refer to the observed sub-population.

We have a sample on $n_{cs} \leq N_{cs}$ students in the class. The peer mean in the sample is computed over observed students only. Hence the plims in terms of the observed variables $\widetilde{x}_{ics}$ and $\widetilde{w}_{cs} = \widetilde{\overline{x}}_{(-i)cs}$ in the sample are

$$\begin{aligned}
\text{plim}\frac{1}{n}\sum_i \left(\widetilde{x}_{ics} - \overline{\widetilde{x}}\right)^2 &= \sigma_\eta^2 + \sigma_v^2 + \sigma_u^2 \\
\text{plim}\frac{1}{n}\sum_i \left(\widetilde{w}_{cs} - \overline{\widetilde{w}}\right)^2 &= \sigma_\eta^2 + \frac{\sigma_v^2}{\overline{n}-1} + \frac{\sigma_u^2}{\overline{n}-1}.
\end{aligned}$$

We will also need various covariance terms below. These are

$$\begin{aligned}
\text{plim}\frac{1}{n}\sum_i \left(\widetilde{x}_{ics} - \overline{\widetilde{x}}\right)\left(\widetilde{w}_{cs} - \overline{\widetilde{w}}\right) &= \sigma_\eta^2 \\
\text{plim}\frac{1}{n}\sum_i \left(\widetilde{x}_{ics} - \overline{\widetilde{x}}\right)\left(w_{cs} - \overline{w}\right) &= \sigma_\eta^2 \\
\text{plim}\frac{1}{n}\sum_i \left(x_{ics} - \overline{x}\right)\left(\widetilde{w}_{cs} - \overline{\widetilde{w}}\right) &= \sigma_\eta^2 \\
\text{plim}\frac{1}{n}\sum_i \left(\widetilde{w}_{ics} - \overline{\widetilde{w}}\right)\left(w_{cs} - \overline{w}\right) &= \sigma_\eta^2 + \frac{\sigma_v^2}{\left(\overline{N}-1\right)}.
\end{aligned}$$

Substituting (1) into (2), taking the plim, and rearranging yields

$$\begin{aligned}
\widehat{\beta}_{OLS} = \; & \frac{\frac{1}{n}\sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)^2 \frac{1}{n}\sum\left(\beta\left(x - \overline{x}\right) + \lambda\left(w - \overline{w}\right) + \left(\epsilon - \overline{\epsilon}\right)\right)\left(\widetilde{x} - \overline{\widetilde{x}}\right)}{\frac{1}{n}\sum\left(\widetilde{x} - \overline{\widetilde{x}}\right)^2 \frac{1}{n}\sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)^2 - \left[\frac{1}{n}\sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)\left(\widetilde{x} - \overline{\widetilde{x}}\right)\right]^2} \\
& - \frac{\frac{1}{n}\sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)\left(\widetilde{x} - \overline{\widetilde{x}}\right) \frac{1}{n}\sum\left(\beta\left(x - \overline{x}\right) + \lambda\left(w - \overline{w}\right) + \left(\epsilon - \overline{\epsilon}\right)\right)\left(\widetilde{w} - \overline{\widetilde{w}}\right)}{\frac{1}{n}\sum\left(\widetilde{x} - \overline{\widetilde{x}}\right)^2 \frac{1}{n}\sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)^2 - \left[\frac{1}{n}\sum\left(\widetilde{w} - \overline{\widetilde{w}}\right)\left(\widetilde{x} - \overline{\widetilde{x}}\right)\right]^2}
\end{aligned}$$

$$
\begin{aligned}
\operatorname{plim}\widehat{\beta}_{OLS} &= \frac{\left(\sigma_\eta^2+\frac{\sigma_v^2}{\overline{n}-1}+\frac{\sigma_u^2}{\overline{n}-1}\right)\left[\beta\sigma_x^2+\lambda\sigma_\eta^2\right]-\sigma_\eta^2\left[\beta\sigma_\eta^2+\lambda\left(\sigma_\eta^2+\frac{\sigma_v^2}{\overline{N}-1}\right)\right]}{\left(\sigma_\eta^2+\frac{\sigma_v^2}{\overline{n}-1}+\frac{\sigma_u^2}{\overline{n}-1}\right)\left(\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right)-\sigma_\eta^4} \\
&= \beta\frac{\sigma_v^2\left(\sigma_v^2+\sigma_u^2\right)+\sigma_\eta^2\left(\overline{n}\sigma_v^2+\sigma_u^2\right)}{\sigma_v^2\left(\sigma_v^2+\sigma_u^2\right)+\sigma_\eta^2\left(\overline{n}\sigma_v^2+\sigma_u^2\right)+\sigma_u^2\left(\left(\overline{n}-1\right)\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right)} \\
&\quad +\lambda\frac{\sigma_\eta^2\left(\frac{\overline{N}-\overline{n}}{\overline{N}-1}\sigma_v^2+\sigma_u^2\right)}{\left(\sigma_v^2+\sigma_u^2\right)\left[\overline{n}\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right]}.
\end{aligned}
$$

In order to study the within class estimator $\widehat{\beta}_W$ consider the deviations from class means

$$
\begin{aligned}
x_{ics}-\overline{x}_{cs} &= v_{ics}-\overline{v}_{cs} \\
\widetilde{x}_{ics}-\overline{\widetilde{x}}_{cs} &= v_{ics}-\overline{v}_{cs}+u_{ics}-\overline{u}_{cs}
\end{aligned}
$$

with analogous transformations for w_{ics} and $\widetilde{w}_{ics}$ and for eq. (1). plims are now taken with $C\rightarrow\infty$, where C is the number of classrooms in the sample, with n_{cs} and N_{cs} fixed. The plims of the sample variance and covariance terms will be as above with two changes. First, the within transformation eliminates η_{cs}, hence the plims for the within variables will correspond to the case with $\sigma_\eta^2=0$. Secondly, the within variance and covariance terms have a small sample bias of $(\overline{n}-1)/\overline{n}$ because classes are small and class sizes fixed. However, in considering plim $\widehat{\beta}_W$ this bias affects the numerator and demnominator proportionately, so that we can obtain plim $\widehat{\beta}_W$ simply from plim $\widehat{\beta}_{OLS}$ by setting $\sigma_\eta^2=0$:

$$
\operatorname{plim}\widehat{\beta}_W=\beta\frac{\sigma_v^2}{\sigma_v^2+\sigma_u^2}.
$$

This is the standard attenuation bias from measurement error.

By a similar argument we obtain for $\widehat{\lambda}$:

$$
\begin{aligned}
\operatorname{plim}\widehat{\lambda}_{OLS} &= \beta\frac{\left(\overline{n}-1\right)\sigma_u^2\sigma_\eta^2}{\left(\sigma_v^2+\sigma_u^2\right)\left[\overline{n}\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right]}+\lambda\frac{\frac{\overline{n}-1}{\overline{N}-1}\sigma_v^2\left(\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right)+\left(\overline{n}-1\right)\sigma_\eta^2\left(\sigma_v^2+\sigma_u^2\right)}{\left(\sigma_v^2+\sigma_u^2\right)\left[\overline{n}\sigma_\eta^2+\sigma_v^2+\sigma_u^2\right]} \\
\operatorname{plim}\widehat{\lambda}_W &= \lambda\left(\frac{\overline{n}-1}{\overline{N}-1}\right)\frac{\sigma_v^2}{\sigma_v^2+\sigma_u^2}.
\end{aligned}
$$

We now turn to the instrumental variables estimator. The instruments

$$
\begin{aligned}
z_{1ics} &= x_{ics}+u_{1ics} \\
z_{2cs} &= \overline{z}_{1cs}=\eta_{cs}+\overline{v}_{cs}+\overline{u}_{1cs}
\end{aligned}
$$

are based on an independent measurement of x_{ics}, i.e. we assume $cov(u_{ics},u_{1ics})=0$.

Similar derivations as before imply

$$\begin{aligned}
\text{plim}\widehat{\beta}_{IV} &= \beta + \lambda \left(\frac{\overline{N} - \overline{n}}{\overline{N} - 1} \right) \frac{\sigma_{\eta}^{2}}{\overline{n}\sigma_{\eta}^{2} + \sigma_{v}^{2}} \\
\text{plim}\widehat{\beta}_{IV,W} &= \beta \\
\text{plim } \widehat{\lambda}_{IV} &= \lambda \left(\frac{\overline{n} - 1}{\overline{N} - 1} \right) \frac{\overline{N}\sigma_{\eta}^{2} + \sigma_{v}^{2}}{\overline{n}\sigma_{\eta}^{2} + \sigma_{v}^{2}} \\
\text{plim } \widehat{\lambda}_{IV,W} &= \lambda \left(\frac{\overline{n} - 1}{\overline{N} - 1} \right).
\end{aligned}$$