Preface

The theory and practice of Swedish education policy has been analysed and discussed earlier in several reports from the Expert Group on Public Finance (ESO). More reports about schools are now being published. This report looks at how the reduction in class size affects pupils' achievement in the short and long term. One of the authors of the report is *Alan B. Krueger*, Professor of Economics and Public Affairs at Princeton University. The other author is *Mikael Lindahl*, Ph.D. in Economics and currently a researcher at University of Amsterdam.

In his research and studies on conditions in the USA, Alan B. Krueger has previously managed to disprove established research findings. However, Alan Krueger is also living proof that seemingly theoretical research can be transformed into practical policy. In the mid-1990s, together with Professor David Card of Berkeley University, Alan Krueger published a pioneering study on minimum wages. The findings of this study have had significant impact on practical policy. Card and Krueger were able to show convincingly that it is possible to raise minimum wages without adversely affecting employment.

In this report, the hypothesis that large class sizes do not have any effect on pupils' achievement is tested. A study conducted by the authors using pupils in the Stockholm area confirmed earlier American research findings on the effects of reducing class sizes. Also in Sweden, school results improve when class sizes are reduced. This finding is especially prevalent for pupils from foreign backgrounds.

Teaching in smaller classes costs more in the short term and it can also be difficult to free up resources straightaway by making savings elsewhere in the school system. However, this increased usage of resources can in the long term generate better returns for society as a whole. The conclusion of the authors is that these additional resources are needed for more teachers in order to make smaller classes possible. There is no evidence to confirm that other types of support resources would produce the same results. Perhaps it is also possible to reconsider priorities between different areas of the education system. No research has been conducted into whether or not the effects are equally significant at higher levels within the education system.

As is the case for all ESO reports, the authors themselves are responsible for the contents.

Stockholm, April 2002

Leni Björklund
Chairman of ESO

Authors' Preface

The authors thank Martin Hörnquist for having initiated this report and the reference group, especially Jan-Eric Gustafsson, for valuable comments. We are also very much acknowledges Marja Lemne. Most of all though, we thank pupils, teachers and principals in the schools in Stockholm who participated in the study for Sweden.

Stockholm, April 2002

Alan B. Krueger ***Mikael Lindahl***

Contents

Figures

Tables

1 Summary

The resources spent on primary and lower secondary education has decreased during the 1990s in Sweden, leading to an increase in the number of pupils per teacher by more than 20 percent between 1991 and 1999. At the same time, in the Swedish public policy debate, some have argued that scholastic achievement has little or nothing to do with the amount of resources devoted to schools. Instead, the debate on the performance of Swedish schools has in recent years focused more on using existing resources more efficiently: for instance, by changing the organization within schools and of the schooling system. Until recently, the consensus among many researchers was also that school resources have little or no influence on student achievement.

1.1 International evidence

This study first evaluates and reinterprets the evidence from Eric Hanushek's influential review of the connection between school resources and pupils' achievement. Hanushek's review concluded that spending more school resources – either by smaller classes or higher expenditure per pupil – is unlikely to have a significant effect on pupils' achievement. In his quantitative analysis, Hanushek extracted as many as 24 estimates from some studies, and only one from others. We show that his main conclusion is due to an over-weighting of studies that find a negative effect of school resources. When we weight each study equally, we find that significantly more studies found positive achievement effects of increased school resources. Hence, unless one weights the existing studies of school resources in a peculiar way, the *average study* tends to find that more resources are associated with greater student achievement.

We then turn to the only large-scale, randomized experiment ever conducted on the class size issue, the Tennessee STAR experiment. A randomized experiment is particularly desirable in this case because randomization should ensure that students in smaller and larger classes are otherwise identical, on average. In STAR pupils and teachers were randomized into small and large classes. The experiment lasted from kindergarten to 3rd grade, after which pupils were reallocated to normal sized classes. We first present evidence that the STAR experiment was indeed implemented properly. We next present results from this experiment which suggest beneficial achievement effects of smaller classes, and that this effect is especially prevalent for disadvantaged groups. It is also the case that even though the experiment only lasted until third grade, the test score gains persisted into later grades, although they were attenuated. We also present evidence that black pupils from small size classes increased their college-test taking probabilities and scored significantly higher on college tests almost 10 years after the experiment ended.

1.2 Swedish evidence

Next we turn to evidence for Sweden. Noting the very sparse and old evidence on the class size issue based on Swedish data, we collected new data on pupils in the municipality of Stockholm. Since we did not have the opportunity to run an experiment like STAR in Sweden, we collected the data with the special purpose of mimicking the design of an actual experiment as closely as possible. We therefore tested the pupils three times, before and after a 10-week summer break, and at the end of the following school year. This design made it possible to relate the change in test scores during the school year to class sizes, controlling for the change in the test scores during the summer. The intuition behind this approach is that schools are closed during the summer, and hence schooling characteristics should not influence the summer test score change, whereas schools are open during the school year, when schooling characteristics are expected to influence test score changes. The experience during the summer months thus provides a way of adjusting for non-school influences on the level and change in student achievement. The results from the study for Stockholm were that smaller class sizes positively affected achievement levels, and that this positive effect was especially prevalent for a group of pupils that on average has relatively low achievement levels, i.e. pupils with non-Swedish parents.

1.3 Recommendations

This conclusion does not, of course, mean that reducing class size is necessarily worth the additional investment. This question requires knowledge of the strength of the relationships between class size and economic and social benefits and information on the cost of class size reduction. We therefore calculated the likely costs and benefits of a 7 pupil class size reduction for the U.S., using the results from STAR, and of a similar reduction for Sweden, using the results from the Stockholm sample. We conclude that decreasing class size can, under reasonable assumptions, indeed have benefits that are at least as great as the costs.

We believe that the actual, state-wide class-size reduction initiative in California could provide a model for Sweden. In this enormous education reform, which was championed by then-Governor Pete Wilson, California school districts that chose to participate received just over $800 for each K–3 student enrolled in a class of 20 or fewer students to encourage smaller classes.[1] Because of the scale of this intervention, many implementation problems were encountered that do not arise in small-scale demonstration studies. Still, Stecher et al. (2000) find that, after two years, the California class-size reduction initiative led to increased math and reading scores.

In this study we have shown that, contrary to what many have argued, the weight of the available evidence suggests that smaller classes increase pupil achievement. We therefore conclude that the increase in pupil teacher ratios observed in Sweden during the 1990s is likely to have had negative consequences for the scholastic achievement of Swedish pupils.

Lastly, we would recommend great care in pursuing large changes in schooling inputs, such as class sizes, which has occurred during the 1990s. A way to decrease the likelihood of making missteps in school policy is to run controlled field experiments on a limited basis, before policies are implemented on a wide scale. We also believe it is important to increase the standard of test score data collection in Sweden. At present, standardized tests are given at several different grades in primary and lower secondary grades.

[1] K-3 students are puils enrolled in grades 1–3 or in the pre school year.

However, only the test in the 9th grade is compulsory and collected centrally.

2 Introduction

The resources spend on primary and lower secondary education has decreased during the 1990s in Sweden, and this have especially had consequences for the number of pupils per teacher. This paper attempts to analyse whether it is likely that this has led to decreased achievement among Swedish pupils. To investigate this, we survey the existing international evidence, conduct new analysis with experimental data for the US and with newly collected data for Sweden. We also attempt to quantify the cost and benefit of decreasing class sizes.

Until recently, the consensus among researchers has been that school resources have little or no influences on student achievement. Hence it is not surprising that in the public policy debate some have argued that the level of scholastic achievement among Swedish pupils has little or nothing to do with the amount of resources devoted to schools. Instead the debate of the performance of Swedish schools has in recent years focused more on using existing resources more efficiently, for instance by changing the organization within schools and of the schooling system. However, in this study we will focus on the effect of class size on pupils achievement level. Since most of the variation in school resources per pupil over time is due to variation in the number of pupils per teacher, we do believe that the effect of class size on achievement is a good indicator of the effect of school resources on achievement. Class size is also a school policy instrument that is easy to understand and to implement, and much of the public policy debate in Sweden did for long focus on this issue.[2] Therefore we later only spend some time on the more general school resource measure; expenditures per pupil.

[2] We will in this study not investigate the effect of other schooling inputs, such as the effect of teacher quality or the organisation of schools. This does of course not imply that we believe these things are unimportant.

Whether smaller class sizes really transfer into higher pupil achievement has been questioned in the public policy debate in Sweden. For instance, Hans Bergström, in the editorial page of Dagens Nyheter, Sweden's biggest morning paper, has written that: ”The research is also unambiguous: class size has almost no connection with school results,” (Bergström, 1998). His statement is not a surprise if one look at what some researchers in Sweden has concluded. For instance, Ingemar Fägerlind, Professor and former director of the Institute of International Education at Stockholm University, has stated that, ”few research results support a connection between the size of groups taught and pupils achievement.” (Fägerlind, 1993). Notably, the Swedish Economy commission, in one of their more than a hundred policy recommendations, did recommend larger class sizes (Lindbeck et al., 1993).

What then are the view of class size among parents and teachers? In 1997, the National Agency for Education in Sweden asked about 5000 pupils (in grades 7–9 and high school), parents and teachers (grades 1–9 and high school) about their view about how their school worked (see National Agency for Education report 144). Among pupils, 17 percents thought their class was too large. This number is about the same for pupils in grades 7–9 as for high school. Parents were asked to evaluate their child's school on several aspects. About 30 percent of the parents were of the view that the size of the class of their child were poor or very poor. Less then 40 percent of the teachers in grades 1–9, were satisfied or very satisfied with the size of their class. Based on these numbers it seems fair to say that many teachers and parents are quite unhappy about the class size situation in Swedish schools.

The next section reviews the trends in resources available to Swedish schools. Section 4 review international non-experimental class size research and reinterpret the results from the most influential survey on the class size issue to date. Section 5 present results from the only large-scale experiment ever conducted on the class size issue, which was done in Tennessee in the US. In section 6 we review earlier research and present new data for Sweden. In section 7, we conduct our own analysis based on the newly collected data for the Municipality of Stockholm. We conduct the analysis using the most common method of analysis, and we also present a new non-experimental method of analyzing class size effects. Section 8 contains a cost-benefit analysis of decreasing class sizes from 22 to 15 pupils in the US and of a similar magnitude in Sweden. Section 9 concludes and discusses policy implications of the results.

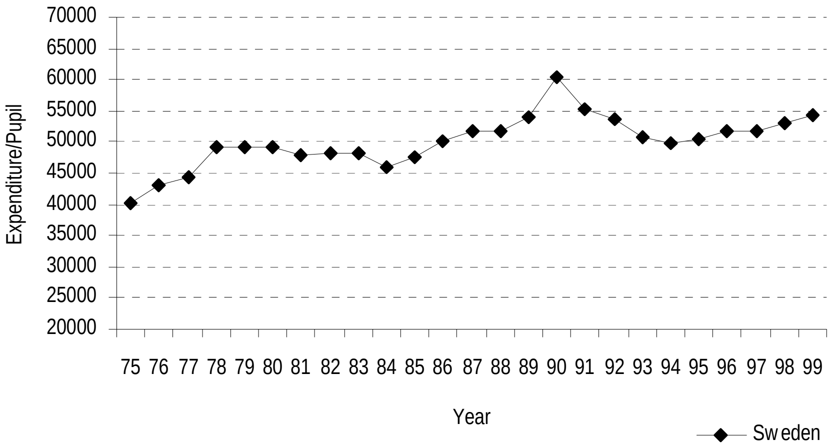
3 Trends in Resources Devoted to Swedish Schools

This section review trends in school resources and pupil-teacher ratios in Swedish elementary and secondary schools during the last 25 years. We also relate the spending on school resources to Swedish GDP figures. A comparison to the development in US during the same time is also made. For Sweden, data from Statistics Sweden (SCB) and the National Agency for Education are used. For US, data from the National Center for Education Statistics (Digest of Education Statistics, 1999) are used.[3]

In Figure 3.1 we show the expenditure per pupil in primary and lower secondary schools (grundskolan) in Sweden, using data from Statstics Sweden (SCB) and the National Agency for Education.

[3] We have for both Sweden and US used the best data available. Data for Sweden is from SCB (SM9201) until 1990, and after that from National Agency for Education publications (Skolverkets ”Beskrivande data om skoverksamheten” for respective year. GDP data for Sweden is from Statistics Sweden (SCB). For Sweden the expenditures on pre-schools and schools for disabled pupils as well as for KOMVUX (municipality based adult education) are not included in the expenditure measure. All data for US are from Digest of Education Statistics, 1999. In expenditures for US, the expenditures related to the last pre-school year is included in some cases.

Figure 3.1. **Expenditure per pupil in Sweden, primary and lower secondary schools (grundskolan), in SEK 1999 prices**

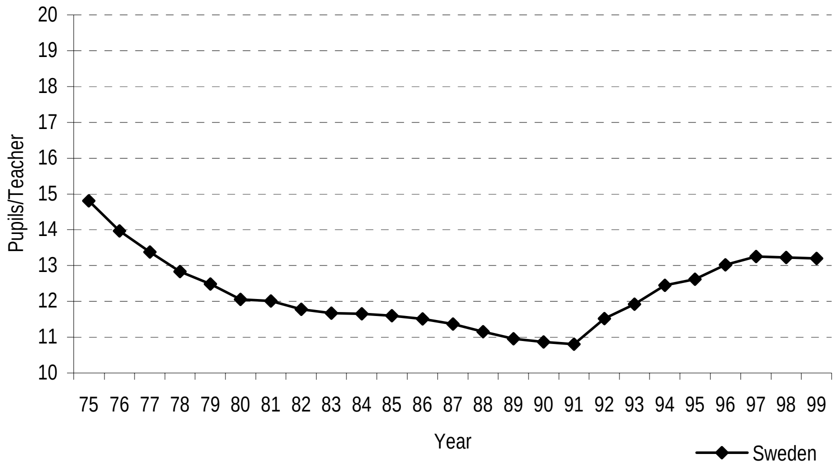


From mid 70s to 1990, the expenditure per pupil increased by 50 percent. However, in the early 90s, expenditure per pupil decreased sharply, so that in mid 90s, it was at about the same level as in the end of the 70s. At the second half of the 90s expenditure per pupil increased somewhat, and in 1999, this number was SEK 54,200 (Swedish Kronor).[4]

In Figure 3.2, the number of pupils per full time equivalent teacher in primary and lower secondary schools in Sweden are shown.

[4] Note that the costs for the years 1975–1990 refer to the fiscal years 1975/76–1990/91. For 1991 and later years the costs are calculated for the calender years. The information sources have also changed between these periods. Also note that in 1995 the calculation of expenditures changed again. From 1995 and onwards the costs per pupil is is 2 percent higher, relative to earlier years, because of this. Hence, cost comparisons between years should be done with care.

Figure 3.2. **Pupils per teacher in Sweden, primary and lower secondary schools (grundskolan)**



As expected, the trend is inversely related to the trend for cost per pupil, although the pattern is less erratic. From mid 70s to 1991, pupils per teacher, decreased from 15 to about 11. From 1991, pupils per teacher increased from 11 to more then 13. These trends are both due to changes in the number of pupils and teachers over time.

Note that number of pupils per teacher can be a crude measure of class size, since this measure also includes special and subject teachers. The more common it is with only one teacher in a class, the more similar are the measures of class size and pupils per teacher. It should be noted that in our own study for Stockholm, we use the number of pupils actually taught together as a measure of class size. This measure should be very close to the number of pupils per teacher. Another measure of class size is the number of pupils in the class that regularly are taught together. Until 1994, the National Agency of Education provided an estimate of the regular class size. In 1994, the average class size was 22.1 in grades 1–9. It is however not obvious how to translate the increase in pupil/teacher ratios into changes in class sizes. It is therefore unfortunate that nationwide data on class sizes no longer is collected for Sweden.

If we compare Figures 3.1 and 3.2 we see that even though the expenditure per pupil has increased slightly during the second half of the 90s, the pupil teacher ratio has surprisingly also increased somewhat. This is partly because of increased teacher wages during the end of this period. However, for most of the period since 1975, there has been an inverse relationship between expenditure per pupil and the pupil/teacher ratio.

In Figure 3.3, the trend in cost per pupil in Swedish primary and secondary schools (Grundskolan + Gymnasium) between 1978–99 is shown. The trend is similar as in Figure 1, but the level is somewhat higher, reflecting the higher cost per student in upper secondary schools.

Figure 3.3. **Expenditure per pupil in primary and secondary schools in Sweden and US. In SEK 1999 prices.**

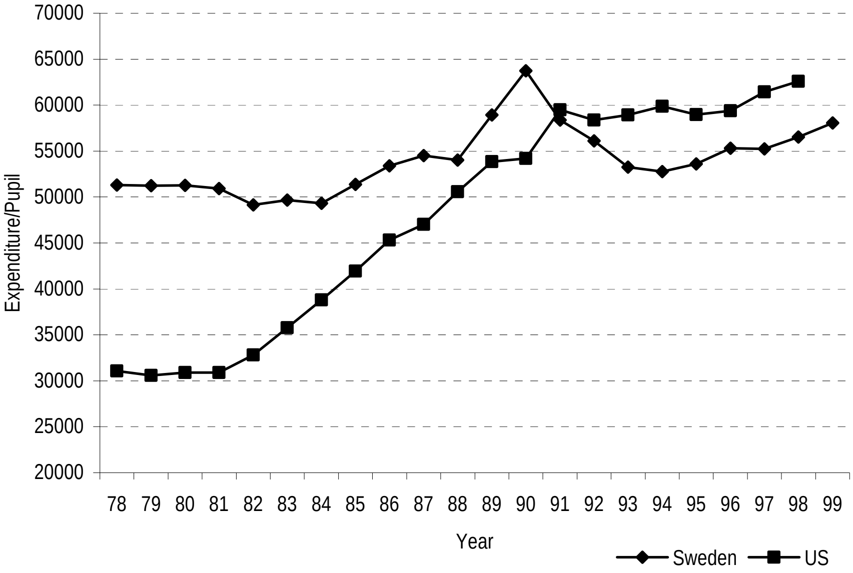
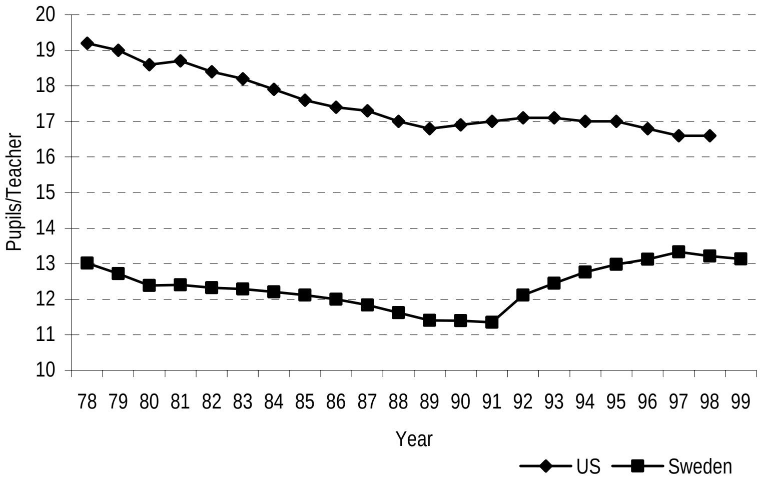


Figure 3.4 shows the pupil teacher ratios for the same schools which looks similar to figure 3.2, although it should be noted that the pupil-teacher ratio decreased somewhat in upper secondary schools during the last half of the 90s.

Figure 3.4. Pupils per teacher in primary and secondary schools in Sweden and US

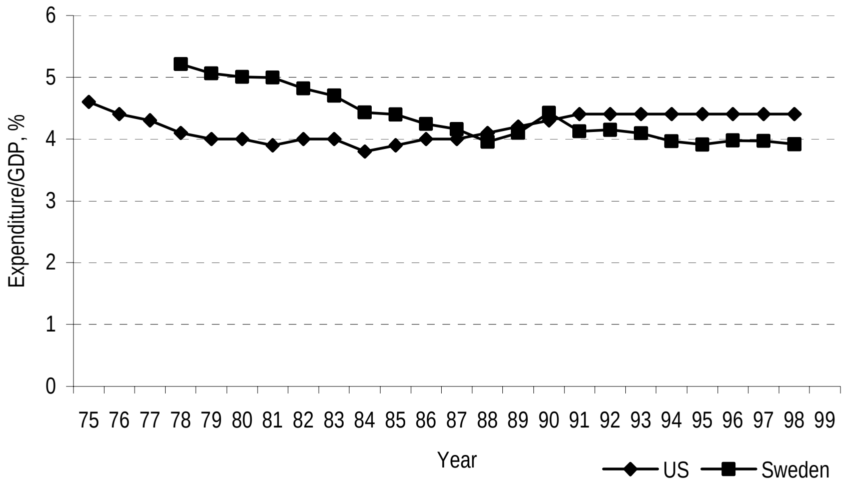


Figures 3.3 and 3.4 also show the cost per pupil and pupils per teacher for US schools. To be able to compare costs between Sweden and US, we convert the US figures by using GDP purchasing power parities (PPP) from OECD. PPP is a number estimated for each year with the purpose of converting prices for identical commodities in different countries. For instance the US-Sweden PPP in 1998 was 9.77, which means that if a commodity did cost $1 in US in 1998, 9.77 Swedish Kronor was needed to buy the same commodity in Sweden in 1998. It should be noted that converting educational costs from one country to another is far from easy, since educational inputs (for instance teachers) are to a high degree country specific. However, we here follow previous researchers in using PPP:s in educational cost comparisons between countries (see for instance Barro & Lee (1997)).

The expenditure per pupil did roughly doubled between 1978 and 1998 in the US. This partly reflects the stronger US currency during this period. However, even if expenditure per pupil in US is measured in fixed US dollars, the increase was about 50 percent. The increase in expenditure per pupil in Sweden during the same period is about 10 percent. Even though US spend more on education per pupil compared to Sweden, the pupil-Teacher ratio is lower in Sweden. One important reason for this is that teachers in US earn higher wages, compared to teachers in Sweden.

When comparing the expenditures on education between countries it is obvious that the relatively richer country can afford to spend more on education. It is therefore important to note that the Swedish economy grew (per capita) by 52 percent between 1978 and 1998, whereas the US growth figure for the same period is 65 percent. Hence, one explanation for the different trends in figure 3.3, is that US has more resources to spend on education compared to Sweden. However, that this is not the only reason for the different developments in expenditure per pupil, can be seen in Figure 3.5, where the trend in expenditure on primary and secondary education in relation to the country-GDP is shown.

Figure 3.5. **Educational expenditure/GDP (primary and secondary schools) in Sweden and US**



In US, about 4.5 percent of GDP is spend on education in primary and secondary schools, and this number is about the same in 1998 as in 1975. In Sweden more then 5.2 percent of the GDP where spent on primary and secondary education in 1978.[5] In 1998, this number has decreased to 3.9 percent. As a comparison, total educational expenditure as a fraction of GDP has in Sweden decreased from 7.9 percent in 1978 to 7.6 percent in 1998.[6] This means that while expenditure on primary and secondary schooling as a fraction of GDP, has decreased during this period, other schooling sectors has increased their expenditure during the same period. This is mostly due to increased expenditures on university education. We also note that the expenditures on KOMVUX was almost one-third of the expenditures on upper secondary schooling in Sweden.

[5] In Eduacation a Glance, OECD 2001, the educational expenditures are presented as a fraction of GDP 1998 for US and Sweden, among other countries, in their Table B2.1c. For Sweden, it is said that 4.5 percent is spent on 'primary & lower secondary' (3.0 percent) and 'upper secondary' (1.5 percent) education. The difference between this number and the number underlying Figure 3.5, which show 3.9 percent for Sweden in 1998, depends mostly on the expenditures to KOMVUX, which is not included in the expenditures underlying Figure 3.5. The OECD-number for all 'primary, secondary and post secondary non-tertiary education' for US is clearly lower, 3.7 percent, than in Figure 3.5, which is 4.4 percent. Halft of the difference is probably due to that in the latter number a large part of the education the last pre-school year in US is included. It is unclear what causes the rest of the difference.

[6] The numbers are from SCB. Note that expenditures on labor market education earlier was included in the education statistics. This expenditure is however not inclued in the 7.9 percent.

4 Reanalysis of Literature Review

The most influential summary of the effect of school resources on pupils' achievement is Eric Hanushek's 1986-article in Journal of Economic Literature. This study is among the most cited studies in Economics.[7] His 1986-survey has however been updated in later studies, so this reanalysis focus on the most recent literature review, Hanushek (1997). Hanushek's all reviews have concluded that more school resources, in terms of smaller classes or higher expenditure per pupil, is unlikely to have any significant effect on pupils achievement. In this section, we show that this result is due to an over-weighting of studies showing no or negative effects of more school resources. When we weight each study equally, significantly more studies found positive effects of increased school resources.

4.1 Hanushek's influential summary of the literature

The classification of estimates and studies underlying the literature summary in 1997 was kindly provided by Eric Hanushek. The same data are used in Hanushek (1996a, 1996b and 1998). Hanushek (1997; p. 142) describes his sample selection as follows:

> This summary concentrates on a set of published results available through 1994, updating and extending previous summaries (Hanushek, 1981, 1986, 1989). The basic studies meet minimal criteria for analytical design and reporting of results. Specifically, the studies must be published in a book or journal (to ensure a minimal quality standard), must include some measures of family background in addition to at least one measure of resources devoted to schools, and must provide information about statistical reliability of the estimates of how resources affect student performance.

[7] The study was cited 336 times in March 2001, according to the Social Science Citation Index.

He describes his rule for selecting estimates from the various studies in the literature as follows:

> The summary relies on all of the separate estimates of the effects of resources on student performance. For tabulation purposes, a 'study' is a separate estimate of an educational production found in the literature. Individual published analyses typically contain more than one set of estimates, distinguished by different measures of student performance, by different grade levels, and frequently by entirely different sampling designs.

Most of the studies included in Hanushek's literature summary were published in economics journals. The modal journal was the Economics of Education Review, which accounted for 22 percent of the articles in his sample and 35 percent of the estimates.

Table 4.1 summarizes the distribution of the estimates and studies underlying Hanushek's literature summary.

Table 4.1. **Distribution of class size studies and estimates taken in Hanushek (1997)**

Number of Estimates Extracted	Number of Studies	Total Number of Estimates	Percent of Studies	Percent of Estimates
(1)	(2)	(3)	(4)	(5)
1	17	17	28.8%	6.1%
2-3	13	28	22.0%	10.1%
4-7	20	109	33,9%	39,4%
8-24	9	123	15.3%	44.4%
Total	59	277	100.0%	100.0%

Note: Column (1) categorizes the studies according to the number of estimates that were taken from the study. Column (2) reports the number of studies that fall into each category. Column (3) reports the total number of estimates contributed from the studies. Column (4) reports the number of studies in the category as a percent of the total number of studies. Column (5) reports the number of studies in the category as a percent of the total numer of estimates used from all the studies.

The first column reports the number of estimates used from each study, classifying studies by whether only one estimate was taken (first row), two or three were taken (second row), four to seven were taken (third row), or eight or more were taken (fourth row). Only one estimate was taken from 17 studies. Nine studies contributed more than seven estimates each. These nine studies made up only 15 percent of the total set of studies, yet they contributed 44 percent of all estimates used. By contrast, the 17 studies from which one estimate was taken represented 29 percent of studies in the literature and only 6 percent of the estimates.

A consideration of Hanushek's classification of some of the individual studies in the literature helps to clarify his procedures, and indicates problems associated with weighting studies by the numbers of estimates extracted from them. Two studies by Link and Mulligan (1986 and 1991), each contributed 24 estimates – or 17 percent of all estimates. Both papers estimated separate models for math and reading scores by grade level (3rd, 4th, 5th or 6th) and by racial background (black, white, or Hispanic), yielding 2 x 4 x 3 = 24 estimates apiece. One of these papers, Link and Mulligan (1986), addressed the merits of a longer school day, using an 8-percent subsample of the data set used in Link and Mulligan (1991). In their 1986 paper, the interaction between class size and peer ability levels was included as a control variable, without a class-size main effect. In their text, however, Link and Mulligan (1986; p. 376) note that when they included class size in the 12 equations for the math scores, it was individually statistically insignificant. In an e-mail communication Eric Hanushek explained that he contacted Link and Mulligan to ascertain the significance of the class-size variable if it was included in their 12 reading equations. This procedure would seem to violate the stated selection rule that restricted estimates to a "set of published results."

Another issue concerns the definition of family background for estimate selection. The Link and Mulligan (1991) paper controlled for no family background variables, although it did estimate separate models for black, white and Hispanic students. Evidently, this was considered a sufficient family background control to justify the extraction of 24 estimates in this case. Card and Krueger (1992), however, reported several distinct estimates of the effect of the pupil-teacher ratio on the slope of the earnings-education gradient using large samples of white males drawn from the 1970 and 1980 Censuses, but only one estimate was selected from that paper. In an e-mail correspondence Hanushek explained that he extracted only one estimate from this study because only one specification controlled for family background information – although all estimates conditioned on race in the same fashion as Link and Mulligan (1986).

No estimates were selected from Finn and Achilles's (1990) analysis of the STAR experiment because it did not control for family background (other than race and school location), even though random assignment of students to classes in that experiment should assure that family background variables and class size are orthogonal (see section 5).

Summers and Wolfe (1977) provides another illustration of the type of researcher discretion that was exercised in extracting estimates. Summers and Wolfe analyze data for 627 sixth-grade students in 103 elementary schools. They mention that data were also analyzed for 533 eighth-grade students and 716 twelfth grade students, with similar class-size results, but these results were not included in Hanushek's tabulation. Summers and Wolfe (1977;Table 1) provide two sets of regression estimates: one with pupil-specific school inputs and another with school-averages of school inputs. They also provide pupil-level estimates of class-size effects estimated separately for subsamples of low, middle and high achieving students, based on students' initial test scores (see their Table 3). Hanushek selected only one estimate from this paper – the main effect from the student-level regression. Why the estimates reported for the various subsamples were excluded is unclear. In addition, because Hanushek (1991) draws inferences concerning the effect of the level of aggregation of the data on the estimates, it is unfortunate that results using both sets of input data (pupil-level or school-level) were not extracted. Contrary to Hanushek's conclusion about the effect of data aggregation, Summers and Wolfe (1977; p. 649) conclude, "when there are extensive pupil-specific data [on inputs] available, more impact from school inputs is revealed."

In a small number of cases, estimates were misclassified and unpublished estimates were selected. Kiesling (1967), for example, was classified as having 3 estimates of the effect of class size, but there is no mention of a class size variable in Kiesling's article. Eric Hanushek informed that Kiesling's estimates were taken from his unpublished thesis, which seems to violate his intention of using published estimates. In Montmarquette and Mahseredjian (1989), the sign of the class-size result was inadvertently reversed.[8]

[8] We have not corrected this error because we want to emphasize that the discrepancy in results comes from the weighting of the studies.

4.2 Alternatively weighted tabulations

Class Size

Table 4.2. Reanalysis of Hanushek´s (1997) literature summary; studies of class size

Result	Class Size Weighted by No. of Estimates Extracted (1)	Equally-Weighted Studies (2)	Studies Weighted by Journal Impact Factor (3)
Posivite & Stat. Sig.	14.8%	25.5%	34.5%
Positive & Stat. Insig.	26.7%	27.1%	21.2%
Negative & Stat. Sig.	13.4%	10.3%	6.9%
Negative & Stat. Insig.	25.3%	23.1%	25.4%
Unkown Sign & Stat. Insig.	19.9%	14.0%	12.0%
Ratio Positive to Negative	1.07	1.57	1.72
P-Value	0.500	0.059	0.034

Notes: Column (1) is from Hanushek (1997; Table 3), and implicitly weight studies by the number of estimates that were taken from each study. Columns (2) and (3) are from Krueger (2000). Column (2) assign each study the fraction of estimates corresponding to the result based on Hanushek´s coding, and calculate the arithmetic average. Column (3) calculates a weighted average of the data in column (2), using the journal impact factors as weights. A positive result means that a small class size is associated with improved student performance. Columns (1)–(3) are based on 59 studies. P-value corresponds to the proportion of times the observed ratio, or a higher ratio, of positive to negative results would be obtained in 59 indeptendent Bernouli trials in which positive and negative resultats were equally likely.

Column 1 of Table 4.2 summarizes Hanushek's tabulation of the estimates he selected from the literature. His approach equally weights all 277 estimates that were extracted from the underlying 59 studies. Following Hanushek, estimates that indicate smaller classes are associated with better student performance are classified as positive results.[9] The bottom of the table reports the ratio of the number of positive to negative results, and the p-value that corresponds to the chance of obtaining so high a ratio from a series of 59 independent Bernoulli trials. The results in column (1) are unsystematic – positive and negative estimates are virtually equally likely to occur. Only one quarter of the estimates are statistically significant, and the statistically significant estimates are also about equally likely to be positive and negative.

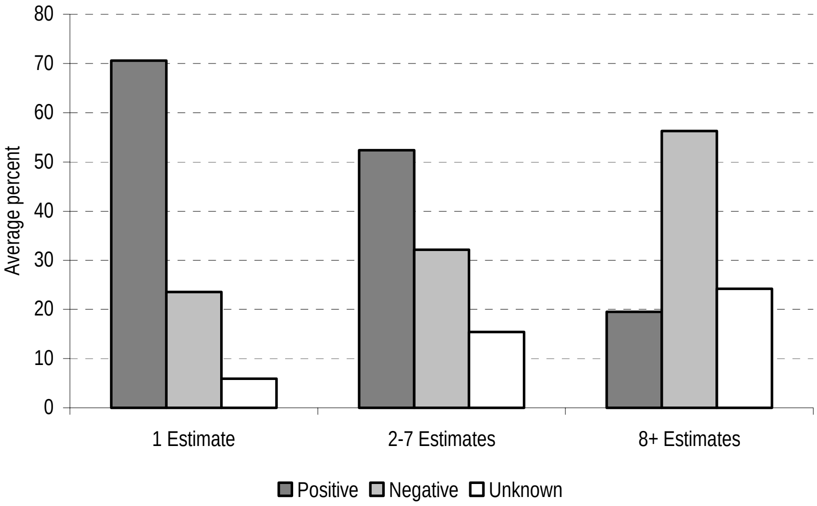
[9] We here follow the practice of using the terms class size and pupil-teacher ratio interchangeably. Also see discussion in section 3.

As mentioned, Hanushek's procedure places more weight on the studies from which he extracted more estimates. There are a number of reasons to question the statistical properties of such an approach. First, the procedure places more weight on estimates that are based on subsamples, all else equal. The optimal weighting scheme would do just the reverse. Second, authors who find weak or negative results (e.g., because of sampling variability or specification errors) may be required by referees to provide additional estimates to probe their findings (or they may do so voluntarily), whereas authors who use a sample or specification that generates an expected positive effect of smaller classes may devote less effort to reporting additional estimates for subsamples. If this is the case, and findings are not independent across estimates (which would be the case if a misspecified model is estimated on different subsamples), then Hanushek's weighting scheme will place more weight on insignificant and negative results.

Third, and perhaps most importantly, the uneven application of Hanushek's stated selection rule raises questions about the discretion of the researcher in selecting many or few estimates from a particular paper. A good case could be made, for example, that more estimates should have been extracted from Summers and Wolfe (1977), and fewer from Link and Mulligan (1986 and 1991). Weighting studies equally lessens the impact of researcher discretion in selecting estimates.

Figure 4.1 provides evidence that Hanushek's procedure of extracting estimates assigns more weight to studies with unsystematic or negative results.

Figure 4.1. **Average percent of estimates positive, negative or unknown sign, by number of estimates taken from study**



Notes: Based on data from Hanushek (1997). Arithmetic averages of percent positive, negative and unknown sign are taken over the studies in each category.

The figure shows the fraction of estimates that are positive, negative or of unknown sign, by the number of estimates Hanushek took from each study. For the vast majority of studies, from which Hanushek took only a small number of estimates, there is a clear and consistent association between smaller class sizes and student achievement. For the 17 studies from which Hanushek took only one estimate, for example, over 70 percent of the estimates indicate that students tend to perform better in smaller classes, and only 23 percent indicate a negative effect. By contrast, for the nine studies from which Hanushek took a total of 123 estimates the opposite pattern holds: small classes are associated with lower performance.

Table 4.3. **Regressions of percent of estimates positive or negative, and significant or insignificant, on the number of estimates used from each study; class size studies**

	Dependent Variable:				
	Percent Positive & Significant	Percent Positive & Insignificant	Percent Negative & Significant	Percent Negative % Insignificant	Percent Unknown Sign & Insignificant
	(1)	(2)	(3)	(4)	(5)
Intercept	35.7	27.4	7.4	21.0	8.5
	(6.4)	(6.0)	(4.5)	(5.9)	(5.6)
Number of Estimates Used	-2.16	-0.07	0.62	0.44	1.18
	(0.96)	(0.89)	(0.66)	0.88)	(0.83)
R-square	0.08	0.00	0.01	0.00	0.03

Notes: Standard errors are shown in parentheses. Sample size is 59 studies. Dependent variable is the percent of estimates used by Hanushek in each result category. Unit of observation is a study.

Table 4.3 more formally explores the relationship between the number of estimates that Hanushek extracted from each study and their results. Specifically, column (1) reports results from a regression in which the dependent variable is the percent of estimates in a study that are positive and statistically significant (based on Hanushek's classification) and the explanatory variable is the number of estimates that Hanushek took from the study. The unit of observation in the table is a study, and the regression is estimated for Hanushek's set of 59 studies. Columns 2–5 report analogous regression equations where the dependent variable is the percent of estimates that are positive and insignificant, negative and significant, negative and insignificant, or of unknown sign, respectively. In Hanushek's summary, there are fewer estimates from studies that tended to find positive and significant results (r = -0.28), and this relationship is stronger than would be expected by chance alone. Moreover, the opposite pattern holds for studies with negative and significant findings: relatively more estimates from studies with perverse class size effects are included in the sample, although this relationship is not significant.

Also notice that in 20 percent of the estimates that Hanushek extracted, the researchers had not reported the sign of the coefficient on the class-size variable. Statistical studies that do not report the coefficient of the class-size variable – let alone its sign – are unlikely to be high quality studies of the effect of class size. Table 4.3 and Figure 4.1 indicate that the incidence of unreported signs rises with the number of estimates extracted from a study, which suggests that the quality of the study does not rise with the number of estimates extracted from it.

The rule that Hanushek used for selecting estimates would be expected to induce a positive association between the prevalence of insignificant results and the number of estimates taken from a study, since studies with more estimates probably used smaller subsamples (which are more likely to generate insignificant estimates). But this sampling bias cannot explain the inverse relationship between the number of estimates taken from a study and the prevalence of statistically significant, positive estimates of class size effects.

As a partial correction for the oversampling from studies with negative estimates, in column (2) of Table 4.2, the underlying studies – as opposed to the individual estimates extracted from the studies – are given equal weight. This is accomplished by assigning to each study the percent of estimates that are positive and significant, positive and insignificant, and so on, and then taking the arithmetic average of these percentages over the 59 studies.[10] This simple and plausible change in the weighting scheme substantially alters the inference one draws from the literature. In particular, studies with positive effects of class size are 57 percent more prevalent than studies with negative effects.

In column (3) an alternative approach is used. Instead of weighting the studies equally, the studies are assigned a weight equal to the 1998 “impact factor” of the journal that published the article, using data from the Institute for Scientific Information. The impact factors are based on the average number of citations to articles published in the journals in 1998. Impact factors are available for 44 of the 59 studies in the sample; the other 15 studies were published in books, conference volumes, or unpublished monographs. Studies not published in journals were assigned the impact factor of the lowest ranked journal. The weighted mean of the percentages is presented in column 3 of Table 4.2. Although there are obvious problems with using journal impact factors as an index of study quality (e.g., norms and professional practices influence the number of citations), citation counts are a widely used indicator of quality, and the impact factor should be a more

[10] For example, if a study was classified as having one estimate that was positive and significant and one that waspositive and insignificant, these two categories would each be assigned a value of 50 percent, and the others would be assigned 0. If a study reported only estimate, the corresponding category would be assigned 100 percent for that study.

reliable measure of study quality than the number of estimates Hanushek extracted. The results are quite similar when either the arithmetic mean or journal-impact-weighted mean is used. In both cases, studies with statistically significant, positive findings outweigh those with statistically significant, negative findings by more than two to one.[11]

Expenditures per Student

Table 4.4 **Reanalysis of Hanushek's (1997) literature summary; studies of expenditures per student**

Result	Expenditures per Student Weighted by No. of Estimates Extracted (1)	 Equally-Weighted Studies (2)	 Studies Weighted by Journal Impact Factor (3)
Posivite & Stat. Sig.	27.0%	38.0%	40.1%
Positive & Stat. Insig.	34.3%	32.2%	28.0%
Negative & Stat. Sig.	6.7%	6.4%	6.3%
Negative & Stat. Insig.	19.0%	12.7%	8.3%
Unkown Sign & Stat. Insig.	12.9%	10.7%	17.3%
Ratio Positive to Negative	2.39	3.68	4.66
P-Value	0.0138	0.0002	0.0001

Notes: Column (1) is from Hanushek (1997; Table 3), and implicitly weight studies by the number of estimates that were taken from each study. Columns (2) and (3) are from Krueger (2000). Columns (2) assigns each study the fraction of estimates corresponding to the result based on Hanushek´s coding, and calculate the arithmetic average. Column (3) calculates a weighted average of the data in column (2), using the journal impact factors as weights. A positive result means that greater expenditures are associated with improved student performance. Columns (1)–(3) are based on 41 studies. P-value corresponds to the proportion of times the observed ratio, or a higher ratio, of positive to negative results would be obtained in 41 indeptendent Bernouli trials in which positive and negative results were equally likely.

Columns 1–3 of Table 4.4 repeat this same exercise using Hanushek's tabulated results for expenditures per student. The first column uses Hanushek's method, which weights studies by the number of estimates he extracted from them. It is more than twice as likely to find an estimate that show a positive effect of increased expenditures per pupil. The second column equally weights each study. Here, studies that find a positive effect of school spending outnumber those that find a negative effect by nearly four to one. In Column 3, where each study is weighted by the number of times it is cited, the probability of observing at least

[11] Also note that if the number of citations to each particular article is used as the weight – which has the advantage of including articles published outside of journals – the results are quite similar.

this many studies with positive results by chance are about one in ten thousand. As a whole, we believe that when fairly summarized the literature does suggest that school resources matter.

It should be emphasized that the results reported in Tables 4.2 and 4.4 are all based on Hanushek's coding of the underlying studies. Although Hanushek (1997) tried to "collect information from all studies meeting" his selection criteria, he notes that, "Some judgment is required in selecting from among the alternative specifications." As mentioned, the selection and classification of estimates in several of the studies is open to question, and could in part account for the curious relationship between the number of estimates taken from a study and the study's findings.

4.3 A closer look at nine studies

Figure 4.1 indicates that the nine studies from which Hanushek extracted 123 estimates are critical for the conclusion that class size is unrelated to student achievement. In view of their importance for Hanushek's conclusion, Table 4.5 summarizes the analysis underlying these studies. Another reason for taking a closer look at these studies is that Hanushek (2000) defends his procedure of placing a disproportionate amount of weight on these studies by arguing that they are of higher quality than the studies from which he extracted relatively few estimates.

***Table 4.5*. Summary of the 9 studies from which 8 or more estimates were extracted**

Study	Description	Hanushek Coding of Class Size Results	Comments
Burkhead (1967)	Stepwise regressions estimated using 3 school-level data sets. Chicago sample is 39 high-school-level observations; dependent variables are 11th grade IQ scores (proportion in stanine 5–9), 11th grade reading scores (proportion in stanine 5–9), residuals of reading and IQ scores from a regression on 9th grade IQ scores, high school dropout rate, and post-high school intentions; independent variables are teacher man-years per pupil, median family income, school enrollment, drop out rates, and 8 other variables. Atlanta sample is 22 high-school-level observations; dependent variables are median 10th-grade verbal achievement test score, residual of 10th-grade verbal score from a regression on the 8th grade IQ score, male dropout rate, and percent enrolled in s chool year after graduation; independent variables include pupils per teacher, expenditures per pupil, teacher pay, median income, and 4 other variables. Sample of 176 high schools from Projekt Talent; dependent variables are average 12th grade reading sco. College attendance rate, and residuals of 12th grade reading scores from a regression on 10th grade scores; explanatory variables included class size, expenditures per student, enrollment, beginning teacher salary, and median income.	11 neg & insig 3 pos & insig	It is unclear how the stepwise procedure was implemented. In many of the final models, none of the independent variables were statistically significant. More parameters are estimated than data points. Effects of pupil-teacher ratio, expenditures per pupil and teacher pay are difficult to separately identify. IQ is supposed to be invariant to environmental factors, so it is an unusual outcome variable. Half of the class-size coefficients in the final models indicate a positive effect of smaller classes; it is unclear how Hanushek coded only 3 as positive. The average standardized effect size is a positive effect of smaller classes.

Study	Description	Hanushek Coding of Class Size Results	Comments
Fowler and Walberg (1991)	Uses a backward stepwise regression procedure in which all explanatory variables are initially entered in the equation and then variables were dropped one by one until only the statistically significant ones remained. 18 deptendent variables were used, ranging from math and reading tests to percent of students constructively employed, and 23 independent variables were used including pupil-teacher ratio, expenditures per student, teacher salary and school size. Sample consists of 199 to 276 NJ high schools in 1985. Some variables are measured at the district level.	1 neg & sig 1 pos & sig 7 unknown & insig	Effect of pupil-teacher ratio is difficult to interpret conditional on expenditures per pupil. Pupil-teacher ratio is included in only 4 of the final 18 models reported. It is unclear how Hanushek selected 9 estimates. Many of the dependent variables are highly related; for example, average math score, percent passing the math exam, and the percent passing both the math and reading exam are used as the dependent variable in separate equations, as are math and reading scores from the Minimum Basic Skills Test and High School Proficiency Test.
Jencks and Brown (1975)	Uses sample of students from 98 high schools from Project Talent data to estimate a two steg model. In first step, high school fixed effects are estimated from a regression that controls for students´ 9th grade characteristics and test scores. In the second step, high school effects are related to class size, expenditures per student, and other scool inputs, as well as mean post-high-school education plans in 9th grade and average SES. Sample size in second step estimation ranges from 49 to 95. Dependent variables are 2 measures of educational attainment (reported 15 months or 63 months after high school), career plans (by sex); occupation (by sex); and vocabulary, social studies, reading and math tests.	3 neg & sig 3 neg & insig 4 unklown & insig	The sample only consists of those who were continuously in high school between 9th and 12th grade. Thus, high school dropouts are truncated from the sample, so any effect of high school characteristics on high school drop out behavior, and related career implications, is missed. Based on the results in Table 9, the four estimates Hanushek classified as unknown signs all have positive effects of smaller classes on test scores.

Study	Description	Hanushek Coding of Class Size Results	Comments
Cohn, Millman and Chew (1975)	Sample consists of 53 Pennsylvania secondary schools from 1972. Eleven goals (test scores, citizenship, health habits, creative potential, etc.) are the outcome variables; exogenous explanatory variables are selected from 31 variables, including class size, instructional personnel per pupil, student-faculty ratio, and average daily attendance. Outputs are measured att 11th-grade level, inputs are measured at the district, school, or 11th-grade level. Stepwise regression is used to select the initial specifications; outcome variables were considered endogenous deteminants of other outcomes if there was a high correlation between them and if ”an a priori argument could support their inclusion in the model”. Two stage least squares, reduce form, and OLS estimates are reported. Instrumental variables are all excluded variables.	1 neg & sig 9 neg & insig 1 pos & insig	Hanushek appears to have selected the OLS model results, which are the weakest for class size. The reduced form estimates indicate 8 positive effects of smaller classes and 3 negative ones, all of which are insignficant. The simultaneous equation models indicate 3 positive and 3 negative coefficients, all of which are insignificant. Procedures to select exogenous explanatory variables, endogenous variables, and exclusion restrictions are open to question.
Link and Mulligan (1986)	Separate OLS regression models for math and reading scores were estimated for 3rd, 4th, 5th and 6th graders, by white, black and Hispanic background, yielding 24 repressions. Explanatory variables are pretest score, interaction between large class (26 or more) and majority-below-average classmates, dummy indicating whether teacher says student needs compensatory education, mother´s education, weekly instructional hours, sex, teacher experience. Student is unit of observation. Sample drawn from Sustaining Effects data set. Median sample size is 237 students.	24 unknown & insig	Models reported include interaction between large class size and peer effects but not class size main effect. The text states that when class size was included as a main effect in the math equations it was not individually statistically significant; no joint test of the class-size-peer group interaction and main effect is reported. The interactions generally indicate that students with weak peers do better in smaller classes. No mention of the main effect of class size in the reading equations is reported, so it is unclear how Hanushek could classify 24 estimates as insignificant. The class-size-peer-group interactions generally indicate that students in classes with low achievers do better in smaller classes.

Study	Description	Hanushek Coding of Class Size Results	Comments
Link and Mulligan (1991)	Separate OLS regression models for math and reading scores were estimated for 3rd, 4th, 5th and 6th graders, by white, black and Hispanic background, yielding 24 regressions. Explanatory variables are pretest score, class size, a dummy indicating whether teacher says student needs compensatory education, weekly instructional hours, sex, same race percentage of classmates, racial busing percentage, mean pre-test score of classmates, standard deviation of pre-test score of classmates. Student is unit of observation. Sample drawn from Sustaining Effects data set. Median sample size is 3,300.	3 neg & sig 8 neg &insig 5 pos &sig 8 pos &insig	No family background variables except race. Standard errors do not correct for correlated effects within classes. Compensatory education variable is potentially endogenous.
Maynard and Crawford (1976)	Study designed to look at effect of family income on children´s outcomes. Data from Rural Income Maitenance Experiment in IA och NC. Dependent variables are days absent (grade 2–9 or 9–12), comportment grade point average, academic GPA (grade 2–9 or 9–12), and standardized achievement tests (deviation from grade equivalents scores or percentile ranks). More than 50 explanatory variables, including expenditures per student (IA), enrollment, log enrollment per teacher, income, log average daily attendance relative to enrollments, average test score for student´s grade and school (NC), remedial program, etc. Student is unit of observation. Estimates equations separately for each state.	2 neg & sig 3 neg & insig 2 pos & sig 4 pos & insig	Class size is just an ancillary variable in a kitchen-sink regression designed to look at the effect of random assignment to an income maintenance plan. Class size effects are difficult to interpret once expenditure per student is held constant. Many of the explanatory variables (e.g., average class performance and attemdamce relative to enrollment) further cloud interpretation of class size effects.

Study	Description	Hanushek Coding of Class Size Results	Comments
Sengupta and Sfeir (1986)	Sample contains 50 school-level observations on 6th graders in California. Dependent variables are math, reading, writing and spelling test scores. Explanatory variables are average teacher salary, average class size, percent minority, and interaction between percent minority and class size. Another set of 4 models also controls for nonteaching expenditures per pupil. Estimates translog production functions by LAD.	7 neg & sig 1 neg & insig	No controls for family background other than percent minority. It is unclear why the specifications are sufficiently different to justify taking 8 as opposed to 4 estimates. In all 8 equations, interactions between class size and percent minority indicate that smaller classes have a beneficial effect at the average percent minority, but only the class size main effect is used.
Stern (1989)	Uses school-level data from CA to regress test scores on average student characteristics, teachers per student, the square root of the number of students, and teacher pay. Math, reading, and writing tests are used in two school years, yielding 12 estimates. Median sample size is 2,360 students.	9 neg & sig 3 post & insig	The 9 equations that yield negative effects of teachers per student in a grade level also control for the number of students in the grade level; the 3 positive estimates exclude this variable. More students in a grade level have a strong, adverse effect on scores. If the teacher-pupil ratio has a nonlinear effect, the number of students in a grade level could be picking it up. In addition, variability in class size in this paper is not due to shocks in enrollment, which many analysts try to use in estimating class size effects.

Table 4.5 describes the analysis in each of these nine studies, summarizes their findings, and comments on their econometric specifications. For a variety of reasons, many of these papers provide less than compelling evidence on class-size effects. For example, Jencks and Brown (1975) analyze the effect of high school characteristics on students' educational attainment, but their sample is necessarily restricted to individuals who were continuously enrolled in high school between 9th and 12th grade. Thus, any effect of class size on high school dropout behavior – a key determinant of educational attainment – is missed in this sample.

At least a dozen of the studies in the full sample, and one third of the studies in Table 4.5, estimated regression models that included expenditures per pupil and teachers per pupil as separate regressors in the same equation. Sometimes this was the case because stepwise regressions were estimated (e.g., Fowler and Walberg, 1991), and other times it was a deliberate specification choice (e.g,. Maynard and Crawford, 1976). In either case, the interpretation of the class-size variable in these equations is problematic. In such a specification, School A can only have a smaller class than School B by paying its teachers less or crimping on other resources – which is not the policy experiment most people have in mind when they think about reducing class size.

Some of the samples used in the nine studies are extremely small. For example, four of Burkhead's estimates use a sample of 22 schools, with only 12 degrees of freedom. Hanushek (2000) argues that the sample sizes are unrelated to the number of estimates he extracted from a study, but his comparison does not adjust for the fact that the unit of observation also varies across the estimates. Studies from which few estimates were extracted tend to analyze more highly aggregated data. Analyses of more highly aggregated data would be expected to have lower sampling variance, because residual variability is averaged out. After adjusting for the level of aggregation, the sample size of an estimate is inversely related to the number of estimates extracted from the study in Hanushek's sample.

In some cases, multiple estimates were selected from papers that used the same data to estimate different specifications, although the specifications were not particularly different (e.g., Sengupta and Sfeir, 1986). In other cases, multiple estimates were selected from models that used different dependent variables, even though the dependent variables were highly related (e.g., Fowler and

Walberg, 1981). Another problem in the selection of some estimates is that studies occasionally included class size and an interaction between class size with percent minority or other variables. Only the class size main effect was selected, although in many of these cases smaller class sizes had a positive effect for students at the mean level of the interacted variable (e.g., Sengupta and Sfeir, 1986).

The imprecision of the estimates in many of the papers also presents a problem. For example, the confidence interval for the change in math scores associated with a reduction in class size from 22 to 15 students in Sengupta and Sfeir (1986) runs from -.06 to .97 standard deviations. This is wide enough to admit a large positive effect or a small negative one.

The review of the studies in Table 4.5 is not meant as a criticism of the contributions of these studies. Many are excellent studies. But problems arise in Hanushek's use of many of the estimates he extracted from these studies because, in many cases, the studies were not designed to examine the effect of class size, *per se,* but some other feature of the education process. Maynard and Crawford, for example, were interested in the effect of exogenous shifts in family income (arising from the Rural Income Maintenance Experiment) on children's academic outcomes, and the study provides persuasive results on this issue; class size and expenditures per pupil were just ancillary variables that the researchers held constant. Indeed, some of the authors (e.g., Jencks and Brown) cautioned against interpreting their class-size variables because of weaknesses in their data or analysis.

It is hard to argue that these nine studies deserve 123 times as much weight as Summers and Wolfe's (1977) AER paper. Indeed, given the discretion used to select the estimates described previously, it would seem to us to be much more sensible to put equal weight on all of the studies, than to weight them by the number of estimates Hanushek extracted.

4.4 Summing up

In response to work by Hedges, Laine and Greenwald (1994), Hanushek (1996b; p. 69) argued that, “Unless one weights it in specific and peculiar ways, the evidence from the combined studies of resource usage provides the answer” that resources are unrelated to academic achievement, on average. Since Hanushek’s results are produced by implicitly weighting the studies by the *number* of “separate” estimates they present (or more precisely, the number of estimates he extracted from the studies), it seems to us that the opposite conclusion is more accurate: Unless one weights the studies of school resources in peculiar ways, the *average study* tends to find that more resources are associated with greater student achievement. This conclusion does not, of course, mean that reducing class size is necessarily worth the additional investment, or that class size reductions benefit all students equally. These questions require knowledge of the strength of the relationships between class size and economic and social benefits, knowledge of how these relationships vary across groups of students, and information on the cost of class size reduction. These issues are taken up in later sections. But the results of this reanalysis of Hanushek’s literature summary should give pause to those who argue that radical changes in public school incentives are required because schooling inputs are unrelated to schooling outputs. When the study is the unit of observation, Hanushek’s coding of the literature suggests that class size is a determinant of student achievement, at least on average.

5 The Tennessee Class Size Experiment

5.1 Experimental versus non-experimental studies

Even though we conclude from the previous section that, for the average study, smaller classes do affect achievement positively; it is worth emphasizing some problems in drawing causal inferences from non-experimental studies. At least three problems are likely to occur; the omission of relevant control variables, the inclusion of variables that themselves are determined by class size and reverse causality. When we estimate the effect of a variable (such as class size) on an outcome variable (such as test scores), we would like to control for all factors affecting both these variables. It is likely that both the present and the entire history of family background factors and schooling characteristics will contribute to achievement in a given year. It is unlikely though that we would ever get data on all these factors, or even get to learn about what all these factors is. Additionally, we also run the risk of including variables that itself are affected by class size. This is for instance likely to be the case with some schooling variables, other then class size. An example already mentioned in the previous section, is the inclusion of expenditure per pupil. However, this might also be the case with variables indicating teacher quality. It is also likely that an additional problem will be reverse causality, by which is meant that the outcome variable will in part determine the effect variable, instead of the other way around. This is the case if local or central governments will assign more school resources to schools with low achieving pupils, and/or if school authorities assign the weakest pupils to the smallest classes.

An advantage with an ideal class size experiment is that pupils and teachers are randomly allocated to classes of different sizes. This means that class sizes would be independent of the regression error, which includes everything else that is affecting achievement. Intuitively, a class size experiment breaks the correlation with the size of the class and anything systematic. We therefore next turn to evidence from the only large-scale class size experiment ever conducted, the Tennessee STAR (Student/Teacher Achievement Ratio) experiment.

5.2 Project STAR

Project STAR was an experiment in which an eventual 11,600 students in their first four years of school (from kindergarten until 3rd grade) were randomly assigned to a small class (target of 13–17 students), regular-size class (target of 22-25 students), or regular-size class with a teacher aide within 79 Tennessee public schools.[12] Teachers were also randomly assigned to class types. The experiment began with the wave of students who entered kindergarten in the 1985–86 school year. Students who entered a participating school while this cohort was in first, second, or third grades were added to the experiment and randomly assigned to a class type. After four years, all students were returned to regular-size classes. Students were supposed to stay in their original class-assignment type for four years, although students were randomly re-assigned between regular and regular/aide classes in first grade.[13] Students who moved along on pace graduated from high school in the Spring of 1998. Mosteller (1995) described Project STAR as ”a controlled experiment which is one of the most important educational investigations ever carried out and illustrates the kind and magnitude of research needed in the field of education to strengthen schools.”

5.3 Sample and population characteristics

Schools were selected to participate in the STAR experiment if they met certain requirements (e.g., sufficient enrolment and geographic criteria), and volunteered to participate. As a consequence, the 79 participating elementary schools were not a random sample of Tennessee elementary schools. To be eligible for the experiment, a school had to be large enough to have at least three

[12] The experiment is described in extensive detail in Word, Johnston, Bain, et al. (1990), Folger and Breda (1989), Finn and Achilles (1990), Krueger (1999) and Achilles (1999).

[13] In addition, about 10 percent of students switched between class types for other reasons. Krueger (1999) examines the impact of these transitions on the experiment, and finds that they have relatively little effect on the main results.

classes per grade so students could be assigned to a small, regular, or regular with teacher's aide class within each school. Furthermore, the state legislature mandated that the sample consist of a specified fraction of schools from inner-city, suburban, urban and rural areas, which led to participation of a higher proportion of inner-city schools than the overall state proportion. To assess how Project STAR schools compare to all schools in Tennessee and in the United States, we present selected characteristics of schools in Table 5.1.

Table 5.1. **Selected population characteristics**

	STAR (1)	Tennesse (2)	United States (3)
Percent minority students	33.1	23.5	31.0
Percent black students	31.7	22.6	16.1
Percent of children below poverty level	24.4	20.7	18.0
Percent of teachers with master´s degree or higher	43.4	48.0	47.3
Average ACT score	19.2	19.8	21.0
Average 3rd grade enrollment across schools	89.1	69.5	67.1
Average current expenditures per student across schools	$3,423	$3.425	$4,477

Notes: With the following exceptions, data are from the 1990 Common Core of Data (CCD) from the Department of Education. For comparability, the Project STAR characteristics were calculated from the CCD. (Nevertheless, the characteristics were very similar when calculated directly from the Project STAR data.) Teacher education data are for 3rd grade teachers from Project STAR data, and for 1993–94 public elementary and secondary school teachers from the Digest of Education Statistics. Statistics on poverty and racial background for the United States are from Census Bureau. ACT scores for Tennessee and United States are from ACT, Inc.

Project STAR schools have a larger minority population than do schools in Tennessee overall, but have a proportion similar to the national average. But most minority students in the STAR experiment are black – only a small fraction of students are Hispanic or Asian – so the proportion of black students in the participating schools is nearly twice the national average. STAR schools are also located in areas with somewhat higher child poverty rates, and teachers are slightly less likely to have completed more than a bachelor's degree. Average student performance as measured by ACT scores is slightly worse for STAR students than for all Tennessee students, and Tennessee performs worse than the nation as a whole.[14]

Since schools in the experiment were required to have at least three classes per grade, the STAR schools are larger than the average school. Average 3rd grade enrolment in Tennessee schools is about 70, whereas STAR schools had almost 90 students per grade – equal to the 72nd percentile of 3rd grade enrolment state-wide. Average current expenditures per student in 1990 were virtually identical in the STAR and Tennessee sample at about $3,425. Per-pupil spending levels in Tennessee were only about three-quarters of the national average.

Most schools in the STAR experiment consisted of students in kindergarten (the typical first year of school) through sixth grade. The average kindergarten student in the experiment was 5.4 years old at the beginning of his or her first school year. Kindergarten attendance was not mandatory in Tennessee when the STAR experiment began, so some students started school in first grade. In addition, some students repeat a year of school (e.g., they are retained in the same grade level), especially in the early years, so additional students joined the wave of students going through the experiment in first, second, and third grade. New students in participating schools were randomly assigned to a class each year. After attending elementary school, students typically attend middle school and then high school. Students graduate from high school after successfully completing 12 years of school beyond the kindergarten level. Most students are 17-18 years old by the time they finish high school.

5.4 A look at random assignment

A limitation of the design of the STAR experiment is that students were not systematically tested prior to entering a small class (see Krueger, 1999; Hanushek, 1999). Random assignment would be expected to produce groups of students that did not differ on average among the three assignment groups, conditional on school and entry grade. If data were available, one could test for significant differences in mean student achievement scores across class types. Nonetheless, if random assignment was implemented correctly, observable characteristics of students and teachers should be similar across class types. This is examined in Panel A of

[14] ACT is a US national college entrance exam.

Table 5.2, which presents a linear regression of student class-type assignment on demographic characteristics.[15]

Table 5.2. **Examination of random assignment, linear probability models**

	A: Students				B: Teachers			
Explanatory variable	Means (SD)	(1)	(2)	(3)	Means (SD)	(4)	(5)	(6)
Intercept	–	0.255 (0.020)	0.311 (0.014)	0.278 (0.014)	–	0.461 (0.131)	0.446 (0.151)	0.463 (0.172)
White/Asian (1=yes)	0.631 (0.483)	0.025 (0.010)	-0.006 (0.016)	-0.011 (0.016)	0.814 (0.389)	0.006 (0.035)	-0.017 (0.043)	-0.032 (0.053)
Female (1=yes)	0.471 (0.499)	0.001 (0.008)	-0.003 (0.008)	0.000 (0.008)	0.988 (0.109)	-0.057 (0.126)	-0.015 (0.140)	-0.011 (0.164)
Free lunch (1=yes)	0.547 (0.489)	-0.018 (0.009)	-0.008 (0.010)	-0.016 (0.010)	–	–	–	–
Master´s degree or higher (1=yes)	–	–	–	–	0.376 (0.485)	-0.047 (0.028)	-0.059 (0.031)	-0.069 (0.037)
Total experience	–	–	–	–	12.027 (8.323)	0.000 (0.002)	-0.000 (0.002)	-0.001 (0.002)
Entry-grade fixed effects	–	No	Yes	No	–	No	Yes	No
School fixed effects	–	No	Yes	No	–	No	Yes	No
School-by-entry-wave fixed effects	–	No	No	Yes	–	No	No	Yes
R-squared	–	0.00	0.03	0.08	–	0.00	0.02	0.04
P-value of significance of Explanatory variables	–	0.000	0.837	0.450	–	0.560	0.392	0.380

Note: White standard errors in parentheses. The free lunch variable measures whether a student was on free or reduced-price lunch during his or her entry year. For columns 1–3, the mean dependent variable is 0.26 and sample size is 11,294. For columns 4–6, the mean dependent variable is 0.39 and sample size is 1,330. For teachers, entry-grade and entry-wave are the grade level they taught. Entry-grade fixed effects are three dummy variables indicating the grade the student first entered the programme.

[15] Although one may object to the use of a linear probability model in this instance (e.g., as opposed to a logit), because the class-type variable is an independent variable in the models that follow, and we are simply interested in whether class-type and personal characteristics are related, the linear model provides appropriate estimates.

The dependent variable is a dummy variable that equals one if the student initially attended a small class, and zero if he or she initially attended a regular or regular/aide class.[16] Each student appears in the sample once, in the year he or she initially joined the experiment.[17] Column 1 only controls for three explanatory variables: racial background, sex, and free-lunch status.[18] Column 2 additionally controls for 78 school fixed effects. Strictly speaking, class-type was randomly assigned within schools for each grade (or entry wave) that the students entered the experiment. Thus, in column 3 we control for 304 school-by-entry-wave dummy variables. When school fixed effects or school-by-entry-wave fixed effects are controlled for, none of the student characteristics predict small-class assignment for the STAR sample (see columns 2 and 3). This finding is consistent with the students being randomly assigned to class types.

An important feature of the STAR experiment is that classroom teachers were also randomly assigned to class types within each participating school. If random assignment of teachers was properly executed, one would not expect a teacher's characteristics to be related to whether or not she taught a small class. Panel B of Table 5.2 reports results from a linear regression of teachers' class assignments on their demographic characteristics, using the sample of 1,330 teachers pooled across all grade levels. The dependent variable equals one if the teacher was in a small class, and zero if she was in a regular or regular/aide class. The results indicate that teachers' education, experience, racial background and gender are essentially uncorrelated with the class type to which they were assigned. Moreover, this result holds irrespective of whether school effects or school-by-grade-level effects are held constant.

Table 5.2 highlights the importance of controlling for school fixed effects, since random assignment of teachers and students was performed within schools. Moreover, students were randomly assigned within schools *in the grade they initially entered Project STAR*, which suggests that it is desirable to control for school-by-

[16] Unfortunately, we do not know which class type students were initially assigned to, as opposed to the class type they initially attended. However, for a subsample of 18 STAR schools, Krueger (1999) finds that 99.7 percent of kindergarten students attended the class type they were randomly assigned to their first year in the experiment. Consequently, henceforth we treat initial assignment and the initial class the student attended interchangeably.

[17] Standard errors have been adjusted for heteroskedasticity that arises in the linear probability model using White standard errors.

[18] Free lunch status means that the pupils' parents have sufficiently low income for the school to guarantee the pupil a free lunch meal at school.

entry-grade effects as in column 3. Most previous analyses of the STAR data have estimated treatment effects controlling for school fixed effects, but not school-by-entry-wave fixed effects. In most of what follows, we control for dummy variables indicating the school students initially attended interacted with dummy variables indicating the grade they entered the experiment (i.e., entry wave).

5.5 Results for kindergarten – 8th grade

To assess the effect of being assigned to a small class on test scores in the STAR data, for each grade we estimate a regression of test scores on SMALL, a dummy variable that equals one if the student initially was assigned to a small class, and zero if he or she was assigned to a regular or regular/aide class. In the estimations we also add controls for students' sex and racial background, and whether the student ever received free or reduced-price lunch in grades K-3, and a set of school-by-entry-wave fixed effects (based on initial school attended). The base group for the small-class-size effect consists of students who were assigned to either regular or regular/aide classes.[19] It is important to stress that class-type is based on the class the student attended the initial year of the experiment, and does not vary over time. As a consequence, the coefficient estimates are not subject to bias because of possible non-random transitions after the initial assignment. Also note that the test scores are the average scores on the Math and Reading Stanford Achievement tests, each scaled in percentile ranks.

The estimations were done separately for the full sample, for students on free or reduced-price lunch, and for the subset of black students.

[19] For students who were present in grades K and 1, we tested this specification against a less restrictive one that differentiated the base group among those who were consistently in regular classes, those who were consistently in regular/aide classes, and those who switched between regular and regular/aide classes. This less restrictive specification typically performed no better than the one reported in the text.

Figure 5.1. **Small-class effect, K-8 in STAR**

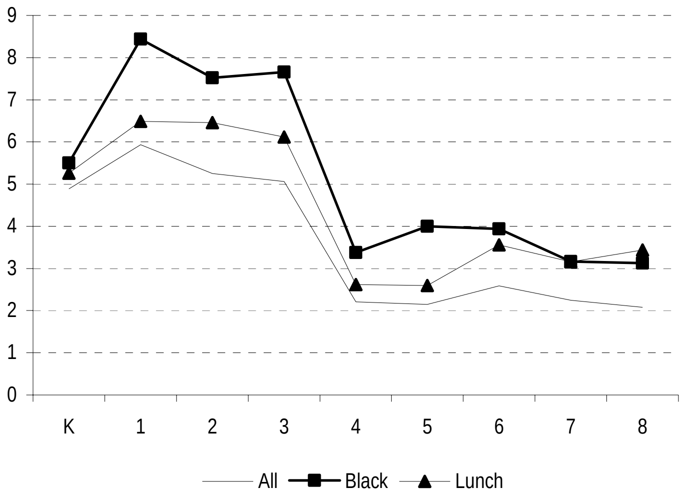


Figure 5.1 summarizes the coefficients on the SMALL dummy variable, using the largest sample of observations available for each group in each year. A 5 percentile-point gap opened up between students in small and regular-size classes by the end of kindergarten, and the gap stayed roughly constant in subsequent grades during the course of the experiment.[20] The small-class advantage was larger for the minority children and those on free lunch. Several studies have found that minority and disadvantaged students benefit more than other students from attending small classes (Summers and Wolfe, 1977; Hanushek, Kain and Rivkin, 1998).

In fourth grade, when the experiment ended and students returned to regular size classes, the effect size in terms of mean percentile ranks was reduced approximately to half to one quarter of its previous magnitude. From teacher reports, we have data on the actual class size for a subset of 520 fourth grade students. Interestingly, the average fourth grade class size for students who were initially assigned to regular size classes was about 0.36 (t = 2.4) student smaller than it was for students initially assigned to small classes, conditional on initial school fixed effects. It is possible that, to some extent, school principals attempted to compensate

[20] Previous work tends to find that the small class advantage expanded between kindergarten and first grade, but that appears to result from the omission of controls for school-by-entry-wave effects.

for the earlier effects of the experiment, which may partially account for the relative improvement of students who were previously in larger classes. In addition, peer effects could have raised the performance of students from regular classes relative to those from small classes after the experiment ended.

One important qualification should be kept in mind while considering changes in the magnitude of the small-class effect in Figure 5.1: the tests are scaled by percentile ranks. Test score percentile ranks are not a cardinal measure. It is possible, perhaps likely, that a given percentile gap implies a larger educational difference in the higher grades than in the lower grades. Indeed, Finn et al. (1999) present evidence that, when the Stanford Achievement Test and CTBS scores are scaled in terms of grade equivalents, the gap between students in small and regular-size classes expands from grade K to 3, and from grade 4 to 8.

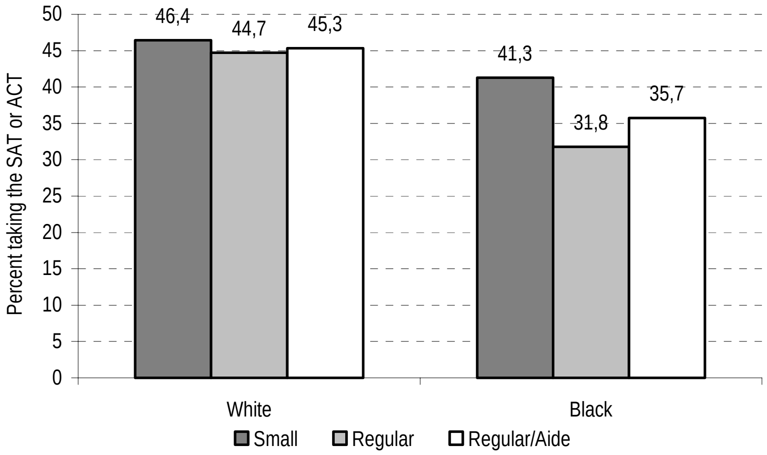
5.6 Results for college entrance exams

In their last or penultimate year of high school, students who intend to enrol in college take the ACT or SAT exam. These are privately administered exams that are required by most colleges for admission. In this section, we investigate whether the probability of taking these exams and the score on them where higher for students in smaller classes.

5.6.1 Test Taking rates

We first examine whether assignment to a small class influences the college-entrance exam test-taking rate. We again hold school effects constant.[21] Our findings are illustrated in Figure 5.2.

Figure 5.2. **Percent of students taking the SAT or ACT by initial class type**



Note: Means are balanced within school using the balanced sample estimator described in Krueger & Whitmore (2001).

This figure reports the percent of students by racial background who took either the SAT or ACT, by the type of class they attended during their first year in Project STAR. For white students, Figure 5.2 indicates that 46.4 percent of students initially attending small classes took a college-entrance exam, compared to 44.7 percent in regular classes and 45.3 percent in regular/aide classes. These differences in rates are not statistically significant. Black students were substantially more likely to take the SAT or ACT if they were assigned to a small rather than regular-size class: 41.3 percent of black students assigned to small classes took at

[21] However, this time we adjust for school effects using a balanced sample estimator (see Krueger & Whitmore (2001)).

least one of the college entrance exams, compared with 31.8 percent in regular classes and 35.7 percent in regular/aide classes. The chance of such a large difference in test-taking rates between the small and regular class students occurring by chance is less than one in 10,000.

To interpret the magnitude of these effects, note that the black-white gap in taking a college entrance exam was 12.9 percentage points for students in regular-size classes, and 5.1 percentage points for students in small classes. Thus, assigning all students to a small class is estimated to reduce the black-white gap in the test-taking rate by an impressive 60 percent.

Findings in Krueger & Whitmore (2001) suggests that small classes matter for black students because of something having to do with the schools they attend, rather than something inherent to individual black students per se. For example, it is possible that black students attend schools that have a disproportionately high number of disruptive students, or students with special needs, which distracts their teachers from instructional time.[22] In this case, white students in those schools would also benefit from smaller classes.

5.6.2 Test scores on college entrance exams

Next, we examined the scores the students attained on the ACT and SAT exams. For students who took the SAT but not the ACT exam, we converted their SAT score to an ACT-equivalent score.[23] For any student who wrote the ACT exam we used the ACT score even if he or she also took the SAT. For students who took an exam more than once we used the first score. Naturally, any analysis of ACT and SAT scores can only be performed on the subset of students who took one of the exams. This creates a potential selection problem. Because a higher proportion of students from small classes took the SAT or ACT exam, it is likely that the group from small classes contains a higher fraction of relatively weak students. That is, stronger students are likely to take an exam regardless of their class assignment, but marginal

[22] See Lazear (1999) for a formal economic model that predicts that smaller classes lead to higher achievement by reducing the number of disruptions in a class.

[23] This was done using a concordance developed jointly by ACT and the College Board (see www.collegeboard.org for the concordance). 121 students – or 2.6 percent of the test-taking sample – took the SAT and not the ACT. For the 378 students in our sample who took both tests, the correlation between their SAT and ACT scores is 0.89.

students who are induced to take the exam because they attended a small class are likely to be lower-scoring students. Such a selection process would bias downward the effect of attending a small class on average test scores. The bias is also likely to be greater for black students, because a higher share of black students were induced to take the exam as a result of attending a small class.

To simplify the analysis, we compare students who initially attended small classes to the combined sample of those who initially attended either regular or regular/aide classes, and we control for school effects instead of school-by-entry-wave effects. Also, because we later implement a Heckman (1976) selection correction, we use raw ACT scores instead of percentile ranks for this analysis. The raw ACT scores in our sample range from 9 to 36 and are approximately normally distributed.

Table 5.3 **Effect of class size on ACT or SAT score with and without selection correction**

	White Students			Black Students		
	No correction	Heckman correction	Linear truncation	No correction	Heckman correction	Linear truncation
Explanatory Variable	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	20.233 (0.138)	16.386 (0.524)	20.242 (0.138)	17.073 (0.275)	7.443 (3.610)	17.164 (0.274)
Small class	0.009 (0.169)	0.209 (0.210)	0.206 (0.167)	0.213 (0.204)	0.834 (0.266)	1.079 (0.203)
Female (1=yes)	0.056 (0.156)	1.787 (0.197)	0.021 (0.156)	0.522 (0.190)	2.229 (0.237)	0.378 (0.191)
Free lunch (1=yes)	-1.434 (0.180)	-4.859 (0.241)	-1.385 (0.179)	-1.715 (0.265)	-3.529 (0.332)	-1.725 (0.263)
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	3198	7124	3173	1427	4117	1357
Effect size	0.002	0.039	0.038	0.039	0.153	0.198

Note: Heteroskedasticity-adjusted standard errors are reported in parentheses for columns (1), (3), (4) and (6). If a student took only the SAT, that score is converted to its comparable ACT score (see text for details). The mean (standard deviation) of the dependent variable in column (1) is 19.9 (4.5), 19.9 (4.4) in column (3), 16.1 (3.5) in column (4), and 16.3 (3.5) in column (6). The effect size is the coefficient on small divided by the standard deviation of test scores among the full sample of students (5.4).

The results are reported in Table 5.3. For the sample of test takers, the average ACT score was virtually identical for students who were assigned to small and normal-size classes. The average white student in a small class scored 19.88, while the average white student in a regular class scored 19.87. Black students in small classes averaged 16.3, while black students in regular classes scored 16.1. The differences between small and regular classes are not statistically significant.

Past studies of state-level data have shown that average test scores tend to decline when more students take a college entrance exam, most likely because the marginal test takers are weaker students than the average student (see, e.g., Card and Payne, 1998). In Project STAR, there were two confounding effects: selection and treatment. One might expect the treatment to result in small-class students scoring slightly higher on the ACT, as they did on previous tests through the 8th grade. But since a larger percentage of students assigned to small classes took the exam, a larger share of weaker students in small classes likely took the test. As a result, it is difficult to interpret the score results because scores are only reported conditional on taking the exam, and the treatment appears to have affected the likelihood of taking the exam – particularly for black students. Columns (2) and (3), and (5) and (6) present two types of estimation results that attempt to adjust for this sample selection problem.

Columns (2) and (4) present results of a standard Heckman-correction procedure for white and black students. Identification in this model is based solely on the assumption of normal errors, as there is no exclusion restriction. We also calculate the “effect size” by dividing the coefficient on the small-class dummy by the standard deviation of ACT scores among all students who took the exam (equal to 5.4). The Heckman correction doubles the point estimate on the effect of attending a small class for white students, but the coefficient is still statistically insignificant and qualitatively small. For blacks, however, column (5) indicates that after adjusting for selection, students in small classes score 0.15 standard deviation higher than those in regular classes.

In columns (3) and (6) we present results from a different approach for adjusting for selection (see Krueger & Whitmore (2001)). Here we have artificially truncated the sample of students from small classes so that the same proportion of students from small and regular-size classes is represented in the test-taking sample. Specifically, we drop from the sample the bottom X percent of students based on their test results, where X is determined so that the share of students from small classes who took the exam equals the share from regular-size classes. This approach is valid if all the additional small-class students induced to take the ACT are from the bottom of the distribution, and if attending a small class did not change the ranking of students in small classes. Although the former assumption is extreme, the results should provide an upper bound on the impact of selection bias, and are an interesting point of comparison to the Heckman-correction results.[24]

The results in columns (3) and (6) are quite similar to the Heckman-correction results in columns (2) and (5). For white students, the linear truncation and Heckman-selection-correction procedure indicate that students in small classes score insignificantly differently from students in normal size classes, with a point estimate corresponding to a 0.04 standard deviation. For black students, the linear-truncation procedure yields and effect size of 0.20 standard deviations, somewhat larger than the 0.15 effect size from the Heckman-correction procedure.

5.7 Summing up

The STAR experiment did raise pupils' achievements by about 5 percentile points in the early grades, and by about 2 percentile points in the later grades of primary schooling. The achievement gains from attending a small class were even higher for black and free-lunch students. Also college-test taking rates and scores on college entrance exams did increase on average. However, this appears to mainly be due to the effect for black students.

Note that we in this section have treated regular and regular/aid classes as being identical. The reason for this is that there appears to be no additional beneficial effect of having a teacher aide (see Krueger, 1999). This is an important result, and it shows that additional teachers per se is not enough to increase performance.

[24] In principle, the Heckman procedure provides an estimate of the effect of attending a small class on test scores for the entire population of students (including those who did not take the test), whereas the linear-truncation approach provides an estimate of the effect of attending a small class on scores for students from regular classes who otherwise would have taken the ACT. If there is a homogeneous treatment effect, the two parameters would be equal.

What appears to be important is to decrease actual class sizes, i.e the number of pupils taught together.

6 Old Swedish Evidence and New Swedish Data

6.1 Previous Swedish class size research

Surprisingly little class-size research has been conducted using Swedish data. Marklund (1962) used two samples of pupils: one from 1959 (Sweden) and one from 1955-1956 (Stockholm). For both samples, he divides the data into three equally sized groups with different class sizes. Controlling for intelligence, the mean test results in these groups are insignificantly different from each other, even between groups with similar social backgrounds. He concludes that: ”we do not suggest that size of class lacks relevance for the achievement of pupils. But it is safe to say that class size has no significant bearing on such achievement in today’s school.”

Lindsey & Cherkaoui (1975) do another study of this matter. They are using Swedish data, collected by IEA in 1962, on mathematics achievement for pupils that are 13 years of age. They fit a nonlinear (in test scores) regression model of mathematics achievement to class size, hours of instruction and sex. The derivative of class size on achievement is positive at the mean of the other variables. It should be noted, however, that they argue that causal inferences based on this result should not be drawn.

The lack of newer Swedish studies on class size and pupil achievement is surprising, given that this issue has been so hotly debated. What comes closest is a study by Skolverket (1999), which utilized data on 92,000 pupils from 900 schools in Sweden. They analysed the effect of weekly teaching hours per pupil on average marks in 9th grade for those pupils leaving basic schooling 1995. The analysis was conducted by regressing school-average marks as a linear function of weekly teaching hours per pupil in the school, conditional on parents education, fraction of boys, fraction of pupils with non-Swedish background and number of lower secondary school students. Note that all variables are measured at the school level. They did find a negative relationship between weekly teaching hours per pupil and average marks for the average school, but a positive relationship for schools with many low-educated parents. They also show that schools with many pupils with low educated parents are assigned more weekly teaching hours per pupil. Based on this compensatory distribution of resources (which would induce a downward biased estimate between weekly teaching hours and marks), they are quite careful in drawing causal inferences.

6.2 New data

Due to the mostly very old and little existing research on class size in Sweden, we collected fresh data on a sample of 556 pupils from 16 schools in Stockholm, Sweden's capital. The sample contains information on scores from an identical math test for the same pupils in the spring 5th grade and in the fall and spring of 6th grade. The sample also contains socio-economic measures from register data and data on pupils' demographics, their class size and teacher characteristics from a questionnaire answered by teachers. The data collection was made with the goal to be able to use more sophisticated methods then has previously been used for Sweden on the issue of the effect of class size on achievement.

6.2.1 Data collection

The Swedish National Agency for Education distributes a math test in the early spring semester of 5th grade to all schools in Sweden. We contacted a number of Stockholm schools at the start of the fall 6th grade semester in 1998. These schools were drawn randomly, with the restriction that two schools from the Stockholm area with the lowest socioeconomic level and two schools from the Stockholm area with the highest socioeconomic level should be included. The schools that declined to participate stated that they were not able to hand over the 5th grade tests because they felt they did not have the time or that they did not want to make the pupils "suffer" once again by making them take this test. Within the participating schools, very few classes chose not to participate.

The math test consisted of selected parts of the test, distributed by the Swedish National Agency for Education in the spring of

1998. We distributed the math test to the pupils at the start and end of 6th grade. The national math test in the spring of 1998 was administered without our participation. It had 10 separate parts of which six were to be done individually, two were to be done in pairs, and two were to be done by groups of pupils. Of the six individual tests, we picked four to be repeated by the pupils at the start and end of the sixth grade. The four test parts were estimated to take an equal length of time to complete. These test parts included math questions on the four fundamental rules of arithmetic, but also included, for instance, math related problems based on a story the pupils had to read. The answers, together with the tests, were handed over to us for grading.[25] One of us was present in most schools during some parts of the test in the fall and spring of 6th grade.

The test taken by the pupils at the start and end of 6th grade were to be done under as similar conditions as possible as the 5th grade test. The conditions, regarding time allowed and help given, varied among pupils within and among classes. We therefore instructed teachers that the 6th grade tests should be conducted under the same conditions as the fifth grade test. Because we in the main analysis use changes in test scores between test occasions, different test conditions between pupils at each test occasion should not be a serious problem. A total of 556 pupils took the test on all three occasions and did the test under similar conditions on all three occasions. [26]

Data on school, class, and teacher characteristics were taken from answers to a questionnaire, which we gave to the teachers at the time of the fall sixth grade test. Teachers were asked to provide information on class sizes, their own education and experience and their pupils' characteristics (gender and number of pupils in the class that had no Swedish parent). Besides the questionnaire, we asked the teachers to list the pupils with no Swedish parent, to elaborate on unclear answers and to state whether or not mathematics was taught in classes other than the regular classes. If so, the

[25] We compared points (based on corrected answers for 50 tests) that we graded with the same tests that an assistant graded. The correlation for all parts of the test was more than 0.92.

[26] Fewer pupils took all four parts of the test at all three test occasions, however. As long as the pupil's knowledge of the math skills tested in a particular test part is unrelated to whether or not this pupil took this test part, this is not a problem in our analysis. This is probably the case at the fall and spring of 6th grade test occasions, whether this also is the case for the first test occasion is not clear.

teacher would state which pupils belonged to which math class. Based on this, actual class sizes in mathematics were calculated.

To get information on the pupil's social background, the addresses of the pupils (from the class lists) were matched with block data on the education and incomes of the parents (from registers).[27] For the purpose of this project, Statistics Sweden calculated block-means (for income) and block-fractions (for educational levels). Education was given in seven categories from which average years of schooling within each block were calculated.[28] The income variable used here is mean family income.[29]

[27] Because the actual parents were not known, Statistics Sweden restricted the calculation on block-means and block-fractions to be for individuals of ages 28–54 and with children ages 10–12. Statistics Sweden required, for safety reasons, that there be at least three observations within each block. So if this was not fulfilled, We restricted the individuals to ages 28–54 with children.

[28] This was done by regressing years of schooling (from a survey questionnaire) on dummies for educational level achieved (from registers) using a sample with the same age restrictions from the 1991 Swedish Level of Living Survey. The estimates from this regression were then used to predict mean years of schooling within each block.

[29] For the 8 pupils with missing address information, their class' average years of schooling and family income were assigned. This was also done in the additional five cases for which family income data was missing.

6.2.2 Data description

Table 6.1. **Descriptive statistics**

	Mean	St. Dev	Min	Max
Test scores (percentile ranks)				
Fifth-grade, spring	47,65	23,07	1	99.5
Sixth-grad, fall	47.93	22.98	1	96.5
Sixth-grade, spring	46.80	22.70	1.5	92
Class size				
Class size, fifth grade (math)	22.91	5.72	3	32
Class size, sixth grade (math)	19.90	4.40	5.5	25
Teacher variables, sixth grade				
Teacher experience in years	16.17	10.82	0.2	33
Teacher exp. (years in the class)	1.62	1.04	0	5
Demographic and family background variables				
Gender (girl=1)	0.50	0.50	0	1
Non-Swedish parents=1	0.23	0.42	0	1
Log (family income)	12.60	0.54	11.19	14.75
Socioeconomic index	0.00	0.93	-2.51	2.52

Notes: Numer of observations is 556. Test scores pertain to the national mathematics tests.

Table 6.1 presents descriptive statistics for the sample used in this study. The minimum possible raw score is 0, and the maximum possible score is 18, for each of the four test parts. To facilitate interpretation and comparison with other studies, we transformed the scores on each test part to percentile ranks, i.e. from a number from 1 to 100. We then calculated the average of these transformed scores. So the test scores used in the estimations are the average of the percentile ranks of the scores on the four test parts. Correlations between the test scores at the different test occasions are between 0.72–0.77. The correlation's between the test parts at the first test occasion range between 0.43–0.56. Assuming that these test parts are independent the reliability ratio for the test is calculated to be 0.79.[30]

[30] This reliability ratio is calculated using Cronbach´s alpha (see Cronbach 1951).

The demographic and family background variables are *girl, non-Swedish parents, parents' years of education, and the logarithm of family income. Non-Swedish parent* is an indicator that equals one if the pupil lack parents with Swedish as their native language, and zero otherwise. Table 6.2 shows correlation matrix for these variables. The *socioeconomic index* is the average of the standardized values of parents' education and logarithm of family income.

The measure of class size used in this study is the number of pupils taught together in math. We also have information on a regular class size measure, which refers to the number of pupils taught together in the typical class. However, since the test that is used in the analysis in this paper reflects math knowledge, we use the math class size measure in our main analysis. In this sample there are a total of 38 math classes. The distribution of this variable is shown in Figure 6.1. Table 6.2 shows the correlations between the class size and the family background variables.

Figure 6.1. **Distribution of math class size in sixth grade**

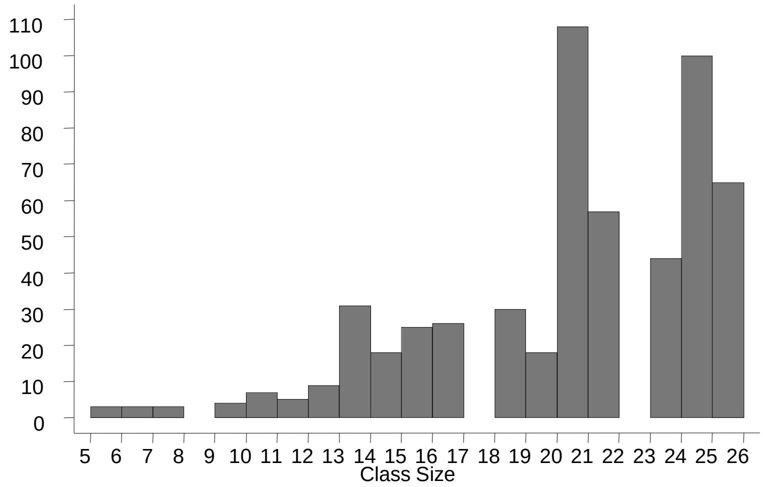


Table 6.2. **Correlation matrix for demographic, family background and class size variables**

	Girl	Non-Swedish parents	Parents´ education	Log (family income)	Class size, fifth grade (math)	Class size, sixth grade (math)
Girl	1.00					
Parents´ nationality (non-Swedish parents=1	-0.02 (0.55)					
Parents´ education	0.03 (0.55)	-0.50 (0.00)	1.00			
Log (family income)	0.01 (0.77)	-0.55 (0.00)	0.72 (0.00)	1.00		
Class size, fifth grade (math)	0.02 (0.69)	-0.40 (0.00)	0.37 (0.00)	0.34 (0.00)	1.00	
Class size, sixth grad (math)	-0.01 (0.86)	-0.47 (0.00)	0.51 (0.00)	0.47 (0.00)	0.60 (0.00)	1.00

Notes: Number of observations is 556. P-values for test of no correlation are in parentheses.

7 An Analysis of the Effect of Class Size in Sweden

The purpose of this analysis is to estimate the causal effect of class size on academic achievement. By estimating a causal effect we mean that we would like to know what achievement the mean pupil would have had, had that pupil gone to a class of different size. It is this estimate that is interesting when a policy is implemented. The preferable way to estimate this class size effect would be to randomly assign pupils to classes of different sizes. This was done in the experiment in Tennessee, which we analysed in section 5. Since no such experiment have been conducted, or has been possible to conduct, in Sweden, we need to use other methods.

The goal is to use a non-experimental method that will, as close as possible, mimic an actual experiment. In doing so we would like to compare pupils that are as equal as possible, both with respect to observable as well as unobservable characteristics. Our preferred method, compare pupils that are the same with respect to unobservable characteristics that affects the level of achievement and the rate of learning as well as with respect to observable previous achievement and observable characteristics in schools, such as the experience of teachers, and observable family background variables. This method of analysing this issue is a new one, and have the advantage, compared too much of previous non-experimental work, that more unobservable pupil characteristics can be controlled for.

7.1 Level and value-added regressions

We do however start our class size analysis by using more “naive” approaches to estimate class size effects. The simplest way to estimate class size effects are to simply associate pupils results on some achievement test with their class sizes. The problem is that pupils are not randomly assigned to classes of different sizes. At least two problems are present here. First, public resources are often distributed toward areas with schools were pupils score low on tests. That this is the case in Sweden has recently been shown in a report from the National Agency for Education (Skolverket, 1999, no. 170). The same is likely to also be true for resources within schools. Second, parents are likely to choose the schools for their children, were classes are the smallest. And the parents who make this choice more actively are likely to be those with high achieving children. The first problem makes the estimate from a simple association between pupils test scores and their class sizes to be biased toward finding a positive association between class size and achievement. The second problem will instead bias the estimate toward showing that a lower class size is associated with higher achievement. It is hard a priori to say which of these problems that dominates, but since school choice in Sweden has been quite restricted and the municipality government in Stockholm actively do re-distribute resources toward low income areas, it might be expected that the former effect dominate. If so, we would expect that simply associating pupils' test scores with their class sizes will give an estimate of a positive effect of larger classes on pupils' achievement, even though the true causal effect could have an opposite sign.

Table 7.1. **Level regressions**

	Dependent variable: Test score in spring of the sixth grade		
	OLS (1)	OLS (2)	OLS (3)
Class size, sixth grade (math)	1.51 (0.31)	0.81 (0.33)	0.84 (0.35)
Teacher experience			0.32 (0.25)
Teacher experience squared			-0.01 (0.01)
Teacher experience in class			-0.92 (0.96)
Additional controls	No	Yes	Yes
R^2	0.085	0.124	0.127

Notes: Number of observations is 556. The standard errors, in parentheses, allow for regression errors that are correlated among pupils in the same school. Test scores measured in percentile ranks. Additional controls are gender, non-Swedish parents, parents´ education, and the logarithm of family income.

In the first column of Table 7.1, test result in 6th grade is associated with class size in the same grade. This leads to a positive and highly statistically significant estimate. One way to improve the analysis somewhat, is to control for the effect of family background variables on class size and test scores. By this we mean that we aim to compare pupils with similar observable family backgrounds. More specifically we condition on pupils' gender and parents social background and whether both parents are non-Swedish. We still get a positive class size estimate, although it is smaller in magnitude (column 2 of Table 7.1). In column 3, we see that conditioning also on teacher variables leave the class size estimate basically unaffected.

The most common way of estimating class size effects is to associate pupils test scores with class size, but controlling for family background variables and previous test scores. This is the model way to estimate class size effects, according to Eric Hanushek. This approach is a bit more data-requiring then the level-model, since it requires data on test scores for the same pupils at two points in time. An advantage with this approach is that it allows us to compare pupils that are similar not only with respect to observable family background variables, but also with respect to the level of achievement attained in the beginning of the period at which test scores were measured. Also note that if previous achievement are controlled for, then, with reasonable assumptions, we control for all previous (unobservable as well as observable) school characteristics (including class size) and family background, which affect the level of achievement, up to that point in time. This method is in the literature called the value-added method. The name reflects that the dependent variable is the value of the test score that are added to the previous test score.

Table 7.2. **Value added regressions**

	Dependent variable: Test score changes from fifth grade, spring, to sixth grade, spring					
	OLS	OLS	OLS	OLS	OLS	OLS
	(1)	(2)	(3)	(4)	(5)	(6)
Class size	-0.01 (0.30)	0.11 (0.30)	0.16 (0.28)	0.17 (0.14)	0.19 (0.16)	0.23 (0.16)
Teacher experience			0.36 (0.38)			0.36 (0.23)
Teacher experience squared			-0.014 (0.011)			-0.013 (0.007)
Teacher experience in current class			-0.55 (1.01)			-0.59 (0.56)
Initial test score control	No	No	No	Yes	Yes	Yes
Additional controls	No	Yes	Yes	No	Yes	Yes
R^2	0.000	0.009	0.019	–	–	–

Notes: Number of observations is 556. In all columns, the standard errors, in parentheses, allow for regression errors that are correlated among pupils in the same school. The test score variable is measured in percentile ranks. In columns 4–6, when control for initial test score (spring of the fifth grade) is made, the estimates and standard errors assumes a true reliability ratio of 0.7878 in the fifth grade test score percentile ranks. Additional controls are girl, non-Swedish parents, parents´ education, and the logarithm of family income.

Table 7.2 shows that if we apply this method to our Stockholm data set we do not get any statistically significant effects of class size on achievement. This still holds if we conditioning for the same variable as we did in Table 7.1. The results in Table 7.2 are in line with much of the literature on class size, using the value-added method (see for instance Hanushek (1986, 1998).

7.2 Results using a new approach to estimating class size effects

There is at least one problem with the value-added model, which can make class size estimates being biased estimates of the true causal effects. If characteristics exists, that we cannot observe, and these characteristics are correlated both with the change in test scores and with class size, the class size estimate will be biased.

In order to control for these unobserved characteristics, we here present a new approach to estimate class size effects. The traditional value added model estimated the level of test scores in the present period as a function of class size in the present period and test scores in a previous period. We instead estimate the change of test scores during a present period, when schools are open, as a function of class size in the present period and the change of test scores during a previous period, when schools are closed. The intuition behind this approach is to use the test score changes during a period when schools are closed as a counterfactual situation to the test score changes when schools are open. In practice, we use the test scores during the summer vacation preceding the school year as the counterfactual to the test score changes during the school year. So we compare the value added to test scores during the school year (when class size varies among pupils), with the value added to test scores during the summer vacation (when class size are constant among pupils). This approach makes it possible to control for unobserved pupil characteristics that affect class sizes and achievement changes. For a more technical presentation of this model, see the technical appendix 1.

Table 7.3. **Difference-in-differences regressions**

	Dependent variable: Test score changes from sixth grade, fall, to sixth grade, spring minus test score changes from fifth grade, spring, to sixth grade, fall					
	OLS	OLS	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Class size, sixth grade	-0.77 (0.23)	-0.95 (0.32)	-0.98 (0.26)	-0.36 (0.14)	-0.38 (0.21)	-0.37 (0.18)
Teacher experience			0.77 (0.81)			0.55 (0.32)
Teacher experience squared			-0.019 (0.025)			-0.016 (0.010)
Teacher experience in current class			-1.04 (1.11)			-0.77 (0.50)
Initial test score change control	No	No	No	Yes	Yes	Yes
Additional controls	No	Yes	Yes	No	Yes	Yes
R^2	0.015	0.020	0.027			

Notes: Number of observations is 556. The standard errors, in parentheses, allow for regression errors that are correlated among pupils in the same school. The test score variable is measured in percentile ranks. In columns 4–6, when control for initial test score change (from spring of the fifth grade to fall of the sixth grade) is made, the estimates and standard errors assumes a true reliability ratio of 0.7878 in the fifth grade test score percentile ranks. Column 4–6 uses test score in the spring of the fifth grade as instrument for the initial test score change. Additional controls are girl, non Swedish parents, parents´ education, and the logarithm of family income.

The estimates using this model are presented in Table 7.3. Without (column 1) or with (columns 2–6) controls, we get a negative class size estimate, i.e. a positive effect of smaller classes on pupils' achievement. In the first three columns we regress the change in the value added during the school year test period minus the change in the value added during the summer test period, on class size. Controlling for the pupils' demographic and social background, and for the observable characteristics of the teacher (column 3), we get that one less pupil per class increases the test score for the average pupil with one percentile rank. It might however makes sense to take into account that the test score value added, both during the school year and summer, could be influenced by the test score in the beginning of these periods. In columns 4–6, where we take this into account, the class size estimates are still negative and statistically significant but smaller in magnitude, compared to columns 1–3.[31] One reason for this could be that previous test scores do affect the test score value added. However, separately regressing the school year and the summer period value added on previous test scores, suggests that this is true only to some degree, and that the estimates in column 4–6, probably underestimates the positive effect of smaller classes on achievement. The technical reason for this is that several econometric assumptions are required to be able to estimate the model underlying columns 4–6 consistently (for details see Lindahl (2002)). Hence, the effect of diminishing class sizes by one pupil, appears, on average, to raise achievement level somewhere between 0.4 and 1 percentile points in this data. This is quite similar as the results from the Tennessee experiment (see section 5).

These and other results are based on a sample of 556 pupils. However, the full sample of pupils is 701. The 145 pupils not included lack test score data on at least one of the three occasions. We do however have family background and schooling variables for the full sample. We therefore implemented a Heckman-correction (see Heckman, 1976) to see whether the results are sensitive to sample selection. However, the sample selection corrected class-size estimates were very similar compared to columns 1–3 of Table 7.3.[32]

Since the summer test period includes a substantial part of the school year, the previous regressions assumed that nothing was learnt during the beginning and end of the school year. We have done several adjustments to this. For instance, assuming linear learning during the school year, one can predict test scores at the exact beginning and end of the school year. Adjusting for the different period lengths, one can use this as the dependent variable

[31] In the estimations underlying Tables 7.2 and 7.3, we have estimated the reliability ratio to equal 0.7878, by estimating the so called Cronbach's alpha (see Cronbach, 1951). This is generally thought of as being a lower bound estimate of the reliability (see Reuterberg & Gustafsson, 1992). If the true reliability ratio is higher, the class size estimates in columns 4 –6 of Table 7.2 is downward biased. However, the class size estimates in Table 7.3 are basically uneffected by a higher reliability. This is because the IV-technique, where we use pre-summer test scores as an IV for the summer change in test scores, already to a high degree corrects for measurement error in the test scores.

[32] Since the Heckman correction rely on normality of the error term in the regression equation, we used the standardized values of the raw scores as dependent variable, instead of percentile ranks. We could then not reject that the type of dependent variable used in Table 7.3 was normally distributed (p-value 0.64 in a skewness-kurtosis test). The class size estimate (st.der.)in column 3 of Table 7.3., for instance, is now – .041 (.009) in standard deviation units. Correcting for sample selection. This estimate (st.dev.) is – .037 (.019).

in the regressions. Doing so actually reinforce the above mentioned results.[33]

Since we now have established a positive effect of smaller classes on achievement using new Swedish data, a natural continuation to this is to ask which pupils that benefit the most from smaller classes. This is done by adding interaction terms to the regressions, consisting of the class size variable, multiplied by the variables' Girl, Non-Swedish parents and the *socioeconomic index.* The results from this are shown in Table 7.4.

[33] Another potential problem could be unobserved schooling characteristics. However, when we did control for school fixed effects, the result that smaller classes generate higher achievement was strengthened.We also investigated whether the class size estimates were sensitive to outliers. Other potential problems are that the same test was used at all three occations, that class size in earlier grades might affect learning during the summer vacation, that the fixed learning effect could be different between summer and school year for each pupil and a correlation between the time varying error term and class size conditional on the fixed learning effect could exist. All these issues are investigated in Lindahl (2002), where the conclusion is that the estimated positive effect of smaller classes on achievement either is unaffected or understated in face of these issues.

Table 7.4. **Difference-in-differences regressions, allowing for heterogeneity in the effect of math class size on achievement**

	Dependent variable: test score changes from sixth grade, fall, to sixth grade, spring minus test score changes from fifth grade, spring, to sixth grade, fall	
	OLS	IV
	(1)	(2)
Class size	-0.23 (0.44)	0.08 (0.27)
Class size * girl	-0.60 (0.55)	-0.36 (0.27)
Class size * non-Swedish	-1.67 (0.49)	-0.51 (0.26)
Class size * socio economic index	0.18 (0.42)	0.42 (0.21)
Initial test score change control	No	Yes
R^2	0.044	

Notes: Number of observations is 556. The standard errors, in parentheses, allow for regression errors that are correlated among pupils in the same school. Test scores are measured in percentile ranks. In column 2, when control for initial test score change (from spring of the fifth grade to fall of the sixth grade) is made, the estimates and standard errors assumes a true reliability ratio of 0.7878 in the fifth grade test score percentile ranks. Column 2 also use test score in the spring of the fifth grade as instrument for the initial test score change. Both columns include controls for are girl, non-Swedish parents, the socioeconomic index, a quadratic in teacher experience and a linear for teacher experience in current class.

The interaction terms in column 1 show that pupils with non-Swedish parents are likely to benefit the most from smaller classes. When controlling for previous test score this is still true, although the estimate goes down to one-third of the previous value. The heterogeneity of the class size effect regarding socioeconomic background is less clear. The estimates are positive in both columns, indicating that pupils with low socioeconomic background gain more from smaller classes, but not statistically significant in column 1. Note that for the mean, pupil, the average class size estimates in both columns in Table 7.4 still gives that smaller classes increase test scores.

8 A Cost-Benefit Analysis of Class Size Reductions

Many studies suggest that education has a causal effect on earnings (see, e.g., Card, 1999, for a survey). Two important benefits of increased school resources, for more smaller classes, are that students learn more and raise their educational aspirations, which pays off in terms of better job placements and higher earnings later on when students join the labour market. Nevertheless, the effect of smaller classes resources on achievement is most commonly measured in terms of student performance on standardized tests. This section converts test outcome measures into US dollars and Swedish Kronor by using the relationship between test scores and earnings. This relationship is used to calculate the internal rate of return from reducing class size.

8.1 Cost-benefit analysis for US

Three recent studies illustrate the magnitude of the relationship between students' test scores while in school and their subsequent earnings in US. Murnane, Willet and Levy (1995) estimate that male high school seniors who scored one standard deviation (SD) higher on the basic math achievement test in 1980 earned 7.7 percent higher earnings six years later, based on data from the High School and Beyond survey. The comparable figure for females was 10.9 percent. This study, however, also controls for students' eventual educational attainment, so any effect of cognitive ability as measured by test scores on educational attainment is not counted as a gain from higher test scores. Currie and Thomas (1999) use the British National Child Development Study to examine the relationship between math and reading test scores at age 7 and earnings at age 33.They find that students who score in the upper quartile of the reading exam earn 20 percent more than students who score in the lower quartile of the exam, while students in the top quartile of the math exam earn another 19 percent more. Assuming normality, the average student in the top quartile scores about 2.5 standard deviations higher than the average student in the bottom quartile, so their results imply that a one SD increase in reading test performance is associated with 8.0 percent higher earnings, while a one standard deviation increase in the math test is associated with 7.6 percent higher earnings. Neal and Johnson (1996) use the National Longitudinal Survey of Youth to estimate the effect of students' scores on the Armed Forces Qualification Test (AFQT) taken at age 15–18 (adjusted for age when the test was taken) on their earnings at age 26–29. They find that a one SD increase in scores is associated with about 20 percent higher earnings for both men and women.

Neal and Johnson find a larger effect of test scores on wages than Currie and Thomas probably for three reasons: (1) students were older when they took the AFQT exam, and Currie and Thomas find some mean regression in test scores; (2) Neal and Johnson examine the effect of only one test score, whereas Currie and Thomas simultaneously enter the reading and math score in a wage equation, and the scores are correlated; (3) the British and American labour markets are different. Based on these three studies, a plausible assumption is that a one SD increase in either math or reading scores in elementary schools is associated with about 8 percent higher earnings in US.

From an investment perspective, the timing of costs and benefits is critical. The cost of hiring additional teachers and obtaining additional classrooms are borne up front, while the benefits are not realized until years later, after students join the labour market. To illustrate the benefits and costs, consider extending the STAR class-size reduction experiment to the average student in the U.S. who entered kindergarten in 1998. In the STAR experiment, classes were reduced from about 22 to about 15 students, so assume that funds are allocated to create 7/15 = 47 percent more classes. The formula for the present value (PV) of the costs discounted to the initial year (1998) is shown in the Technical appendix 2.

Probably a reasonable approximation is that the cost of creating and staffing 47 percent more classrooms is proportional to the annual per pupil cost.[34]

[34] Folger and Parker (1990) tentatively conclude from the STAR experiment that proportionality is a reasonable assumption.

Table 8.1. **Discounted present value of costs and benefits of reducing class size from 22 to 15 in grades K-3 in US (1998 Dollars),**

Discount rate	Cost ($)	Increase in income assuming annual productivity growth rate of:		
		None	1 Percent	2 Percent
(1)	(2)	(3)	(4)	(5)
0.02	$7,787	$21,725	$31,478	$46,294
0.03	$7,660	$15,174	$21,667	$31,403
0.04	$7,537	$10,784	$15,180	$21,686
0.05	$7,417	$7,791	$10,819	$15,238
0.06	$7,300	$5,718	$7,836	$10,889

Note: Figures assume that a 1 standard deviation increase in math test scores or reading test scores in grades K-3 is associated with an 8 percent increase in earnings, and that attending a small class in grades K-3 raises math and reading test scores by 0.20 SD. Real wages are assumed to grow at the same rate as productivity. Costs are based on the assumption that students are in a smaller class for 2.3 years, as was the average in the STAR experiment. Note that in 1998, 9.77 Swedish Kronor was needed to buy 1 $, according to PPP-figures from OECD.

We assume the additional cost per pupil each year a pupil is in a small class equals $3,501, or 47 percent of $7,502, which was the nationwide total expenditures per student in 1997–98.[35] Although the experiment lasted 4 years, the average student who was assigned to a small class spent 2.3 years in a small class.[36] As a consequence, we assume the additional costs are $3,501 in years one and two, 30 percent of $3,501 in year three, and zero in year four. Column (2) of Table 8.1 provides the PV of the costs for various values of the discount rate.

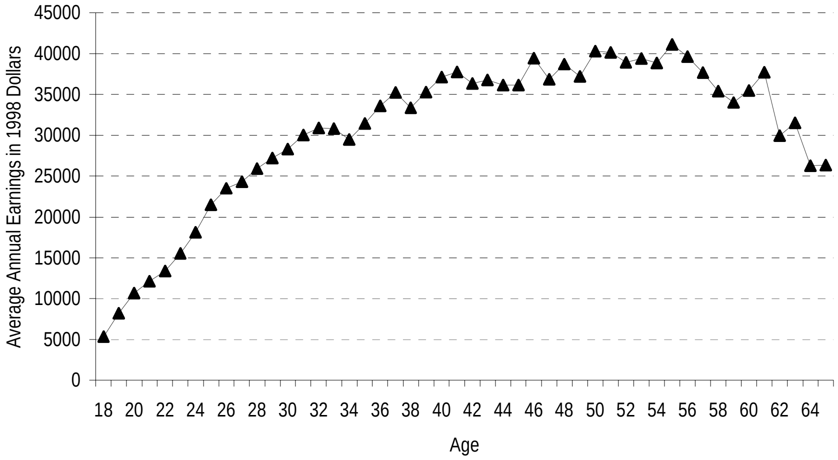
The pecuniary benefits of reduced class size are harder to quantify, and occur further in the future. Figure 8.1 illustrates the age-earnings profile for workers in 1998.[37]

[35] See *Digest of Education Statistics, 1998*, Table 169.

[36] Students spent less than four years in a small class because half the students entered the experiment after the first year, and because some students moved to a new school or repeated a grade, causing them to return to regular size classes.

[37] The figure is based on data from the March 1999 Current Population Survey. The sample consists of all civilian individuals with any work experience in 1998.

Figure 8.1. Age earnings profile in US, 1998



The figure displays average annual earnings for workers at each age between 18 and 65. As is commonly found, earnings rise with age until workers reach the late 40s, peak in the early 50s, and then decline. Average earnings are quite low until workers reach their mid 20s.

Suppose for the time being that the earnings of the current labour force represents the exact age-earnings profile that the average student who entered kindergarten in 1998 will experience when he or she completes school and enters the labour market. Using the age-earnings profile, we can calculate the average real earnings each year after age 18. The preceding discussion suggests that 8 percent is a reasonable estimate of the increase in earnings associated with a one standard deviation increase in either math or reading test scores. The STAR experiment suggests 0.20 SD is a reasonable figure to use as an estimate of the increase in test scores for both math and reading due to attending smaller class in grades K-3 (see, e.g., Finn and Achilles, 1990, Mosteller, 1995, Krueger, 1999, and section 5 in this paper).[38,39] The addition to annual earnings must be discounted back to the initial year to account for the fact that a dollar received in the future is less valuable than a dollar received today. Assuming students begin work at age 18 and retire at age 65, the present value of the higher earnings stream due to smaller classes is calculated as shown in technical appendix 2. Using these assumptions, column (3) of Table 8.1 reports the PV of the additional earnings due to reducing class size by 7 students for various values of the discount rate.

One important issue, however, is that real average earnings are likely to grow substantially between 1998 and when the average kindergartner of 1998 retires. That is, when the kindergartners of 1998 enter the labour market, their average real earnings will be greater than that depicted in Figure 8.1. Real wages typically grow in step with labour productivity. Over the 20th century, real earnings and productivity have typically grown by 1 or 2 percent per year. The estimates of test scores on earnings discussed above are all based on earnings long after students started school, which reflect the effect of higher productivity growth on earnings. Consequently, columns (4) and (5) present discounted benefits assuming either 1 or 2 percent annual productivity and real wage growth after 1998 (see the Technical appendix 2 for details). The latest U.S. Social Security Trustees' intermediate projection is for real wages to grow by slightly less than 1 percent per year over the next 75 years, so column (4) probably provides a reasonable forecast of future earnings.

The next question is which discount rate should one use to discount costs and benefits from age 5 until 65? The current yield on essentially risk-free long-term inflation-indexed government bonds is just under 4 percent. If we assume an interest rate of 4 percent (row 3), then the benefits of reducing class size from 22 to 15 in the early grades would be 43 percent greater than the

[38] Work by Krueger and Whitmore (2001) and Nye, Zaharias, Fulton, et al. (1994) suggests that the improved test performance of small class students in Project STAR may have fallen to about 0.10 standard deviations by the end of high school. Although we suspect that some of the initial gain from small classes in the STAR experiment faded after students returned to regular-size classes, the calculations reported in Table 16 are probably still reasonable. The reason for this is that Currie and Thomas's estimate of β is based on test scores at age 7. They find some regression to the mean in test scores as students age. If the 0.10 SD gain at older ages is used in the calculations, then the appropriate estimate to use for β would be higher.

[39] Note that in section 5, we scaled the tests in percentile ranks. In this section we instead interpret the class size effects in standard deviation units. This is simply done by dividing the estimated effect, using percentile ranks, with the estimated standard deviation of the percentile rank test scores, which approximately will equal 28.

costs absent real wage growth, and 100 percent greater than the costs if real wages grow by 1 percent per year. If society desires to reflect some risk in the interest rate used to discount future benefits of reduced class size – because the payoff is uncertain – a higher discount rate would be desired. With a discount rate of 6 percent and 1 percent annual productivity growth, the costs of reducing class size from 22 to 15 students are predicted to almost equal the benefits.

The internal rate of return, r*, can be calculated by solving for the discount rate that equates the benefits and costs, as shown in the technical appendix 2. If earnings grow by 1 percent per year, as expected by the Social Security Trustees, the internal rate of return is 6.2 percent.[40]

In Krueger and Whitmore (2001), a comparison of the effects of attending a smaller class with the effect of private school vouchers, as estimated by Howell, Wolf, Peterson and Campbell (2000), is made. They conclude that, when comparable samples are considered, black students who attended a small class for two years in the STAR experiment improved their test performance by around 50 percent more than the gain experienced by black students who attended a private school as a result of receiving a voucher in the New York, Dayton and Washington voucher experiments.

8.2 Cost-benefit analysis for Sweden

In a similar fashion as in the previous section, we can calculate PV of cost and benefits of a class size reduction for Sweden. We estimated the relationship between test scores and future earnings for Sweden using data from the UGU-project (see Härnqvist, Emanuelsson, Reuterberg & Svensson (1994)). The scores are from math, reading and IQ tests for 6th grade pupils born 1948 and 1953, and the earnings are from 1993. In all the estimations we controlled for fathers and mothers education, and for pupils gender.

[40] The internal rate of return on an investment is the discount rate that equate the PV of the investment benefits and costs. The investment is here the resources spent on a class size reduction. If the internal rate of return on this investment is higher than for an alternative investment, the class size reduction is preferred.

The test scores are standardized with mean zero and standard deviation one.[41] The regression results are shown in Table 8.2.

Table 8.2. **Regressions of log(earnings) on test-scores, cohorts 1948 and 1953**

	Dependent variable is Log(earnings)			
	(1)	(2)	(3)	(4)
A. 1948 cohort				
Math test score	0.13 (0.01)	0.10 (0.01)	0.11 (0.01)	0.09 (0.01)
Reading test score		0.05 (0.01)		0.04 (0.01)
IQ test score			0.05 (0.01)	0.03 (0.01)
Controls for gender, mothers´ and fathers´ education	Yes	Yes	Yes	Yes
R2	0.178	0.181	0.180	0.182
B. 1953 cohort				
Math test score	0.09 (0.01)	0.06 (0.01)	0.06 (0.01)	0.05 (0.01)
Reading test score		0.05 (0.01)		0.04 (0.01)
IQ test score			0.04 (0.01)	0.02 (0.01)
Controls for gender, mothers´ and fathers´ education	Yes	Yes	Yes	Yes
R^2	0.150	0.153	0.151	0.153

Notes: Number of observations is 7760 for 1948 cohort and 6687 for 1953 cohort. Math, reading and IQ test scores are in standard deviations. The reading score are in 1948 calculated as the average of standardized scores from tests of reading and writing in Swedish and of an English test. For 1953, the reading score are the average of standardized scores from tests of Swedish and English. Robust standard errors are in parentheses. Controls are four dummies for mothers´ education, four dummies for fathers´ education and a dummy for gender. The mean(st.dev) of Log(earnings) are 12.02 (0.59) in Panel A. and 11.93 (0.62) in Panel B.

[41] Since the reading and IQ scores are based on average scores on separate test parts, each part is standardized with mean zero and standard deviation one. The standard deviations of the average reading scores are still very close to one, whereas for IQ they are about 0.75 for both years. The IQ estimates should therefor be multiplied by about 1.3 if interpreted in standard deviation units.

Without reading and IQ scores included, a one standard deviation increase in math test score is associated with 9-13 percent higher earnings. If scores on the IQ test are included as additional controls, the association is 6–11 percentage, but still highly significant. The effect of the sum of the math and reading test is 11–15 percent, without IQ as control, and 9–13 percent, with IQ as control. The latter figures are probably too low to use in the PV calculations however, since class size reductions might also affect IQ scores, and since the IQ and math scores are highly correlated. Note that an advantage with using these estimates is that the tests were conducted at the same age in the UGU-data, as in the data used in analysis in section 7. We believe that it is reasonable to assume that a one SD increase in either math or reading scores in primary schools in Sweden is associated with about 6 percent higher earnings, where 6 percent is the average between math and reading scores in columns 2 and 4. Controlling for reading scores the average math score estimate is 7.5 percent. Controlling for math scores the average reading score estimate is 4.5 percent.

In the STAR experiment, classes were reduced from about 22 to about 15 students, so we assumed that funds were allocated to create 7/15 = 47 percent more classes. For Sweden, we consider the policy of decreasing class sizes by the same amount, and continue to assume that that the cost of creating and staffing 47 percent more classrooms is proportional to the annual per pupil cost. For Sweden, we assume the additional cost per pupil each year a pupil is in a small class equals SEK 24,700, or 47 percent of SEK 52,900, which was the nationwide total expenditures (including cost of housing facilities and inventories in 1998.[42]

An important difference to the STAR experiment, is that the class size reduction there was done for several years. In our analysis for Sweden, the effect of a class size reduction is estimated only for one grade. Hence the average student could be assumed to have spent only 1 year in a small or large class. This is clearly not the case though, since class sizes are correlated across grades. We do however have information about pupil's class sizes in both the 5th and 6th grades. The average class size in 6th grade is 19.9. For the pupil's with class sizes less than 18 pupils, the average class size is 13.3. A decrease of classes from 19.9 to 13.3 require 50 percent more classes, approximately equal to the class size reduction in

[42] See National Agency of Education, 1999, no.173.Table XXX.

STAR. Hence we label a class size of less than 18 pupils, a small class. 55 percent of the pupils in a small class in 6th grade was also in a small class in 5th grade. Hence, we assume that these pupils spent on average 3.75 [$=(0.55 \cdot 5)+1$] years in a small class in grades 1–6. For the mean pupil, there was about 20 percent chance of spending the 5th grade in a small class. Hence, the average pupil spent 1 [$=0.20 \cdot 5$] year in a small class in grades 1–5. So, the "small-class"-pupil spent 2.75 more years in a small class in grades 1-6, compared to a "regular-class"-pupil. So the costs should be based on the assumption that pupils are in a small class for 2.75 years, instead of just being in a small class for one year in grade 6. Column (2) of Table 8.3 provides the PV of the costs for Sweden for various values of the discount rate.[43]

Table 8.3. **Discounted present value of benefits and costs of reducing class size from 19.9 to 13.3 in grades 1–6 in Sweden (1998 SEK).**

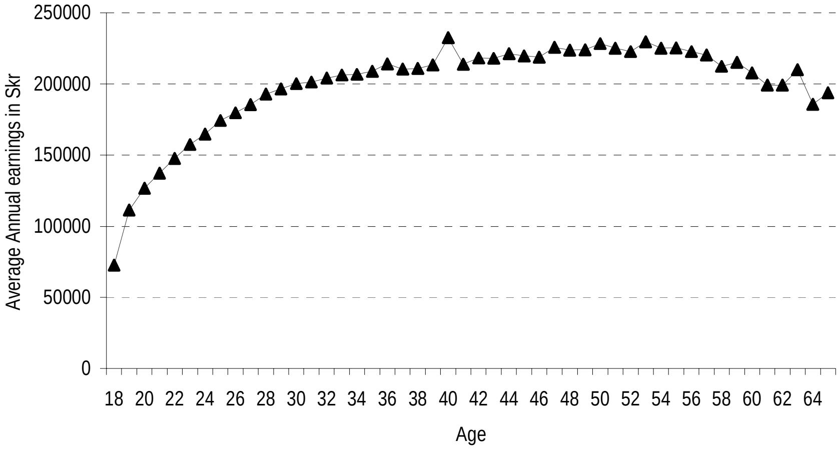
Discount rate	Cost (SEK)	Increase in income assuming annual productivity growth rate of:		
		None	1 Percent	2 Percent
(1)	(2)	(3)	(4)	(5)
0.02	63,264	108,354	154,222	223,229
0.03	61,183	78,659	110,155	156,836
0.04	59,206	58,155	80,184	112,343
0.05	57,326	43,735	59,423	81,982
0.06	55,537	33,413	44,783	60,890

Note: Figures assume that a 1 standard deviation increase in math or reading test scores in grades 1–6 is associated with an 6 percent increase in earnings, and that attending a small class in grades 1–6 raises math and reading test scores by 0.19 SD. Real wages are assumed to grow at the same rate as productivity. Costs are based on the assumption that students are in a smaller class for 2.75 years, as was estimated to be the average in the study for Stockholm.

[43] Note that the costs for US in Table 8.1 and Sweden in Table 8.3 are not exactly comparable. For US, the costs are based on students being in a small class for 2.3 years o grades k-3, whereas for Sweden they are based on being in a small class for 2.75 years in grades 1–6.

Also regarding the pecuniary benefits of reduced class size, we follow the path of the previous section. Figure 8.2 illustrates the age-earnings profile for all workers in 1998.[44]

Figure 8.2. Age earnings profile in Sweden, 1998



The figure displays average annual earnings for individuals at each age between 18 and 65. The age-earnings profile is similar as in the US, although somewhat less concave.

As in the analysis for US, suppose that for the time being that the earnings of all current individuals represents the exact age-earnings profile that the average 1th grade student in 1998 will experience when he or she completes school and enters the "labor market". The preceding discussion suggests that 6 percent is a reasonable estimate for the value of the increase in earnings associated with a one standard deviation increase in either math or reading test scores. The analysis in section 7 suggests that decreasing classes with 1 pupil, is associated with about 0.4–1.0 percentile points higher math scores. Using the average of the two class size estimates in columns 2 and 5 of Table 7.3, which is 0.67, and dividing this number by 23, which is the standard deviation (SD) of the 6th grade math scores, give us that a one pupil decrease

[44] The figure is based on data from the LINDA data base in 1998 (see Edin and Fredriksson, 2000). The sample consists of almost 98,200 workers, between ages 18–65 in 1998. The yearly earnings is calculated as the monthly wages times the percentage of full time the respondent work (both reported by the employer) times 12.

is associated with a 0.029 SD increase in math scores. Decreasing class sizes by a similar amount as US, but here from 19.9 pupils, which is the average class size in 6th grade in our Stockholm data, equates to decreasing class sizes to 13.5 pupils per class. Hence, a 6.4 pupils smaller class size is associated with about 0.19 SD increase in math scores, on average. We were never able to separately investigate the effect of class size on reading scores. Based on the results from STAR, we here assume the effect of class size on reading and math scores to be the same.[45] If as before, we assume that students begin work at age 18 and retire at age 65, using the formula in technical appendix 2, we can calculate the present value of the higher earnings stream due to smaller classes. Using these assumptions, column (3) of Table 8.3 reports the PV of the additional earnings due to reducing class size by 6.4 students in Sweden, for various values of the discount rate.

As for US, columns (4) and (5) present discounted benefits assuming either 1 or 2 percent annual productivity and real wage growth after 1998. Regarding which discount rate one should use, we note that, the current yield on governmental bonds in Sweden are a bit higher then 4 percent. But if we also for Sweden assume an interest rate of 4 percent (row 3), then the benefits of reducing class size would be about the same as the costs absent real wage growth, and almost 35 percent higher than the costs if real wages grow by 1 percent per year. With a discount rate of 5 percent and 1 percent annual productivity growth, the costs of reducing class size are predicted to almost equal the benefits, the internal rate of return being 5.1 percent.

[45] This might seem as a very strong assumption. However, if we instead disregard reading scores altogether (and thereby assume class size to have no effect on reading scores) and hence use the estimate of math scores on earnings, unconditional on reading scores, the results only changes slightly. In this case, using average math score estimates from columns 1 and 3 of Table 8.3 which is 0.10, and multiply this by 0.19, which was the effect of 6.4 fewer pupils per class on math scores, give us 0.019 instead of $0.12 \cdot 0.19 = 0.023$. The benefit figures in Table 8.3, would then be 83 percent of the present numbers.

8.3 Caveats

The cost-benefit calculations presented here are subject to many qualifications. We consider the following to be most important:

- The effect of test score gains on earnings in the future may turn out to be different than the value assumed. Indeed, because this value was estimated from cross-section relations it could reflect the effect of omitted characteristics.[46] In addition, general equilibrium effects could affect the value assumed if class size is reduced on a wide scale. It is also likely that school resources influence noncognitive abilities, which in turn influence earnings especially for blue collar workers (see Cawley, et al., 1996), but are not reflected in test scores.
- Class size probably influences other outcomes with economic consequences, such as crime and welfare dependence, and there may be externalities from human capital, so the economic benefits could be understated. In addition, improved school quality probably has non-economic private and social benefits, such as improved citizenship and self-enlightenment.
- It is unclear how much real earnings will grow in the future, although the 0 to 2 percent annual growth figures probably provide a reasonable range.
- The cost of reducing class size in the early grades may be different than assumed here. For example, expenditures per student are typically lower in grammar school, yet expenditures per students in all grades was used.
- The quality of teachers could decline (at least in the short run) if class size is reduced on a wide scale. This could however be partly offset by increased supply of teachers when class sizes decreases.[47]

[46] Note, however, that Jencks and Phillips (1999) find that math test score gains between 10th and 12th grade have about the same impact on subsequent earnings as cross-sectional differences in scores of equivalent magnitude in 10th grade.

[47] Some support for this is given by the Swedish Teachers Union, who report that one reason for why newly examined teachers is unwilling to stay for long in the teaching profession, is too large classes in Swedish schools (Lärarnas Riksförbund, Arbetsmarknadsenkät 2000).

- The calculations in Tables 8.1 and 8.3 neglects fringe benefits, which are about one third of total compensation. If fringe benefits are proportional to earnings, the reported benefits are understated by about one third. The calculations for US also assume that everyone works for pay, at least part year, which tends to overstate the economic benefit.
- The cost-benefit calculations do not take into account distributional effects.

8.4 Summing up

The cost-benefit calculations described above subject to the many qualifications listed there, suggest that the internal rate of return from increasing the number of classes by about 50 percent in grades K-3 for US and grades 1–6 for Sweden is about 6 and 5 percent respectively. The "critical effect size" – or minimum gain for the benefit of a reduction of class size by this amount to equal the costs – would equal 0.10 standard deviation units for US if productivity grows at 1 percent per annum, and a 4 percent real discount rate is assumed. The corresponding critical effect size for Sweden is only somewhat higher, 0.14 standard deviations. These would be natural null hypothesises against which to test estimates for these countries to judge their economic significance.

The relationship between class size and achievement is not as robust as, for example, the relationship between years of education and earnings. But this is probably because relatively small achievement gains from smaller classes translate into positive benefit-cost differentials. Even subtle effects could be economically important. Anyone who expects much larger gains from reducing class size than was found in this study is expecting extra-normal returns. Although such returns are possible, economists are usually sceptical that such large returns are available.

9 Conclusions

In this study we have shown that, contrary to the public policy view, smaller classes do increase pupil achievement. We therefore conclude that the increase in pupil teacher ratios observed in Sweden is likely to had negative consequences for the scholastic achievement among Swedish pupils.

This study has noted that in Sweden resources spent on basic schooling has decreased by 10 percent since 1990. For primary and secondary schooling, expenditure per pupil has decreased by 9 percent during the same period. We note however, that in 1990 the expenditure per pupil was extraordinarily high historically in Sweden. From 1978 to 1998, the expenditure per pupil in primary and secondary schooling in Sweden increased by 10 percent. When we compared with the development for the US, we noted that for the same period and the same schooling levels, expenditure per pupil has increased by about 50 percent in fixed US dollars, and by about 100 percent if we convert US dollars to Swedish Kronor. We also noted that while in US the fraction of GDP spent on primary and secondary education increased by 7 percent from 1978 to 1998, in Sweden this fraction decreased by 25 percent.

Next, we reinterpreted the evidence from the most influential existing review on the connection between school resources and pupils achievement. This review concluded that more school resources, in terms of smaller classes or higher expenditure per pupil, is unlikely to have any significant effect on pupils achievement. We showed however that this result is due to an overweighting of studies showing no or negative effects of more school resources. When we weight each study equally, we found that significantly more studies found positive effects of increased school resources, either measured as smaller classes or higher expenditures per pupil.

We then turned to the only large-scale experiment ever conducted on the class size issue, the Tennessee STAR experiment. In STAR pupils and teachers were randomized into small and large classes. The experiment lasted from kindergarten to 3rd grade, after which pupils were reallocated to normal sized classes. We first presented evidence that the design of STAR were such as intended, i.e. the randomization seems to have been done properly. We next showed that this experiment indeed has beneficial achievement effects of smaller classes, and that this effect is especially prevalent for disadvantaged groups. It is also the case that even though the experiment only lasted until third grade, the beneficial effects did last until later grades, even though they did decrease. We also presented evidence that for some disadvantaged groups, the pupils from small size classes scored significantly higher on college tests almost 10 years after the experiment ended.

We then turned to evidence for Sweden. Noting the very sparse and old evidence on the class size issue using Swedish data, we collected new data on pupils in the municipality of Stockholm. Since we did not have the opportunity to run an experiment, like the STAR, in Sweden, it is very difficult to be able to draw conclusions with the same certainty where non-experimental data is used. We did however collect data with the purpose of drawing conclusions of the effect of pupils' achievement, from implementation of a class size reduction. We therefore tested the pupils three times, before and after a 10 week summer break, and at the end of the following school year. This design made it possible to relate the change in test scores during the school year to class sizes, controlling for the change in the test scores during the summer. The intuition behind this way of estimating class size effects is that schools are closed during the summer, and hence schooling characteristics should not influence summer test score changes, whereas schools are open during the school year, when schooling characteristics might influence test score changes. The results from the study for Stockholm were that smaller class sizes indeed positively affected achievement levels, and that this positive effect were especially prevalent for a group of pupils that on average has relatively low achievement levels (pupils with non-Swedish parents).

Lastly, we calculated the costs and benefits of a class size reduction for US, using the results from STAR, and for Sweden, using the results from the Stockholm pupil data. We found that decreasing class sizes can, with reasonable assumptions, indeed have benefits that are larger then the costs. For US we found that for benefits to be larger then the costs when implementing a class size reduction of about 50 percent, the required test score effect size is about 0.10 standard deviations. For Sweden the required test score effect for the same class size reduction is somewhat higher, about 0.14 standard deviations.

In the US, former President Bill Clintons initiated a gigantic nation-wide class-size reduction plan, starting in 1999. Even though it is too early to know what effects this might have, an example of an actual, state-wide class-size reduction initiative that could provide a model for other countries is the one in California (see Stecher, et al. (2000)). In this enormous education reform, which was championed by then-Governor Pete Wilson, California school districts that chose to participate received just over $800 for each K-3 student enrolled in a class of 20 or fewer students to encourage smaller classes. Because of the scale of this intervention, many implementation problems were encountered that do not arise in small-scale demonstration studies. For example, some higher income school districts reportedly raided teachers from lower income districts. In addition, many new classrooms had to be built to accommodate smaller classes, and temporary structures were often used. Nonetheless, Stecher, et al. find that, after two years, the California class-size reduction initiative led to a 0.10 S.D. increase in math scores and a 0.05 S.D. increase in reading scores on the SAT-9 exam for third graders. Both of these effect sizes were statistically significant. They also found some evidence that schools with a larger share of minority students had larger effect sizes.

Technical appendix 1:
A formal presentation of the estimated models

In this section, we first show the level model and the variant of it which is used in estimating class size effects using the experimental STAR data. We also show the most common estimation technique when non experimental data is used; the value added model. We then show a new way to estimate educational production functions, which in section 7 is applied to data from Stockholm.

The level model

The simplest model to use when estimating class size effects is to simply regress achievement level on a measure of class size. This can be expressed as:

$$(1) \qquad A_{it} = \theta + \beta CS_{it} + u_{it},$$

where A_{it} is the achievement level for pupil *i* at the end of grade *t*; CS_{it} denote the class size for pupil *i* in grade *t*, and u_{it} is a regression error, consisting of everything that contributes to the variation in A_{it}, except for class size. θ is the intercept and β is the parameter of interest.

The model used for the experimental data

If ideal experimental data is available, we can simply estimate (1) to get the causal effect of class size on achievement level, i.e. β. The reason is that in an ideal class size experiment, pupils and teachers are randomly allocated to classes of different sizes. This means that class sizes would be independent to the regression error, which includes everything else that is affecting achievement. Intuitively, a class size experiment makes the size of the class unaffected by anything systematic.

Since the randomization in the Tennessee experiment was done within schools and student entried the experiment in different grades, interactions between school- and entry-by-grade fixed effects are added to model (1). Also, in some specifications, family background variables are added. Hence the following model is estimated in section 5:

$$A_{ist} = \theta_t + \beta_t CS_{ist} + \o_t F_{is} + m_{sw} + u_{ist}, \tag{2}$$

where A_{ist} denotes the achievement level for pupil *i* in school *s* in grade *t*; CS_{ist} denote the class size for pupil *i* in school *s* in grade *t*; F_{is} denotes a vector of demographic, family background and neighborhood characteristics for pupil *i* in school *s*; $\theta_{t,}$,β_t and $\o_t$ are grade specific intercepts, class size effects and ”family” effects, respectively; m_{sw} is a set of school-by-entry-wawe fixed effects (based on initial school attended) and u_{ist} is a random error term.

In estimation using the STAR data, *CS* is represented by SMALL, a dummy variable that equals one if student *i* initially was assigned to a small class and zero if he or she was assigned to a regular or regular/aide class. The reason for using initially assignment to class type is for the class size estimates to not be biased due to non-random transitions after the initial assignment.

Note that if achievement level tests are available for the same pupils in several earlier grades, the scores on these tests could be added to equation (2), but will not affect β if the randomization is done correctly.

The value-added model

If only non-experimental data are available, estimation of equations (1) or (2) will typically give inconsistent estimates of class-size effects. There are many reasons for this (see the discussion in section 5.1). A very popular framework for estimating class size effects, using non-experimental data, is to estimate a so-called value-added model, which can be expressed as:

$$A_{it} - A_{it-1} = \theta + \beta CS_{it} + \rho S_{it}^{/} + \phi F_{it} + v_i + u_{it}, \tag{3}$$

where $A_{it} - A_{it-1}$ is the change in achievement level for pupil *i* that has occurred between the end of grade *t-1* and the end of grade *t* and *S'*$_{it}$ denotes a vector of schooling variables other than class size (for instance teacher quality) in grade *t*. The error term in (3) is assumed to consist of two parts, v_i which is a (time) fixed learning effect that captures family background, innate ability, and everything else that has constant influences on achievement change for pupil *i* during period *t* and u_{it}, which is a random error term that is assumed to be orthogonal to F_{it}, S_{it} and v_i .

If lagged achievement level is allowed to affect the change in achievement between grades we can instead write (3) as:

$$A_{it} = \theta + \beta CS_{it} + \rho S_{it}^{/} + \phi F_{it} + \lambda A_{it-1} + v_i + u_{it}. \tag{4}$$

In both equations (3) and (4), the lagged achievement level, A_{it-1}, captures all the previous observed and unobserved pupil, family, neighborhood, and school characteristics, as long as these characteristics affects the level and not the change in achievement. These characteristics, including any unobserved fixed achievement level effect before school starts, do hence not biasing parameter estimates of equation (3). Equations' (3) and (4) are in this study referred to as the *value-added* specifications.

A potential drawback with the value-added model is that it fails to eliminate the fixed learning effect, v_i. The reason for this is that in (3) and (4), we have allowed unobservable time-constant factors to have an effect on achievement growth through the fixed learning effect, besides a one-time effect on achievement level. If the fixed learning effect is correlated with CS_{it}, all parameter estimates will be biased. The approach outlined in the next section attempts to eliminate biases due to both fixed learning effects, as well as fixed achievement level effects.

The difference-in-differences model

So far, we have assumed that achievement level could only be measured at the end of each grade level, *t* and *t-1*. Suppose achievement level could also be observed at the start of each school year. For expository purposes, assume that each grade level consists of two parts of equal length, the summer vacation and the school period. In reality the summer period is much shorter then the school period (in Sweden the summer vacation is 10 weeks) but this can easily be dealt with in empirical applications. The part of grade t, when school is in session, is denoted j=2, and the part of the grade, where school is out of session, i.e. the summer vacation, is denoted j=1.

Assuming that previous achievement level do not affect the change achievement during the summer and during the school year, equation (3), at grade t (for j = 1, 2) can then be expressed as:

$$\Delta A_{it,1} = \kappa_1 + \alpha_1 F_{it} + \delta_i + \varepsilon_{it,1} \tag{5}$$

$$\Delta A_{it,2} = \kappa_2 + \beta CS_{it} + \rho S_{it}^{/} + \alpha_2 F_{it} + \delta_i + \varepsilon_{it,2}, \tag{6}$$

where $\Delta A_{it,1} = A_{it,1} - A_{it-1,2}$ is the achievement change during the summer period; $\Delta A_{it,2} = A_{it,2} - A_{it,1}$ is the achievement change during the school period; $A_{it,1}$ is achievement level at the start of the school year in grade *t*; $A_{it-1,2}$ is achievement level at the end of the school year in grade *t-1*; $A_{it,2}$ is achievement level at end of school period in grade *t*; and κ_1 and κ_2 is intercepts allowing the average achievement change to be different during the school and summer periods. The error terms are assumed to consist of two parts; δ_i which is the fixed learning effect and ε_{itj} which are random error terms. The latter terms are assumed not to be correlated with F_{it} , S_{it} and δ_i.[1]

Equation (5) expresses summer learning as a function of family background and the fixed learning effect. Equation (6) expresses learning over the school period as a function of family background, school characteristics and the fixed learning effect. The important difference between equation (5) and equation (6) is that schooling characteristics are allowed to affect achievement only when schools are in session, whereas family background characteristics can influence achievement when schools are in session and when they are not. Note that equation (3) is a special case of equations (5) and (6), since the difference is that in equations (5) and (6), grade level t is divided into a summer, when j=1, and a school period, when j=2.

Note that since schools are out of session during the summer, school characteristics are expected not to influence learning in grade *t*, when j=1. This makes the achievement level in grade *t*,

[1] Note that we have assumed that F_{it} is the same at the end of the summer and at the end of the school year.

when j = 1, depend on cumulative schooling factors only until time period *t-1*. $A_{it,1}$ and $A_{it-1,2}$ are both functions of all previous pupil, family, neighborhood and school characteristics, including an individual-specific achievement effect that captures the unobserved achievement level before the school starts.

In the following, we will assume that in equations (5) and (6), the parameters linking family background to achievement, are the same at the end of the summer and at the end of the school period, i.e. $\alpha_1 = \alpha_2$.

We can eliminate the fixed learning effect by taking the difference between (6) and (5) to get:

$$\Delta A_{it,2} - \Delta A_{it,1} = \kappa' + \beta CS_{it} + \rho S_{it}^{'} + \Delta\varepsilon_{it,2}, \tag{7}$$

where the dependent variable is the difference between learning during the school and summer periods; $\Delta\varepsilon_{it,2} = \varepsilon_{it,2} - \varepsilon_{it,1}$; and $\kappa' = \kappa_2 - \kappa_1$. Estimation of equation (7) will produce consistent estimates of the effect of class size on pupils' achievement levels, i.e of β, if the assumption that lagged test scores do not affect changes in test scores, conditional on the fixed learning effect and family- and schooling characteristics, is correct.

The principal identification strategy still holds, but becomes more complicated, if lagged achievement level is allowed to have an effect on the achievement change.

Equations (5) and (6) are then expressed as:

$$A_{it,1} = \kappa_1 + \alpha_1 F_{it} + \gamma_1 A_{it-1,2} + \delta_i + \varepsilon_{it,1} \tag{8}$$

$$A_{it,2} = \kappa_2 + \alpha_2 F_{it} + \beta CS_{it} + \rho S_{it}^{/} + \gamma_2 A_{it,1} + \delta_i + \varepsilon_{it,2}, \tag{9}$$

where equations (8) and (9) are generalizations of equation (5) and (6). Taking the difference between (9) and (8), assuming $\alpha_1 = \alpha_2$, we get:

$$\Delta A_{it,2} = \kappa' + \beta CS_{it} + \rho S_{it}^{/} + \gamma_2 \Delta A_{it,1} + \Delta\gamma A_{it-1,2} + \Delta\varepsilon_{it,2} \tag{10}$$

where $\Delta\gamma = \gamma_2 - \gamma_1$. It is not possible to estimate the class size parameter in equation (10) consistent unless some restriction is imposed.

If we assume that $\gamma_1 = \gamma_2 = \gamma$, that is, previous test score level has the same effect on the change in test scores during the summer and during the school year, we can rewrite (10) to get:

$$(11) \qquad \Delta A_{it,2} = c' + \beta CS_{it} + \rho S_{it}^{/} + \gamma \Delta A_{it,1} + \Delta \varepsilon_{it,2}$$

Due to the correlation between $\Delta A_{it,1}$ and $\Delta \varepsilon_{it,2}$ (since $\mathrm{cov}(A_{it,1}, \varepsilon_{it,1}) \neq 0$), the parameter estimates will be biased if equation (11) is estimated by OLS. So we instead estimate this equation by using $A_{it-1,2}$ as an instrument for $\Delta A_{it,1}$.[2]

The main difference between equations (3) and (4) and equations (7) and (11), is that the last two specifications eliminates the unobservable fixed learning effect, whereas the first two specifications do not. In this study, equations (7) and (11) are referred to as the *difference-in-differences* approach to estimate educational production functions. In Lindahl (2002), there is shown that if $\gamma_2 > \gamma_1$, equation (11) is likely to give an underestimate of the effect of *smaller* classes on achievement level.

[2] This is the recommended method for estimating dynamic panel-data models, based on the analysis by Anderson and Hsiao [1981] and Arellano [1989].

Technical appendix 2: Cost-Benefit calculations

The Present value of the costs of reducing class sizes by 7 pupils in STAR, and by an equal percentage amount for Sweden, is calculated as:

$$(12) \qquad PV\ of\ Costs = \sum_{t=1}^{j} C_t \,/\, (1+r)^t\ ,$$

where C_t. is the cost of reducing class size in year t and r is a real discount rate. Costs are discounted back to the initial year of class size reduction, 1998. In STAR, the experiment lasted 4 years (from kindergarten to 3rd grade), hence j=4. Since the average "small-class"-pupil in STAR spent 2.3 years in a small class, we assume full cost first and second year, and 30 percent of the full cost in the 3rd year. For Sweden, the analysis is done for 1 year (6th grade). However, since class sizes are correlated across grades, we calculate the costs for more than 1 year. As described in the text we arrive at the number 2.75 as being a likely number of more years that the average "small-class"-pupil has spent in a small class, compared to the average "regular-class"-pupil. In Sweden the "experiment" can therefore be assumed to have lasted from 1st to 6th grade, hence j=6. For Sweden we assume 46 percent (2.75/6) of the cost per pupil for each of the sixth years.

The present value of the higher earnings stream due to smaller classes is:

$$(13) \qquad PV\ of\ Benefits\ =\ \sum_{t=18-j}^{65-j} Y_t \times \beta(\delta_M + \delta_R) \,/\, (1+r)^t,$$

where Y_t represent the average real earnings (adjusted for productivity growth) each year after age 18 (until 65), ß represents the increase in earnings associated with a one standard deviation increase in (either math or reading) test scores in STAR. δ_M and δ_R represents the increase in test scores (in SD units) in math and reading due to attending smaller class in grades K-3 in STAR, and r is a real discount rate. We assume $\delta = \delta_R = \delta_M$, which earlier research has found using STAR data (see Krueger, 1999). For US, ß = 0.08 and δ=0.20. For Sweden, ß represents the increase in earnings associated with the average of a one standard deviation increase in math and a one standard deviation increase in reading test scores using the UGU material. For Sweden, ß = 0.06 and δ=0.19, so we also for Sweden assume $\delta = \delta_R = \delta_M$ which was found in STAR. This last assumption could be questioned since for Sweden we only investigated the effect of class size on math test scores. If for Sweden we would set $\delta_R = 0$, and let ß represents the increase in earnings associated with the average of a one standard deviation increase in math test scores (not controlling for reading scores) using the UGU material, the calculated benefits would however only be slightly lower (83 percent of the figures in Table 8.3). So, by using the combined test scores we inflate ß by about 20 percent. Note that $Y_t = E_t(1+\gamma)^t$, where E_t is the unadjusted average real earnings in year t, and γ is the assumed percentage productivity growth.

The internal rate of return, r*, can be calculated by solving for the discount rate that equates the benefits and costs in the following equation:

$$\sum_{t=1}^{j} C_t \,/\, (1+r^*)^t = \sum_{t=18-j}^{65-j} Y_t \beta(\delta_M + \delta_R) \,/\, (1+r^*)^t \,, \tag{14}$$

References

Achilles, C. (1999). *Let's Put Kids First, Finally: Getting Class Size Right.* Thousand Oaks, CA: Corwin Press.

Anderson, T. W. and Cheng Hsiao, “Estimation of Dynamic Models With Error Components,” *Journal of the American Statistical Association,* LXXVI (1981), 598-606.

Arellano, Manuel, “A Note on the Anderson-Hsiao Estimator for Panel Data”, *Economics Letters,* XXXI (1989), 337-341.

Barro and Lee, 1997:Schooling Quality in a cross section of countries ,NBER, Working paper 6198.

Bergström, Hans, “Om konsten att lyfta den Svenska skolan,” Artiklar och föredrag om tillståndet I och utvecklingen av den Svenska skolan, Bertil Ohlin Institutet, 1998.

Burkhead, J. 1967. *Input-Output in Large City High Schools.* Syracuse, NY: Syracuse University Press.

Card, David. 1999. “The Causal Effect of Schooling on Earnings,” forthcoming in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics (*Amsterdam: North Holland, forthcoming).

Card, David and Alan B. Krueger. 1992. “Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States.” *Journal of Political Economy* 100 (February): 1-40.

Card, David, and Alan B. Krueger. 1992. “School Quality and Black-White Relative Earnings: A Direct Assessment.” *Quarterly Journal of Economics* 107(1): 151-200.

Card, David and Alan B. Krueger. 1996. “Labor Market Effects of School Quality: Theory and Evidence.” In Gary Burtless, editor, *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington D.C.: Brookings Institution, pp. 97-140.

Card, David and Abigail Payne. 1998. “School Finance Reform, the Distribution of School Spending and the Distribution of SAT Scores.” U.C. Berkeley, Center for Labor Economics, Working Paper, forthcoming, *Journal of Public Economics.*

Cawley, John, James Heckman and Edward Vytlacil. 1999. “On Policies to Reward the Value Added by Educators.” *Review of Economics and Statistics* 81(4), 720-27.

Cohn, E., S.D. Millman, and Chew. 1975. *Input-Output Analysis in Public Education.* Cambridge, MA: Ballinger.

Cronbach, Lee J., “Coefficient alpha and the internal structure of tests,” *Psychometrika*, XVI (1951), 297-334.

Currie, Janet and Duncan Thomas. 1999. “Early Test Scores, Socioeconomic Status and Future Outcomes.” NBER Working Paper No. 6943, February.

Digest of Education Statistics, 1998 and 1999.

Edin, P.-A and Fredriksson, P. (2000). *LINDA – Longitudinal INdividual DAta for Sweden*, Uppsala University, WP 2000:19.

Finn, Jeremy D. and Charles M. Achilles. 1990. “Answers and Questions About Class Size: A Statewide Experiment.” *American Educational Research Journal* 27 (Fall): 557-577.

Finn, J. D., Gerber, S., Achilles, C. M. and Boyd-Zaharias, J. (1999). ‘Short- and long-term effects of small classes.’ SUNY Buffalo, mimeo.

Folger, J. and Breda, C. (1989). ‘Evidence from Project STAR about class size and student achievement.’ *Peabody Journal of Education*, vol. 67, pp. 17-33.

Folger, John and Jim Parker. 1990. “The Cost-Effectiveness of Adding Aides or Reducing Class Size,” Vanderbilt University, mimeo.

Fowler, W. And H. Walberg, “School size, characteristics, and outcomes,” *Educational Evaluation and Policy Analysis* 13 (2), pp. 189-202.

Fägerlind, “Utbildningen i Sverige och det mänskliga kapitalet,” rapport till Ekonomikommissionen, bilaga 9 i Nya villkor för Ekonomi och politik, 1993.

Greenwald, Rob, Larry V. Hedges, and Richard D. Laine, The Effect of School Resources on Student Achievement, *Review of Educational Research* 66(3), 1996.

Hanushek, Eric A. 1986. “The Economics of Schooling: Production and Efficiency in Public Schools.” *Journal of Economic Literature* 24 (September): 1141-1177.

Hanushek, Eric A. 1996a. "A More Complete Picture of School Resource Policies." *Review of Educational Research,* LXVI: 397-409.

Hanushek, Eric A. 1996b. "School Resources and Student Performance." In Gary Burtless, editor, *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* Washington D.C.: Brookings Institution, pp. 43-73.

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19(2): 141-164.

Hanushek, Eric A. 1998. "The Evidence on Class Size." Occasional Paper Number 98-1, W. Allen Wallis Institute of Political Economy, University of Rochester, Rochester, NY, February.

Hanushek, E. (1999). 'Some findings from the Tennessee STAR experiment and other investigations of class size reductions.' University of Rochester, mimeo.

Hanushek, Eric A. 2000. "Evidence, Politics, and the Class Size Debate." Mimeo., Hoover Institute, August.

Hanushek, Eric A., John F. Kain and Steven G. Rivkin. 1998. "Teachers, Schools, and Academic Achievement." NBER Working Paper 6691, August 1998.

Heckman, J. (1976). 'The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models.' *Annals of Economic and Social Measurement,* vol. 5, pp. 475-92.

Hedges, Larry V., Richard Laine and Rob Greenwald. 1994. "Does Money Matter? A Meta- Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Education Researcher* 23, no. 3 (April): 5-14.

Howell, Wolf, Peterson, and Campbell, "Test score effects of school vouchers in Dayton, Ohio, New York City, Washington D.C.: Evidence from randomized field trials." Program on Education Policy and Governance Research Paper, August 2000.

Härnqvist, Kjell, Ingemar Emanuelsson, Sven-Eric Reuterberg and Allan Svensson, "Dokumentation av Projektet "Utvärdering Genom Uppföljning," Rapport nr. 1994:03, Institutet för pedagogik, Göteborgs Universitet, 1994.

Jencks, Christopher S. and M. Brown. 1975. "Effects of High Schools on their Students." *Harvard Educational Review* 45(3): 273-324.

Jencks, Christopher S. and Meredith Phillips. 1999. "Aptitude or Achievement: Why Do Test Scores Predict Educational

Attainment and Earnings?" Forthcoming in Susan Mayer and Paul Peterson, eds., *Learning and Earning: How Schools Matter,* Brookings Institution Press, 1999.

Kiesling, H. J. 1967. "Measuring a Local Government Service: A Study of School Districts in New York State." *Review of Economics and Statistics* 49: 356-367.

Krueger, Alan B. 1999. "Experimental Estimates of Educational Production Functions." *Quarterly Journal of Economics* 114, no. 2: 497-532.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence From Project STAR." *Economic Journal* 111: 1-28.

Krueger, Alan B. and Diane M. Whitmore, "Would Smaller Classes Help Close the Black-White Achievement Gap?, Princeton IR-section WP, March 2001.

Lazear, Edward P. 1999. "Educational Production." NBER Working Paper No. 7349, Cambridge, MA.

Lindahl, Mikael, *Studies of Causal Effects in Empirical Labor Economics,* Dissertation, SOFI, Stockholm University, 2000.

Lindahl, 2002, Home versus School Learning: A New Approach to Estimating the Effect of Class Size on Achievement, mimeo, University of Amsterdam, March 2002.

Lindbeck, Assar, Molander, P., Persson, T., Petersson, O., Sandmo, A., Swedenborg, B. and Thygesen, N, *Nya villkor för ekonomi och politik,* SOU 1993:16.

Lindsey, J.K. and M. Cherkaoui, 1975. "Some Aspects of Social Class Differences in Achievements among 13-Year-Olds," *Comparative Education* 11(3).

Link, Charles R. and James G. Mulligan. 1986. "The Merits of a Longer School Day." *Economics of Education Review* 5(4): 373-381.

Link, Charles R. and James G. Mulligan. 1991. "Classmates' Effects on Black Student Achievement in Public School Classrooms." *Economics of Education Review* 10(4): 297-310.

Marklund, Sixten, 1962. *Skolklassens storlek och struktur,* Stockholm.

Marklund, Sixten, 1989. *Skolsverige 1950-75.* D.6. Stockholm.

Lärarnas Riksförbund, Arbetsmarknadsenkät 2000

Maynard, Rebecca and D. Crawford. 1976. "School Performance." *Rural Income Maintenance Experiment: Final Report.* Madison, WI: University of Wisconsin.

Montmarquette and Mahseredjian, 1989. Could Teacher Grading Practises Account for Unexplained Variation in School Achievements?. *Economis of Education Review 4*, 1989, p. 181-193.

Mosteller, Frederick. 1995. "The Tennessee Study of Class Size in the Early School Grades." *The Future of Children: Critical Issues for Children and Youths* 5, (Summer/Fall): 113-27.

Murnane, Richard, John Willet and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics*, LXXVII, 251-266.

National agency for Education, "Kommunernas kostnader för skolverksamheten – vad påverkar dem?," Skolverkets rapport 31, 1993.

National Agency for Education, Skolverkets rapporter 144, 170.

National Agency for Education, "Beskrivande data om skolverksamheten," 1993-1999 (skolverkets rapporter 8, 52, 75, 107, 135, 157, 173).

National Agency for Education, "Barnomsorg och skola: Jämförelsetal för huvudmän," Skolverkets rapport 189, 2000.

Neal, Derek and William Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differentials," *Journal of Political Economy* 104, October, pp. 869-95.

Nye, B., Zaharias, J., Fulton, B. D., et al. (1994). 'The lasting benefits study: a continuing analysis of the effect of small class size in kindergarten through third grade on student achievement test scores in subsequent grade levels.' Tennessee State University, Center of Excellence for Research in Basic Skills.

OECD, Education at a Glance: OECD Indicators, 2001.

Pate-Bain, H., Fulton, B. D., and Boyd-Zaharias, J. (1999). 'Effects of class-size reduction in the early grades (K-3) on high school performance.' HEROS, Inc., mimeo.

Reuterberg, S.E., & J. E. Gustafsson, 1992. Confirmatory Factor Analysis and Reliability; Testing measurement model assumptions. *Educational and Psychological Measurement*, 52, pp. 795-811.

Sengupta, J.K. and Sfeir, R.E. 1986. "Production Frontier Estimates of Scale in Public Schools in California," *Economics of Education Review*, 5 (3), pp. 297-307.

Stecher, B. M. and G. W. Bohrnstedt. 2000. *Class size reduction in California: The 1998-99 Evaluation Findings.* Sacramento, CA: California Department of Education, August.

Stern, D. 1989. “Educational Cost Factors and Student Achievement in Grades 3 and 6: Some New Evidence,” *Economics of Education Review,* 5 (1), pp. 41-48.

Summers, Anita and B. Wolfe. 1977. “Do Schools Make A Difference?” *American Economic Review,* 67 (4), pp. 649-52.

Winkler, D. 1975. “Educational Achievement and School Peer Group Composition.” *Journal of Human Resources,* 10(2), 189-204.

Word, E., Johnston, J., Bain, H., et al. (1990). ‘The state of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: technical report 1985-1990.’ Tennessee State Department of Education.