The causal effect of class size on scholastic achievement: distinguishing the pure class size effect from the effect of changes in class composition*

SIMONE DOBBELSTEEN, JESSE LEVIN and HESSEL OOSTERBEEK

Department of Economics, University of Amsterdam

I. Introduction

As in many other countries the causal effect of class size on pupils' school performance is an important issue in the ongoing educational debate in the Netherlands. In 1997 the Dutch government committed to undertake a major investment to reduce average class size in primary schools. Although opinion leaders and politicians were easily persuaded that this is a good investment, this measure is not backed by convincing empirical evidence pertaining to the Dutch situation. In the process of policy preparation one study was published supporting the reduction of class size, but as we will show in this paper the results of this study should be questioned.

Traditionally, the study of the effect of class size on school performance has been dominated by educational researchers. Only recently economists entered this field. A notable early exception is the influential survey article by Hanushek (1986) who, after reviewing a large number of studies, concludes that there is no positive effect of schooling expenditures on pupils' school performance.

The key methodological problem to assess the causal effect of class size on pupils' school performance, is that assignment of pupils to classes of different sizes need not be random. Consequently, differences in performance between

*We gratefully acknowledge valuable comments from Joshua Angrist, Roel Bosker, Alan Krueger, Henry Levin, Wim Meijnen, Erik Plug, Dinand Webbink, an anonymous referee, and seminar participants in Amsterdam, Dublin and Jerusalem. We thank the Dutch organization for Scientific Research (NWO) for giving permission to use the PRIMA data, and Jaap Roeleveld for actually supplying the data. We also thank the NWO Priority Program "Scholar" for financial support. The authors take full responsibility for any remaining errors.
pupils from small and large classes can be caused by factors other than differences in class size. Especially, two mechanisms are likely to play a role. Firstly, parents may choose the school to which they send their children based partly on class size. It is likely that parents who care more about their children's achievement in school are more inclined to base their choices on class size. If that is the case and the degree to which parents care about their children's achievement has an own independent effect on children's school performance but is unobserved, then the class size effect will be biased downwards. Secondly, if there is more than one class at a particular grade level, the size and composition of each class depend on choices made by the school. The school might decide to form a larger and a smaller class at the same grade level and assign weaker pupils to the smaller class. If controlling for this assignment process is not possible, the class size effect will be biased upwards. Hence, an uncorrected estimate of the class size effect can be biased upwards or downwards depending on which of the two sources of bias plays the larger role in a particular setting.

The most attractive way to encompass these problems is by running a real field experiment. This has been done in the state of Tennessee, where the Student/Teacher Achievement Ratio (STAR) experiment was held. Here, students were randomly assigned to classes of different sizes. But even in that case, serious problems such as re-assignment and attrition, which possibly biased the results, occurred. In a careful re-analysis of the STAR experiment, Krueger (1999), concludes that in this experiment class size reductions boost students' performance. In a follow-up analysis, Krueger and Whitmore (2001) analyse the long-term effects of the intervention and find that these are present. Of course, organising a field experiment is generally not within the reach of researchers. Therefore, to collect information applicable to other settings than those in Tennessee, different approaches have to be applied. In a recent paper Angrist and Lavy (1999) propose a method which is closely related to the real experiment type of study. They use an instrumental variable approach, where the instrument is based on a rule currently applied in Israel which establishes a discontinuous relation between enrolment and class size. This rule produces exogenous variation in class size. Isolating the exogenous portion of class size and estimating its effect on achievement can be considered analogous to running a field experiment.

In the current paper we apply a methodology close to the one applied by Angrist and Lavy to the Dutch situation. The instrument used here is based on the rules which link total enrolment in the school to the number of teachers a school receives payment for. These rules too cause discontinuities in the relation between enrolment and average class size and pupil-teacher ratio. These discontinuities are clearly smaller than those in Israel, but still lead to a considerable degree of exogenous variation of class size. The dataset used
The causal effect of class size on scholastic achievement

here contains individual information on more than 12,000 pupils over three different grade levels (4, 6, and 8).

Corrected for a large number of observables and for endogeneity bias, our estimation results indicate that pupils in smaller classes do not perform better (and sometimes even worse) than pupils in larger classes. In a subsequent analysis we attempt to explain this finding, by decomposing the class size effect into two underlying effects. Reduction of class size causes two changes. One change is that the number of pupils per teacher decreases. This allows the teacher to spend more teaching time per pupil which is likely to improve the pupils' achievement. A second change caused by reducing class size is that the number of classmates one has is reduced. If pupils do not only learn from their teacher but also from their classmates, this change may harm the pupils' achievement. Social cognitive learning theory strongly suggests that pupils' achievement may benefit from a larger number of classmates with similar levels of cognitive ability. To disentangle the two effects, we construct a new variable which measures for each pupil in the sample the number of classmates with similar IQ. Inclusion of this variable in the analysis reveals that the number of similar pupils has indeed a significantly positive impact on performance. Moreover the class size effect has now in most cases the proper negative sign. It is important to stress that our measure of similar pupils counts the absolute number of similar pupils within a class. This variable has an expected value that is increasing in class size, and is different from usual measures of class homogeneity.

The organization of the remainder of this paper is the following. In the next section we summarize in some more detail the methods and results of recent related studies. We focus on two studies in particular. The first is the aforementioned study by Angrist and Lavy (1999) and the second by Bosker and Hox (1996), the latter being a Dutch study that played an important role in persuading Dutch politicians that reducing class size is an effective means to boost scholastic achievement. Their study uses the same dataset as we will use for our own analyses. In Section III we describe the dataset and the construction of variables; the section also contains an account of the creation of our instrumental variable. Section IV describes, discusses and explains the empirical findings. Section V is devoted to the robustness of the findings and discusses the outcomes of different specifications and different sub-samples. Section VI summarizes and concludes.

II. Related Studies

Class sizes in public schools in Israel are partly determined by a rule proposed by the 12th century scholar Maimonides. According to this rule up to 40 pupils
may be put in one class; if there are more than 40 pupils an additional class will be created. Consequently, if there are 40 pupils at a grade level, the average class size equals 40. With one additional pupil the average class size drops to 20.5. When enrolment rises further average class size increases again to 40 when enrolment equals 80. The 81st pupil causes the creation of a third class, so that average class size drops again to 27.

Applying Maimonides' rule leads to a discontinuous relation between enrolment and class size. In practice, this rule is not strictly followed; actual class sizes differ from class sizes predicted by Maimonides' rule. But as long as actual class size is partly determined by this rule, the rule serves as a credible source of exogenous variation of actual class size. As such it can be used to create an instrumental variable which purges actual class size of its endogenous component. Class size according to the rule is a valid instrument if there is no effect of the instrument on pupils' school performance, other than the one caused by the relation between the instrument and actual class size.

Angrist and Lavy (1999) regress class average achievement scores on school characteristics and class size. This equation is estimated separately for 3rd, 4th and 5th grade classes in Israeli primary schools using both OLS and IV. The OLS results show a significantly positive effect of class size on reading and math scores when no control for socio-economic status (SES) is included. The class size effects get much smaller, although some remain significantly positive when SES is controlled for. For 5th and 4th grades, results change dramatically with the IV approach. In almost every specification the effect of class size on reading and math scores becomes significantly negative. This is not the case for 3rd grade pupils. As a possible explanation for this latter effect Angrist and Lavy mention that the effect of small classes operates cumulatively. It is likely that pupils who are in small classes in 5th grade were also in small classes in 4th and 3rd grades. This explanation is supported by the finding that class size effects are larger for 5th graders than for 4th graders.

Some other recent studies also apply the IV approach to the effect of class size on pupils' school performance. Akerhielm (1995) uses the average class size for a given subject in the student's school and the 8th grade student enrolment in the school as instruments for class size. She reports significantly negative effects of class size on school performance. Boozer and Rouse (1995) construct instruments from the maximum special education class size as dictated by state law. They too find significantly negative effects of class size on school performance. Hoxby (2000) uses two independent identification strategies. The first exploits the random component of population variation, which causes exogenous variation in class size. Her second method is very close to the one developed by Angrist and Lavy and is based on the discontinuous changes in class size occurring when a small change in
enrolment triggers a maximum or minimum class size rule. For both identification strategies Hoxby finds no effect of class size on achievement. The estimates are sufficiently precise to even rule out modest effects.

While the results of Akerhielm, and Boozer and Rouse are consistent with Krueger's (1999) findings from the field experiment organized in Tennessee, Hoxby's results are at odds with these findings. A possible explanation, suggested by Hoxby, is that in a field experiment participants are aware of being evaluated which may provide them with an incentive to make good use of the opportunities. Such explicit incentives are absent in the natural experiment approach employed in Hoxby's study. This leaves unexplained, however, why the results obtained by Hoxby deviate from those reported in the other two instrumental variable studies. Although in this respect it should be pointed out that the exogenous variation exploited by Hoxby is more credible than the exogenous variation used in the other two studies.

In February 1996 the Dutch vice-minister of Education asked an independent committee to advise her on the implication of smaller class sizes in primary education. The report includes an empirical study by Bosker and Hox (1996), using the so-called PRIMA-data. Our own empirical analysis is based on the same dataset, which is be described in greater detail in the next section.

The PRIMA survey contains a large amount of information on pupils in grades 2, 4, 6 and 8. For all pupils individual school performance is measured in the fields of arithmetic and language. These individual scores serve as the dependent variables and Bosker and Hox regress these, in a multi-level framework, on a series of dummy variables measuring different class size brackets. The brackets are for the following class sizes: 5–9, 10–14, 15–19, 20–24, 25–29, 30–34 and 35–39. Although this specification captures some degree of possible non-linearity in the relation between class size and performance, the specification also ignores any variation within these brackets. The authors admit that their results may be biased because of endogeneity of class size, but no attempt is made to correct for this other than including a number of covariates. These include controls on: pupil's gender and social background; teacher's experience, gender and attitudes; and class averages of pupils' IQ, gender and social background.

The findings of Bosker and Hox provide a mixed picture. For grade 2 performance in arithmetic goes down when class size increases from 20–24 to 25–29; also, in classes in the size brackets 30–34 and 35–39 arithmetic scores are lower than in the reference category. It is important to note however, that increasing class size from 25–29 to 30–34 pupils performance actually increases and is higher than in both the 25–29 and 15–19 categories. For the effect of class size on language performance in grade 2 a similar pattern emerges. Scores are significantly lower in the 25–29 and 35–39 size brackets,
but not in the 30–34 category. Pupils in classes with 30–34 children perform as well on languages as pupils in classes with 5–9, 10–14, 15–19 or 20–24 children. This result is especially important as 21% of the second grade classes in the sample fall in the 30–34 range, and another 27% in the 25–29 range.

In grades 4, 6 and 8 pupils in classes with more than 35 children in it (about 5% of the classes in the sample) do worse for both language and arithmetic. Adding 10 or more pupils to a class of 25–29 students causes a drop in average achievement by about 4 percentiles. Noticeable also is that for arithmetic pupils in the class size bracket 30–34 have the second highest scores.

These results do not support the hypothesis that, in general, smaller class sizes lead to better performance in the fields of arithmetic and language. Only for classes with over 35 pupils a negative effect of class size is found, but for classes with less than 35 children in it the pattern is erratic. The main shortcoming of the study is, however, that the results are not purged for possible endogeneity bias and, as the results in the study by Angrist and Lavy show, this may change the results dramatically. The current paper assesses whether this is indeed the case.

III. Data and Construction of Variables

This paper uses data from the first wave of the PRIMA survey which contains information on Dutch pupils who were enrolled in grades 2, 4, 6 and 8 in the school-year 1994/95. Several instruments have been used for collection of the data: administrative sources, tests, and questionnaires for teachers, parents and school headmasters. 800 primary schools participated in the survey, 400 of which form a national representative sample of regular schools. This paper only uses data on the representative sample of 400 schools. Moreover, we will concentrate only on grades 4, 6 and 8, as we find the information on class size for grade 2 too unreliable.

The PRIMA survey provides information at the pupil, class, and school levels. We use pupils' individual scores on arithmetic and language tests transformed into percentile rankings to measure school performance, and regress these scores on various pupil, class and school characteristics. At the
individual level we include gender and weight factors accounting for SES as covariates. The weight factor of a pupil indicates his/her social background, and is used to determine the amount of resources a school is given for a certain pupil; schools get more money for pupils with a disadvantaged background than for pupils with a “normal” background. The weight factor ranges from 1 to 1.9 and is defined as follows: 1.9-pupils are pupils with foreign-born parents and whose father or mother has at most completed lower vocational education or whose primary earnings parent has a job involving physical labor or has no income from labor; 1.7-pupils are pupils whose parents are transients. 1.4-pupils live in a boarding school or a foster home, and their parents are masters of a ship. 1.25-pupils are pupils of whom one parent has at least an education at VBO-level of education. All other pupils have a weight equal to one.

At the class level we control for the following: class size, teacher’s gender and experience, class averages of pupil’s gender and weight factor, dual teacher class (whether more than one teacher has taught the class), and multi-grade class (whether the class combines individuals from more than one grade level). School level variables include three dummy indicators denoting the denomination of the school, the average SES within a school, and total enrolment of the school.

The dataset also contains a measure of pupils’ IQ. This IQ-score is based on tests taken at the same time as the tests that measure language and arithmetic scores. Consequently, this IQ measure can not be regarded as an independent measure of innate ability and is therefore not included as a control variable in the regressions. In our analysis we use, however, the variable “number of classmates with similar IQ” which is constructed on the basis of pupils’ IQ scores. This variable will be used to identify a peer group effect motivated by social cognitive learning theory (see the next section).

As we want to investigate if pupils in smaller classes perform better than pupils in larger classes, our most crucial explanatory variable is class size. Because of the possible endogeneity of class size, OLS-estimates of the class-size effect on pupils’ test scores may be biased. We therefore use an instrumental variable approach, using an instrument for class size based on the Dutch rules that link total enrolment in the school to the number of teachers a school receives funding for. These rules are decreed by the Dutch Ministry of Education, Culture and Science in the “Formatiebesluit WBO”. In order to determine the number of teachers a school gets paid for, first the weighted number of students (WNS) of a school is calculated using the following formula:

$$WNS = \frac{\text{Total Enrollment}}{\text{Average SES}}$$

Note that Angrist and Lavy’s instrument is based on enrolment at the grade level, whereas our instrument is based on total enrolment at the school level.
where $N$ is total enrolment in the beginning of a school year and $w_i$ is the weight factor of pupil $i$ as described above. The weighted sum of all students is reduced by 9% of the total number of students (as it is assumed that the school should be able to cope with this proportion of disadvantaged students without additional resources), and rounded down to the nearest integer. If this number is lower than the actual (unweighted) number of students then the actual number of students $N$ is used instead. The outcome is increased by 3% to account for the possible increase in total enrolment during the school year due to new entrants. The number of teachers that can be appointed given the weighted number of students can then be looked up in tables. These show that schools with a number of weighted students up to 30 can appoint 2 full-time teachers, and the number of teachers increases by 0.2 or 0.3 full-time units for every 6 to 11 additional weighted students. Figure 1 shows that the resulting pupil-teacher ratio function is discontinuous; the pupil-teacher ratio increases to 15 when the weighted number of students goes up to 30, it then drops to about 14, climbs up again to more than 16, drops to 15.4, etc.

When comparing the graph of our instrument with the graph of Angrist and Lavy's instrument, two differences are apparent. Firstly, the differences at the points of discontinuity are much larger in their case than in ours. The distance from a peak to a trough represents a change in average class size by 1 or 2 with our instrument, while using Maimonides' nile it changes abruptly from 40 to 20.5. While it seems to be more attractive to have large exogenous variations to identify the effect of class size reduction, it is
important to stress that the current Dutch policy proposes long run class size reductions of 3 to 4 pupils. In that case small reductions of 1 or 2 is more on the mark than reductions exceeding 15 pupils. Secondly, the number of discontinuities is much smaller in the Israeli case than in the Dutch case. With a small number of discontinuities only the observations near one of these (discontinuous) points are relevant for the identification of the class size effect. Consequently, Angrist and Lavy present separate results for those observations that are in a \(-5/\pm 5\) range of the points of discontinuity (40, 80, 120 pupils per grade level). With so many evenly spread points of discontinuity, as implied by the Dutch funding scheme, all observations are in a relevant range of potential treatment. Hence, we do not have to narrow our sample.

For each school we know the number of pupils with various weight factors, and so we can derive the number of group teachers that can be appointed by the school based on the ministerial rules. The instrument we use in our estimations is predicted average class size, calculated as total school enrolment divided by the predicted number of group teachers using the ministerial rules. Henceforth, we refer to this instrument as the "Enrolment Function Determined" class size, or EFD class size for short. A nice feature of this instrument is that it is a discontinuous function of the school's average SES-weight factor and total enrolment. This allows us to include both the school's average SES-weight factor and total enrolment as regressors in the second stage equation. Consequently, it is unlikely that the instrument is picking up any direct effects of these two variables. Hence for our instrument to be valid it is only necessary that the particular transformation of average SES-weight and total enrolment affects achievement only through its effect on class size, controlling for linear average SES-weight and total enrolment.\(^4\)

**IV. Empirical Analysis**

For the analysis of the effect of class size on pupils' school performance in Dutch primary schools, we estimate equations of the form

\[
y_{ics} = X_i'x + Z_c'z + W_{s}^t'\nu + n_{ics}\mu + \epsilon_{ics},
\]

\[
n_{ics} = X_i'\beta + Z_c'\gamma + W_{s}^t\delta + EFD_{s}v + \rho_{ics}
\]

where \(y_{ics}\) is the percentile score (in arithmetic or language) of pupil \(i\) in class \(c\) of school \(s\); \(X_i\), \(Z_c\) and \(W_{s}\) are vectors of variables pertaining to the individual, the class and the school, respectively; \(n_{ics}\) is the number of pupils

\(^4\)In Section IV we report results which show that including higher order terms of total enrolment and school average SES-weight do not affect our findings.
in pupil i's class c within school s; EFD_s is the enrolment determined average class size of school s; and $\beta_{ics}$ and $\epsilon_{ics}$ are random components of the error term specific to the pupil. As pupils within the same school may have interdependent error terms, we calculated robust standard errors which allow for correlated error terms of pupils within the same school. The difference between the equations estimated in this paper and the ones estimated by Angrist and Lavy is that the equations in this paper take individual pupils as the unit of analysis while in the Angrist and Lavy paper the unit of observation is the class.\(^5\)

Our main findings are comprised only from the results using actual class size as reported by the teacher. In the next section we explore whether there are differences when alternative measures such as pupil teacher ratio and average class size (calculated as administrative reported enrolment divided by number of group teachers) are implemented. In addition, results for other specifications and for different sub-samples are reported. We performed all analyses separately for grade 4, grade 6, and grade 8. The numbers of observations are large enough to allow that, and we do not want to impose any a priori restrictions on the effects of the regressors on school performance at different grade levels.\(^6\)

The three parts of Table 1 contain our main results. For each grade level five columns with results are reported. The first columns report the first stage estimation results. The second and fourth columns give the respective OLS (baseline) estimates for arithmetic and language achievement while columns three and five contain the corresponding IV results. We do not report the full estimation results, but report only the coefficients that are of prime interest from the perspective of the causal effect of class size on achievement. Hence, we do not report the results for all of the covariates, but only those for total enrolment and the school average SES-weight. These are reported because our instrumental variable is a direct transformation of the two. Included in each cell of the tables is the estimated coefficient with the robust standard errors corrected for grouped (at the school level) heteroskedasticity in brackets.

Starting with grade 4 pupils, the reduced form results in the first column show that EFD-class size is highly correlated with the potentially endogenous actual class size at grade 4. The partial correlation coefficient equals 0.38. The OLS results in columns 2 and 4 show an insignificant positive effect of class size on pupils' arithmetic percentile scores, and an insignificant negative effect on language percentile scores. These are results from an equation including a

---

\(^5\)We checked for our data whether raising the unit of analysis leads to different conclusions; this is not the case.

\(^6\)Tests of differences in class size effects on achievement between grades reveal significant differences using both OLS and IV for language and using IV for arithmetic.
### TABLE 1

<table>
<thead>
<tr>
<th></th>
<th>4th grade</th>
<th></th>
<th>6th grade</th>
<th></th>
<th>8th grade</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arithmetic</td>
<td>Language</td>
<td>Arithmetic</td>
<td>Language</td>
<td>Arithmetic</td>
<td>Language</td>
</tr>
<tr>
<td></td>
<td>First stage</td>
<td>OLS</td>
<td>IV</td>
<td>OLS</td>
<td>IV</td>
<td>First stage</td>
</tr>
<tr>
<td>EFD</td>
<td>1.313</td>
<td>(0.182)**</td>
<td></td>
<td>1.384</td>
<td>(0.189)**</td>
<td></td>
</tr>
<tr>
<td>Class size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.149)</td>
<td>(0.368)</td>
<td>(0.123)</td>
<td>(0.373)</td>
<td>(0.135)</td>
<td>(0.268)</td>
<td>(0.117)*</td>
</tr>
<tr>
<td>Total enrolment</td>
<td>0.007</td>
<td>0.019</td>
<td>0.015</td>
<td>0.011</td>
<td>0.007</td>
<td>0.004</td>
</tr>
<tr>
<td>(0.005)</td>
<td>(0.012)</td>
<td>(0.014)</td>
<td>(0.011)</td>
<td>(0.014)</td>
<td>(0.005)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>SES-weight</td>
<td>(1.609)</td>
<td>(3.219)</td>
<td>(3.371)</td>
<td>(3.320)**</td>
<td>(3.511)*</td>
<td>(1.404)*</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.3314</td>
<td>0.0875</td>
<td>0.0861</td>
<td>0.1461</td>
<td>0.1438</td>
<td>0.4313</td>
</tr>
<tr>
<td>Partial correlation coefficient</td>
<td>0.3777***</td>
<td></td>
<td></td>
<td>0.4725***</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable in first stage is actual class size as reported by the teacher, dependent variables in second stage equations are pupils’ percentile scores for arithmetic and language. All regressions also include a constant, four individual SES-class dummies, a dummy for pupil’s gender, percentage of girls in pupil’s class, teacher’s gender and experience, dummy for dual teacher class, dummy for multi-grade class and three dummy variables for school’s denomination. Robust standard errors taking account of correlated disturbance terms within schools are reported in brackets. ***/**/* indicate significance at the 1%/5%/10%-levels.
large number of control variables. Without control variables, the class size effects in both equations are significantly positive. Although the IV point estimates in columns 3 and 5 differ largely from the OLS results, the differences are not significant as indicated by Hausman tests which accept the null hypothesis that any endogeneity of class size has a negligible effect on scholastic achievement. For both fields of study (arithmetic and language) the IV procedure yields increased but nevertheless insignificant point estimates of the class size effect. Hence, for pupils in grade 4 the estimation results which correct for endogeneity bias give no support for the view that pupils benefit from being placed into smaller classes.

The middle part of Table 1 reports results for 6th graders. EFD-class size is again highly correlated with actual class size (with a partial correlation now equal to 0.47). The OLS and IV results for percentile scores in arithmetic are similar to those obtained for 4th graders. The estimates of the class size effect are not significantly different from zero, and going from OLS result to IV result turns the point estimate from negative into positive. For languages, the results are somewhat different; the OLS-estimate is significantly negative (at the 10% level) indicating that a reduction of class size is beneficial for language achievement. The IV-estimate suggests, however, that this inference might be attributed to endogeneity bias.

Lastly, results for 8th grade pupils are reported in the right hand part of Table 1. Again EFD-class size is highly correlated with the potentially endogenous variable actual class size; the partial correlation equals 0.43. Notice that for 8th graders the first stage equation includes also significantly positive coefficients for total enrolment and school average SES-weight, while in the first stage equations for 4th and 6th graders this was not the case. The significance of total enrolment and school average SES-weight do not, however, result in a weaker instrumental variable. The OLS and IV results for 8th graders differ from those obtained for grades 4 and 6. The OLS estimates of the class size effect do not differ significantly from zero, but the IV estimates do. For both arithmetic and language, we find that 8th graders benefit from being placed in larger classes. For both IV estimations, the Hausman-test strongly rejects the null hypothesis that any possible endogeneity in the class size variable has a negligible effect on scholastic achievement. Thus, using an IV procedure is warranted.

The fact that endogeneity is only a problem for 8th grade and not for 4th and 6th grades suggests that the richness of the dataset already solves a lot of the potential endogeneity problems. As a result the OLS estimates for these grade levels are not significantly different from the IV estimates. This in turn supports the approach followed by Dearden et al. (2002) who identify the causal effect of school quality on educational attainment and wages by using a very extensive and rich set of control variables.
The results reported in Table 1 for different grade levels reveal a remarkably consistent pattern. The point estimates of the class size effect are never significantly negative and in some cases they are even significantly positive. Krueger (1999) reports an improvement of 5–7 percentile point from being placed in a small (13–17) instead of a large (22–25) class. Thus, the class size effect equals a 0.7 percentile point gain per one pupil reduction. When we take this class size effect – which is arguably the most convincing result in the class size literature – as the null-hypothesis, we reject it in 11 out of 12 cases.

Our findings indicate that Dutch primary school pupils do not benefit from being placed in smaller classes. As the results reported in Section V below show, the findings are very robust. Including higher order terms of total enrolment and school average SES-weight, using different measures of class size, allowing for nonlinearities or limiting the sample to specific subgroups does not affect the main results.

The findings reported here not only contradict the “common wisdom” dominant among educationalists but are also at odds with the results from other recent studies using the IV method (but are in line with Hoxby’s results). Although the results should be taken seriously as they stand, the findings will gain importance when their underlying mechanism can be unveiled. It is to this issue that we now turn.

When pupils in larger classes learn no less than pupils in smaller classes, an explanation might have something to do with children learning from one another. More specifically, when there is a larger number of pupils in a class there are potentially more children from whose presence a pupil may benefit. Social cognitive learning theory strongly suggests that the achievement of pupils may benefit from the presence of similar classmates. This theory emphasises that people acquire knowledge, skills and attitudes by observing others (Bandura, 1986; Schunk, 1991). Moreover, educational practices like the use of peer models and ability grouping of pupils are based on the idea that pupils learn by observing and working together with other similar individuals. In this literature, evidence can be found of positive effects of both ability grouping (Slavin, 1987) and peer modelling (Schunk, 1987) on the school performance of pupils.

Research on peer modelling suggests that both the competencies and number of role models can affect pupils’ performances. Similarity in competence (which we proxy with IQ) appears important for purposes of self-evaluation: France-Kaatrude and Smith (1985) show that children that were allowed to compare their performances with a similarly performing peer compared themselves more often, and demonstrated greater task persistence than children that were offered comparisons with superior or inferior peers. Observing the success of individuals similar to oneself raises the observer’s
self-efficacy and motivates her or him to perform the same tasks. This is especially true among children who have experienced difficulties in learning skills (Schunk and Hanson, 1985). Moreover, Schunk et al. (1987) and Bandura and Menlove (1968) find advantages for multiple peers, presumably as these increase the probability that observers will perceive themselves as similar in competence to at least one of the role models. Especially when subjects doubt their learning capabilities, observation of multiple peers succeeding may better promote subjects’ self-efficacy (Bandura, 1986).

Given the pupil’s own IQ, the expected number of other pupils in the class with similar IQ increases with class size. Whether this mechanism indeed explains a zero net effect of class size reduction can be tested by including the number of pupils in the class with similar IQ as an additional regressor in the test score equations. As the PRIMA data include for each pupil in the sample its own IQ and the IQ-score of each of its classmates, we are able to construct the variable “number of classmates with similar IQ”. This variable has an expected value which increases with the size of the class. Our hypothesis is thus twofold: inclusion of this variable should reduce (or possibly even reverse) positive class size effects on achievement, and the new variable should have a positive direct effect on achievement. To make this test operational we have defined students with similar competence as those in a class that fall within a range of plus or minus half a grade-level standard deviation (+/-2 IQ-points) around the given pupil’s IQ. The correlation coefficient between this variable and actual class size equals 0.46 in grade 4, 0.31 in grade 6, and 0.40 in grade 8. It should also be noted that the number of similar pupils is not a direct measure of class homogeneity. Class homogeneity is best measured by a statistic like the variance. Although the number of pupils in the class with similar IQ and the

---

7Another possible mechanism leading to a positive relation between performance and the number of similar classmates is that teachers may target their instruction to groups of children with similar competence. If instruction time is divided in proportion to the size of the group, members of larger groups receive more instruction time.

8In a recent paper Lazear (1999) proposes another peer group mechanism in relation with class size effects. The driving force in his model is the presence of disruptive children who cause negative externalities (congestion). To explain our findings, we are, however, in search for a mechanism causing positive externalities.

9An alternative hypothesis is that children learn from individuals that are smarter than themselves. We tested this hypothesis by including number of students in an individual's class with an IQ higher than their own. Even after controlling for pupils' own IQ, this served merely as a proxy for IQ rank within a class producing a large negative effect with no significant decrease in the positive class size effect.

10As mentioned earlier, IQ was measured at the same time as the test scores. Although individual IQ scores have probably been affected by the size of the class, there seems no reason to assume that class size affects the number of classmates with similar IQ, in another way than the simple mechanism that with a larger number of classmates the expected number of similar classmates increases. Notice that with IQ measured afterwards, explicit ability tracking on the basis of measured IQ is prevented.
The causal effect of class size on scholastic achievement

The variance of IQ in a class are negatively correlated, this correlation is weak: -0.25 in grade 4, -0.18 in grade 6, and -0.16 in grade 8.

Table 2 presents estimation results from specifications identical to those reported in Table 1, except that the additional control variable “number of classmates with similar IQ” is included.\(^\text{11}\) In all (but one) cases the newly created variable has the predicted positive effect on performance, and is highly significant. For grade 4, we find a drop in the class size effects on arithmetic and language achievement for both the baseline OLS estimates as well as those produced by the IV procedure. Most notably, for language achievement the negative OLS baseline estimate proves to be significant at the 10% level. The results for grade 6 are even stronger: all point estimates of the class size effect drop considerably, and three out of four are significantly negative. In all cases the number of classmates with similar IQ has the predicted positive effect on achievement. The results for grade 8 also support the hypothesis set out above; the number of classmates with similar IQ has the predicted positive effect on achievement. The OLS estimates of the class size effect change from positive into negative after including the number of similar classmates as additional regressor. The two IV estimates of the class size effect for grade 8 which were significantly positive without a control for number of classmates with similar IQ are no longer significantly different from zero.

In conclusion, the non-negative, and sometimes even positive, relation between class-size and school performance in Dutch primary schools can – at least partially – be attributed to the fact that reduction of class size also reduces the (expected) number of pupils in the class with a similar level of competence. This reduction apparently limits a pupil’s scope to learn from her or his classmates.

V. Robustness

This section addresses the robustness of the main results by experimenting with different measures of class size, including higher order terms of total enrolment and school average SES-weight as covariates, a non-linear functional form of class size, and limited sub-samples.\(^\text{12}\)

Alternative measures of class size

As alternatives to the actual class size as reported by the teacher, the pupil-teacher ratio for each class has been constructed in addition to average class

\(^{11}\)We also estimated the same models including the class average IQ as an additional covariate.

\(^{12}\)To save space, we do not present tables with results of these different specifications. These tables are available from the authors.
<table>
<thead>
<tr>
<th></th>
<th>4th grade</th>
<th></th>
<th>6th grade</th>
<th></th>
<th>8th grade</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>First stage OLS</td>
<td>IV</td>
<td>OLS</td>
<td>IV</td>
<td>First stage OLS</td>
</tr>
<tr>
<td><strong>EFD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.161</td>
<td>(0.157)**</td>
<td></td>
<td></td>
<td></td>
<td>1.311</td>
</tr>
<tr>
<td><strong>Class size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.144</td>
<td>(0.152)</td>
<td>0.074</td>
<td>(0.420)</td>
<td>-0.240</td>
<td>0.051</td>
</tr>
<tr>
<td><strong>Number of students</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.406</td>
<td></td>
<td>1.078</td>
<td>0.975</td>
<td>0.558</td>
<td>0.421</td>
</tr>
<tr>
<td><strong>with similar IQ</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.057)**</td>
<td>(0.154)**</td>
<td>(0.241)**</td>
<td>(0.235)**</td>
<td>(0.318)**</td>
<td>(0.055)**</td>
</tr>
<tr>
<td><strong>Total enrolment</strong></td>
<td></td>
<td></td>
<td>0.003</td>
<td>0.011</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.004)</td>
<td>(0.012)</td>
<td>(0.013)</td>
<td>(0.013)</td>
</tr>
<tr>
<td><strong>SES-weight</strong></td>
<td></td>
<td></td>
<td>(1.474)</td>
<td>(3.160)</td>
<td>(3.311)</td>
<td>(3.313)**</td>
</tr>
<tr>
<td><strong>R-squared</strong></td>
<td></td>
<td></td>
<td>0.4023</td>
<td>0.1067</td>
<td>0.1053</td>
<td>0.1513</td>
</tr>
<tr>
<td><strong>Partial correlation</strong></td>
<td></td>
<td></td>
<td>0.3534***</td>
<td>0.4529***</td>
<td>0.4529***</td>
<td>0.4529***</td>
</tr>
<tr>
<td><strong>coefficients</strong></td>
<td></td>
<td></td>
<td>0.5061***</td>
<td>0.5061***</td>
<td>0.5061***</td>
<td>0.5061***</td>
</tr>
<tr>
<td><strong>P-value</strong></td>
<td></td>
<td></td>
<td>-0.249</td>
<td>-0.333</td>
<td>-0.258</td>
<td>0.184</td>
</tr>
</tbody>
</table>

Dependent variable in first stage is actual class size as reported by the teacher, dependent variables in second stage equations are pupils' percentile scores for arithmetic and language. All regressions also include a constant, four individual SES-class dummies, a dummy for pupil's gender, percentage of girls in pupil's class, teacher's gender and experience, dummy for dual teacher class, dummy for multi-grade class and three dummy variables for school's denomination. Robust standard errors taking account of correlated disturbance terms within schools are reported in brackets. ***/**/* indicate significance at the 1%/5%/10%-levels.
size within a school. The results for these two alternative class size measures are remarkably similar to those reported above in our main findings. In all but a few cases both the sign and significance level of the point estimates are the same. Therefore, we conclude that the findings are robust to differing measures of class size.

**Inclusion of higher order terms of total enrolment and school average SES-weight**

The instrumental variable EFD-class size plays a key role in determining the causal effect of class size on achievement. For this variable to be a valid instrument it has to be the case that it has no other effect on achievement other than through its effect on class size. As the instrumental variable is derived from official ministerial tables which translate SES-weighted total enrolment into the number of class teachers, it is important that the instrument does not pick up any direct effects from a school's SES-weighted total enrolment on achievement. In Section 4 we controlled for this by including the school's average SES-weight and total enrolment as covariates, but perhaps it is insufficient to enter these terms in linear form only.

We estimated specifications including quadratic terms in total enrolment and average SES-weight as well as an interaction term between total enrolment and average SES-weight. These results are fairly similar to those in Tables 1 and 2. For the estimations without "number of classmates with similar IQ", the main difference is that the IV estimates of the class size effects for 8th graders are not significantly positive, while they are so in Table 1. For the specification including the "number of classmates with similar IQ", inclusion of the higher order terms of total enrolment and school average SES-weight cause a drop of the coefficient of EFD in the first stage equation. Although also the partial correlation coefficient is lower, we still end up with a fairly strong instrument. The terms in enrolment and school average SES-weight do in most cases not have a significant impact on school performance.13

**Non-linear class size**

The above findings implicitly assume that the relationship between class size and scholastic achievement is of a linear form which may not be the case. To address the possibility of non-linearities, we changed the functional form of our model by using a two-step selection model with class size measured in

---

13This remains the case even after further smoothing the achievement/enrolment relation by including a cubic term in enrolment or a series of dummy variables for different total enrolment categories.
7 different categories. Following Bosker and Hox (1996), we broke class size into the following categories: 5–9, 10–14, 15–19, 20–24, 25–29, 30–35, 35 and above. In the first step we treated class size as an ordinal variable, and applied an ordered probit equation, in the second step the percentile achievement scores were just as before treated as cardinal variables. The identifying variable in this approach is the instrument EFD-class size.

When “number of classmates with similar IQ” is omitted, the results for 4th grade math achievement suggest that relative to classes with 25 to 30 students, those with more than 35 have a significant positive effect, while being placed in small classes with 5 to 9 pupils has a significant negative impact on achievement. The selectivity correction term is insignificant indicating that the endogeneity of class size has no effect on the OLS point estimates. The only significant effect of class size on 4th grade language achievement is for individuals in class sizes with 20 to 24 students (one category lower than the reference) who are estimated to score approximately 4 percentile points higher. For 6th graders, the relation between class size categories and achievement appears to be completely flat, with the exception of the negative effect of being placed in a very small class (5–9) for the arithmetic score. The estimates for the 8th grade make for a completely different story. Here, significant effects on math and language achievement of negative 3 to 4 percentile points are associated with the downsizing of a class from the range of 25–29 to 20–24 students.

With “number of classmates with similar IQ” included, results confirm those obtained before. The significantly negative effect on arithmetic scores of being placed in a small 4th grade-class vanishes, as does the significantly positive effect of being placed in a very large class. The same holds (in most cases) for the effects on arithmetic and language scores of being in an 8th grade class of 20–24 students.

The class size effects presented in Table 1 point to a flat achievement/class size profile. The results from the non-linear specification suggest that – with a few exceptions – the profile is flat everywhere. Formal tests confirm this: not only are most of the class size coefficients not significantly different from zero, they are also not significantly different from each other. Of the 90 possible pairwise comparisons of IV estimates, only 13 point to a significant difference.

**Homogeneity with respect to SES**

One possible explanation as to why we do not find the common negative correlation between class size and achievement may lie in composition of class with respect to SES. Indeed, policy made by the Ministry of Education encourages a higher allocation of resources for those with lower SES. It would
then be no surprise that those students with lower SES end up in smaller classes and have lower achievement on average, driving the result of a positive class size/achievement relationship. To test for this, we restrict our sample by preserving observations that are in relatively homogeneous classes with respect to SES. To be more precise, we limit the observations only to those unweighted individuals (SES weight equals unity) that are in classes with an average SES weight of 1.35 or less.

For the specification without “number of classmates with similar IQ”, results are almost identical to those in Table 1. For the specification with “number of classmates with similar IQ”, results differ, however, somewhat from those in Table 2. None of the class size effects is significantly negative. Also the point estimates of the effect of the number of classmates with similar IQ are somewhat smaller (but remain highly significant in all but four cases). This suggests that pupils from more disadvantaged social background benefit more from both the “pure” class size effect, and the presence of classmates with similar IQ.

Common support of “number of classmates with similar IQ” over the range of class sizes

To estimate the effect of class size on achievement requires to have the full range of class sizes at every value of the number of classmates with similar IQ. As the expected number of classmates with similar IQ is increasing with class size, common support over the range of class sizes may be problematic.\(^\text{14}\) For grades 4 and 6 common support is achieved when class size exceeds 17 pupils and when the number of classmates with similar IQ does not exceed 10. For 8th grade the respective numbers are 12 and 11. Restricting the sample to cases satisfying these criterions, we are left with 3515 observations for 4th grade, 3158 for 6th grade and 3579 for 8th grade. With regard to the estimates of the effect of class size on achievement signs and significance levels are identical to those reported in Table 2. The same holds for the estimates of the effect of the “number of classmates with similar IQ”; these remain all positive and highly significant.

Schools with one class per grade level

Endogeneity bias of the estimated effect of class size on achievement may arise from two different sources. One is school-selection by the parents, the other is sorting of pupils over classes by the schools. An easy remedy against biases from this second source is to restrict the sample to observations which

\(^{14}\)We thank the referee, who brought this to our attention.
have only one class at the grade level. As a result of the small scale of Dutch primary schools, selecting on the basis of this criterion leaves us with 80 percent of the observations. Repeating the analysis for this 80 percent generates results which are almost identical to the results reported in Tables 1 and 2.

VI. Conclusion

This paper analyses the causal effect of class size on scholastic performance in Dutch primary schools. Empirical estimation of this effect is difficult due to the possible endogeneity of class size. Two sources of endogeneity may play a role: choices by parents, where parents who care more about their children's achievement opt to send their children to schools with smaller classes, and selection by schools where weaker pupils are allotted to smaller classes. We correct for these biases via an instrumental variable approach based on ministerial rules determining the number of class teachers as a function of enrolment. The resulting class size function is discontinuous.

The estimation results show that after correcting for endogeneity, pupils in large classes do no worse — and sometimes even better — than identical pupils in small classes. This result holds for a large variety of functional forms, specifications and using restricted samples.

As an explanation for this result, we formulate a hypothesis based on the social cognitive learning literature that pupils' achievement benefits from a larger number of classmates with similar levels of competence. By reducing class size, the (expected) number of classmates with similar IQ falls which may have a negative effect on achievement. We test this hypothesis by constructing a new variable which counts for each individual the number of pupils in the class with the same IQ. In all specifications this new variable has the predicted positive effect on performance and is in most cases highly significant. Moreover, of the twelve class size effects that we estimate, two change from significantly positive into insignificantly positive when the number of similar classmates is included, two estimated effects remain insignificantly positive, four effects change from insignificantly positive to insignificantly negative, and four effects change from insignificantly negative to significantly negative.

It is important to realise that the new variable "number of classmates with similar IQ" differs from conventional measures of class homogeneity. Class homogeneity is measured in terms of the spread of characteristics and is not intended to be a measure varying with the number of pupils in the class. Instead, our measure explicitly intends to do so as we were in search of a mechanism explaining the non-negative effect of class size on performance.
The important policy conclusion to be drawn from the findings in this paper is the following: class size reduction at the scale currently implemented in the Netherlands is unlikely to be an effective intervention to boost performance. The reason is that reduction of class size does not only decrease the pupil-teacher ratio, but also decreases the (expected) number of classmates with a similar level of competence. It is tempting to draw policy conclusions based on the finding that a larger number of classmates with similar IQ increases achievement. This finding may even be regarded as a plea for tracking. It is important to realise, however, that the two achievement measures used in this paper (arithmetic and language) do not cover the entire range of skills and attitudes learned in schools. Schools also play a role in developing good citizenship which also includes (in our view) interacting with others who are dissimilar in a diverse environment.

Perhaps the best way to summarize the findings of this paper is to consider the following situation. A child comes home and tells her parents that the principal, in an effort to reduce class sizes, has decided to take three children out of her class. Should the parents of this child be happy with the message? According to our results, this depends on which children are taken out of the class. If these are children who stimulate the child’s learning behaviour, the parents should not favour this class size reduction.

Date of Receipt of Final Manuscript: July 2001

References


