The Economic Journal, **113** (*February*), F34–F63. © Royal Economic Society 2003. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

ECONOMIC CONSIDERATIONS AND CLASS SIZE*

Alan B. Krueger

This paper examines evidence on the effect of class size on student achievement. First, it is shown that results of quantitative summaries of the literature, such as Hanushek (1997), depend critically on whether studies are accorded equal weight. When studies are given equal weight, resources are systematically related to student achievement. When weights are in proportion to their number of estimates, resources and achievements are not systematically related. Second, a cost-benefit analysis of class size reduction is performed. Results of the Tennessee STAR class-size experiment suggest that the internal rate of return from reducing class size from 22 to 15 students is around 6%.

Apart from the opportunity cost of students' time, the number of teachers hired per student is the main determinant of the economic cost of education. Looking across school districts in Texas, for example, variability in the pupil-teacher ratio accounts for two-thirds of the variability in expenditures per student.[1] If reducing class size does not increase student achievement, then variations in overall spending per pupil are unlikely to matter either because the pupil-teacher is such an important determinant of overall spending.

In a series of influential literature summaries Hanushek (1986, 1989, 1996*a, b,* 1997, 1998) argues, 'There is no strong or consistent relationship between school inputs and student performance'.[2] Although Hanushek never defines his criterion for a strong or consistent relationship, he apparently draws this conclusion from his finding that estimates included in his sample are almost equally likely to find positive effects of small class sizes on achievement as they are to find negative effects, and a majority of the estimates are statistically insignificant. Hanushek's findings are widely cited as evidence that attending a smaller class confers few benefits for students. For example, Finn and Petrilli (1998) cite Hanushek's literature summary to argue, 'there's no credible evidence that across-the-board reductions in class size boost pupil achievement'. In addition, several authors, for example, Chubb and Moe (1990), have argued that the presumed failure of the education system to convert inputs into outputs implies that incentives in public education are incapable of producing desired results.

The next section reanalyses the data in Hanushek's literature reviews, including his paper in this symposium. Hanushek's quantitative summary of the literature on class size is based on 277 estimates drawn from 59 studies. More estimates were drawn from some studies than others. Hanushek's analysis applies equal weight to every estimate, and therefore assigns much more weight to

* The author is grateful to Diane Whitmore, Mark Votruba and Michael Watts for excellent research assistance, to David Card, Victor Fuchs and three anonymous referees for helpful comments, and to Eric Hanushek for providing the data used in Section 1. Naturally, the author bears sole responsibility for the views expressed in the paper and any errors.

[1] This is based on a regression of log expenditures per pupil on the pupil-teacher ratio using data from the 1992-93 National Center for Education Statistics Common Core of Data.

[2] This quote is from Hanushek (1997, p. 148).

some studies than others. The main finding of my reanalysis is that Hanushek's pessimistic conclusion about the effectiveness of schooling inputs results from the fact that he inadvertently places a disproportionate share of weight on a small number of studies that frequently used small samples and estimated misspecified models. This problem arises because Hanushek used a selection rule that would extract more estimates from studies that analysed subsamples of a larger data set than from studies that used the full sample of the larger data set *and* because considerable discretion was exercised in the application of the selection rule.

For example, if one study analysed a pooled sample of third to sixth graders, it would generate one estimate, whereas if another study using *the same data* analysed separate subsamples of third, fourth, fifth, and sixth graders, that study would generate four estimates. Moreover, if the second study estimated separate models for black, white and Hispanic students it would 12 estimates, and if it further estimated separate regressions for mathematics and reading scores for each subsample, as opposed to the average test score, it would yield 24 estimates. As a consequence, most of the estimates were extracted from a small minority of studies. It is not uncommon for some of the estimates to be based on as few as 20 degrees of freedom. Estimates based on smaller samples are likely to yield weaker and less systematic results, other things (e.g., the level of aggregation) being equal. Moreover, a close examination of the nine studies that provided the largest number of estimates (123) in Hanushek's sample suggests that the data and econometric specifications used in the lion's share of the estimates are not capable of detecting class-size effects of reasonable magnitudes, even if they exist.[3] For example, one third of these studies estimated specifications that controlled for both expenditures per student and students per teacher, which effectively imposes that schools that have fewer students per teacher spend less money on teacher salaries or other inputs to achieve their student-teacher ratio than schools with larger student-teacher ratios; this is not the policy experiment that most observers have in mind.

Even using Hanushek's classification of the estimates – which in many cases appears to be problematic – class size is systematically related to student performance if the various studies in the literature are accorded equal weight.

A more general point raised by the reanalysis of Hanushek's literature summary is that not all estimates are created equal. Hedges *et al.* (1994) and other formal meta-analyses of class-size effects reach a different conclusion than Hanushek apparently because they combine estimates across studies in a way that takes account of the estimates' precision. Although their approach avoids the statistical pitfalls generated by Hanushek's method, it still yields uninformative results if the equations underlying the studies in the literature are misspecified. Research is not democratic. In any field, one good study can be more informative than the rest of the literature. There is no substitute for understanding the specifications underlying the literature and conducting well designed experiments. A similar point was

[3] This is not meant as a criticism of these studies. Many of them were not about class size and only control for it as an ancillary variable in a regression designed to answer another question.

made much more eloquently by Galileo some time ago (quoted from Sobel (1999, p. 93)).

> I say that the testimony of many has little more value than that of few, since the number of people who reason well in complicated matters is much smaller than that of those who reason badly. If reasoning were like hauling I should agree that several reasoners would be worth more than one, just as several horses can haul more sacks of grain than one can. But reasoning is like racing and not like hauling, and a single Barbary steed can outrun a hundred dray horses.

Insofar as sample size and strength of design are concerned, I would argue that Tennessee's Project STAR is the single Barbary steed in the class size literature. Project STAR was an experiment in which 11,600 students and their teachers were randomly assigned to small- and regular-size classes during the first four years of school. According to Mosteller (1995), for example, project STAR 'is one of the most important education investigations ever carried out and illustrates the kind and magnitude of research needed in the field of education to strengthen schools'. Research on the STAR experiment indicates that primary-school students who were randomly assigned to classes with about 12 students performed better than those who were assigned to classes with about 22 students, even when they were observed at the end of secondary school.

Section 2 considers the economic implications of the magnitude of the relationship between class size and student performance using results from Project STAR. The key economic question of, 'How big an improvement in student performance is necessary to justify the cost?' is rarely raised. A method is presented for estimating the internal rate of return from reducing class size. The estimates suggest that the (real) internal rate of return from reducing class size from 22 to 15 students is around 6%. At a 4% discount rate, the benefits of reducing class size are estimated to be around twice the cost.

1. Reanalysis of Literature Review

Hanushek kindly provided me with the classification of estimates and studies underlying his 1997 literature summary. These same data are used in his contribution to this symposium. 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.

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 function 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% of the articles and 35% of the estimates.

Table 1 summarises the distribution of the estimates and studies underlying Hanushek's literature summary. 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.[4] Nine studies contributed more than seven estimates each. These nine studies made up only 15% of the total set of studies, yet they contributed 44% of all estimates used. By contrast, the 17 studies from which one estimate was taken represented 29% of studies in the literature and only 6% 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

Table 1

Distribution of Class Size Studies and Estimates Taken in Hanushek (1997)

Number of estimates extracted (1)	Number of studies (2)	Total number of estimates (3)	Percent of studies (4)	Percent of estimates (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

Notes: Column (1) categorises 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 number of estimates used from all the studies.

[4] Many of these studies reported more than one estimate, but only one estimate was selected because the other estimates were not judged sufficiently different in terms of sample or specification. Hanushek (1997) notes that as a general rule he tried to 'reflect the estimates that are emphasized by the authors of the underlying papers'.

studies by Link and Mulligan (1986; 1991) each contribute 24 estimates – or 17% of all estimates. Both papers estimated separate models for mathematics and reading scores by grade level (3rd, 4th, 5th or 6th) and by race (black, white, or Hispanic), yielding $2 \times 4 \times 3 = 24$ estimates apiece. One of these papers, Link and Mulligan (1986), addressed the merits of a longer school day, using an 8% subsample of the data set used in Link and Mulligan (1991). In their 1986 paper, the interaction between class 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 mathematics scores, it was individually statistically insignificant. In an e-mail communication, 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.[5] This procedure would seem to violate the stated selection rule that restricted estimates to a 'set of published results'. Moreover, because the estimates in Link and Mulligan (1986) is based on a small subset of the data in Link and Mulligan (1991), the additional estimates add little independent information.

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. Likewise, 'percent minority' was the only family background variable in Sengupta and Sfeir (1986), from which eight estimates were taken. 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 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), and more flexibly than Sengupta and Sfeir (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.

Hanushek selected 11 OLS estimates from Cohn and Millman (1975), but excluded estimates that corrected for simultaneity bias. The latter estimates were consistently more positive and were the authors' preferred specification. The authors' preferences were over-ridden, however. Moreover, the OLS estimates

[5] Hanushek (2000, Appendix) explains that, 'Link and Mulligan (1986) included an ambiguous footnote about whether teacher-pupil ratio was included in all 24 equations in their paper or just 12' which prompted his contracts with Link and Mulligan. This explanation is puzzling, however, because none of the four footnotes in Link and Mulligan (1986) concerns class size, and the text is quite clear that the reference to class size refers to the 12 mathematics equations. It is also unclear why Hanushek did not try to ascertain the sign of the 24 unreported estimates.

which Hanushek selected controlled for *both* the average class size and pupil-teacher ratio in a secondary school, a clear specification error.

Summers and Wolfe (1977) provides another illustration of the type of researcher discretion that was exercised in extracting estimates. Summers and Wolfe analyse data for 627 sixth-grade students in 103 elementary schools. They mention that data were also analysed 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.[6] 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. It could be argued that studies that report models with interaction between class size and student characteristics yield multiple estimates for subgroups more efficiently than studies that estimate separate models for subsamples, yet only one estimate was taken from interactive models. Furthermore, because Hanushek (2003) 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'.

Hanushek classified six estimates from Smith (1972) as having unknown sign, which was particularly puzzling because there is no reference to estimates of the effects of class size or the pupil-teacher in Smith's paper. When I inquired, Hanushek provided the following rationale: 'Mike reports reproducing the Coleman report results, showing that pupil-teacher ratios have no effect'. While Smith reports having replicated 'most' of the Colemann Report results, he makes no specific reference to results concerning the pupil-teacher ratio. Moreover, Smith concludes that his analysis puts 'into question any findings at the secondary level about relationships between school resources and student achievement' from the Coleman Report. Other papers in Hanushek's sample did analyse the Coleman data, so it would seem unnecessary to classify six nonexistent estimates from Smith's paper as insignificant.

In a smaller number of cases, estimates were misclassified and unpublished estimates were selected. Kiesling (1967), for example, was classified as having three estimates of the effect of class size, but there is no mention of a class size variable in Kiesling's article. Hanushek informed me that Kiesling's estimates were taken from his unpublished thesis, which seems to violate his intention of only using published estimates. In Montmarquette and Mahseredjian (1989), the sign of the

[6] Their paper mentions that a full set of estimates for the additional samples was included in a Philadelphia Federal Reserve Bank publication, but this paper was not included in Hanushek's sample. Their footnote 22 also provides some description of the class-size results in the other samples.

class-size result was inadvertently reversed. I have not corrected this or other any other error that I detected because I want to emphasise that the discrepancy in results comes from the weighting of the studies. Moreover, once one starts down the road of correcting estimates there is no end. A corrected sample in my view would include different estimates from Cohn and Millman (1975), fewer estimates from Link and Mulligan (1986), no estimates from Smith (1972), some estimates from Finn and Achilles (1990), and on and on. My objective is not to derive the best data set possible, but to see how sensitive Hanushek's (1997) results are when alternative assumptions are used to aggregate his original sample. Correcting all the errors in Hanushek's selection and coding of estimates would undoubtedly weaken his conclusion that school inputs do not matter for student achievement.

1.1. *Alternatively Weighted Tabulations*

Column 1 of Table 2 summarises 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.[7] 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.

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.[8] 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 mis-specified 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 researchers in selecting many or few estimates from a particular paper. A good case could be

[7] I also follow the practice of using the terms class size and pupil-teacher ratio interchangeably. The difference is primarily a question of how one aggregates micro data.

[8] If the weights were selected to minimise the sampling variance of the combined estimate, the optimal weights would be the inverse of the sampling variances of the individual estimates; see Hedges and Olkin (1985).

Table 2

Reanalysis of Hanushek's (1997) Literature Summary of Class Size Studies

Result	Hanushek's weights (1)	Studies equally weighted (2)	Studies weighted by journal impact factor (3)	Regression-adjusted weights (4)
Positive & stat. sig. (%)	14.8	25.5	34.5	33.5
Positive & stat. insig. (%)	26.7	27.1	21.2	27.3
Negative & stat. sig. (%)	13.4	10.3	6.9	8.0
Negative & stat. insig. (%)	25.3	23.1	25.4	21.5
Unknown sign & stat. insig. (%)	19.9	14.0	12.0	9.6
Ratio positive to negative	1.07	1.57	1.72	2.06
p-value*	0.500	0.059	0.034	0.009

Notes: See text for full explanation. Column (1) is from Hanushek (1997, Table 3), and weights studies by the number of estimates that Hanushek extracted from them. Columns (2), (3) and (4) are author's tabulations based on data from Hanushek (1997). Column (2) weights each estimate by the inverse of the number of estimates taken from that study, thus weighting each study equally. Column (3) calculates a weighted average of the data in column (2), using the 'journal impact factor' as weights; articles that are not published in a journal are assigned the lowest journal impact factor. Column (4) uses the regressions in Table 3 to adjust for sample selection (see text). Table is 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 independent random draws in which positive and negative results were equally likely.

made, for example, that more estimates should have been extracted from Summers and Wolfe (1977), and fewer from Link and Mulligan, (1986, 1991). Weighting studies equally lessens the impact of researcher discretion in selecting estimates.

Figure 1 provides evidence that Hanushek's procedure of extracting estimates assigns more weight to studies with unsystematic or negative results. 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 size and student achievement. For the 17 studies from which Hanushek took only one estimate, for example, over 70% of the estimates indicate that students tend to perform better in smaller classes, and only 23% indicate a negative effect. By contrast, for the nine studies from which he took a total of 123 estimates the opposite pattern holds: small classes are associated with lower performance.

Table 3 more formally explores the relationship between the number of estimates that Hanushek extracted from each study and their results. Specifically, column (1) reports a bivariate regression in which the dependent variable is the percentage 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 in which the dependent variable is the percentage of estimates that are positive and insignificant, negative and significant, negative and insignificant, or of unknown sign, respectively. Hanushek extracted fewer estimates from studies that found positive and significant effects of

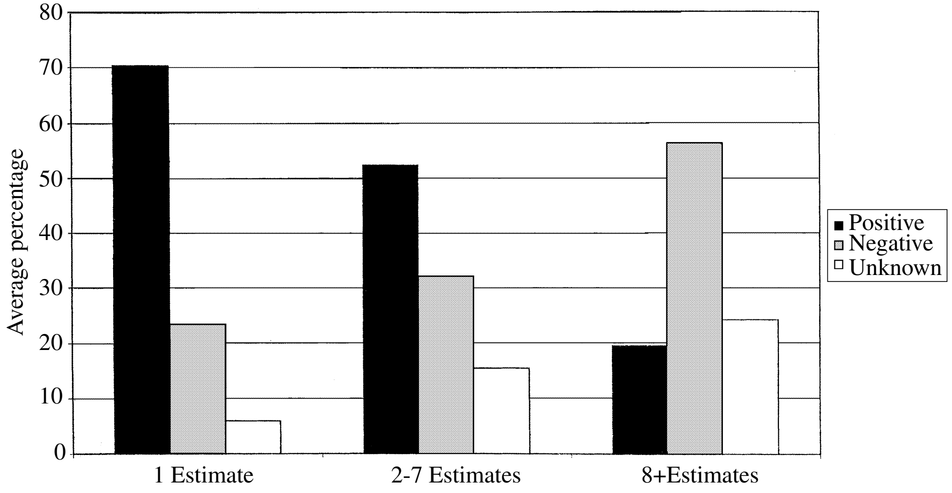

Fig. 1. *Average Percentage of Estimates Positive, Negative or Unknown Sign, by Number of Estimates Taken from Study*

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

smaller classes (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% 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 3 and Figure 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. The uneven application of the estimate selection rule could explain this pattern. Precisely to avoid the undue influence of researcher discretion in quantitative literature summaries, Stanley (2001) gives the following advice for carrying out a meta-analysis, 'In order not to give undue weight to single study, one estimate should be chosen or averaged from many comparable estimates'.

As a partial correction for the oversampling from studies with negative estimates, in column (2) of Table 2, the underlying studies – as opposed to the individual

Table 3

Regressions of Percentage of Estimates Positive or Negative, and Significant or Insignificant, on the Number of Estimates Used from Each Study; Class Size Studies

	Dependent Variable				
	% Positive & significant (1)	% Positive & insignificant (2)	% Negative & significant (3)	% Negative & insignificant (4)	% Unknown sign & insignificant (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^2	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 percentage of estimates used by Hanushek in each result category. Unit of observation is a study.

estimates extracted from the studies – are given equal weight. This is accomplished by assigning to each study the percentage 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.[9] 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% 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. Not surprisingly, the journals with the highest impact factors in the sample are the AER, QJE and JPE. 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 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 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.[10]

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

[10] 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 journals – the results are quite similar.

Another approach to adjust for estimate selection bias is to use the regressions in Table 3 to generate predicted percentages for all studies under the hypothetical situation in which one estimate was extracted from each study. This approach would be preferable to the equally-weighted-studies approach in column (2) if the marginal estimates extracted from studies are systematically different than their average estimates. Such a pattern could arise, for example, if the first estimate that each study presents is for its most relevant outcome measure (e.g., curriculum-based tests) and subsequent estimates are for less relevant outcomes (e.g., IQ, which is supposed to measure inherent intelligence). These weights could also help overcome certain forms of uneven researcher discretion in selecting estimates. A linear approximation to what the average study would find if one estimate were extracted from all studies is derived by adding together the intercept and slope in each of the regression models in Table 3.[11] These results give a prediction of the outcome that would have been obtained if one estimate had been selected from each study.

Results are reported in column (4) of Table 2. This approach for adjusting for the selection of estimates from the studies indicates even stronger and more consistently positive effects of class size. After adjusting for selection, studies with positive results are twice as likely as studies with negative results; the probability of observing at least this many studies with positive results by chance is less than one in a hundred. Among studies with statistically significant results, positive results outnumber negative results by four to one.

It should be emphasised that the results reported in Table 2 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.

1.2. *A Closer Look at Nine Studies*

Figure 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 the estimate-level survey. In view of their importance for Hanushek's conclusion, Table 4 summarises 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 describes the analysis in each of these nine studies, summarises their findings, and comments on their econometric specifications. For a variety of

[11] The intercept in column (1), for example, gives the expected percentage positive and significant if there are zero estimates. Adding the slope gives the expected percentage if one estimate is extracted per study. These expected percentages are reported in column 4.

Table 4

Summary of the 9 Studies From Which 8 or More Estimates Were Extrated

Study	Description	Hanushek Coding of Class Size Results	Comments
Burkhead (1967)	Stepwise regressions estimated using 3 school-level data sets. Chicago sample is 39 secondary-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, secondary school dropout rate, and post-secondary 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 secondary-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 school year after graduation; independent variables include pupils per teacher, expenditures per pupil, teacher pay, median income, and 4 other variables. Sample of 176 secondary schools from Project Talent; dependent variables are average 12th grade reading score, secondary school dropout rate, college attendance rate, and residuals of 12th grade reading scores from a regression on 10th grade scores; explanatory variables include 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 identify separately. 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 standardised effect size is a positive effect of smaller classes.

Table 4

Continued

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 dependent variables were used, ranging from mathematics and reading tests to percentage 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 secondary 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, percentage passing the mathematics exam, and the percentage 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 secondary schools from Project Talent data to estimate a two-step 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 school 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), careers plans (by sex); occupation (by sex); and vocabulary, social studies, reading and mathematics test.	3 neg & sig 3 neg & insig 4 unknown & 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 behaviour, and related career implications, is missed. Based on the results in Table 9, the four estimates Hanushek classified as having unknown signs all show positive effects of smaller classes on test scores.

Table 4

Continued

Study	Description	Hanushek Coding of Class Size Results	Comments
Cohn and Millman (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 at 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 determinants 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, reduced 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 seleted 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 insignificant. 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 mathematics 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, 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.

Table 4

Continued

Study	Description	Hanushek Coding of Class Size Results	Comments
Link and Mulligan (1991)	Separate OLS regression models for mathematics 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 clasmates, 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 Maintenance Experiment in IA and 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 standardised 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 programme, 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 attendance relative to enrollment) further cloud interpretation of class size effects.

Table 4

Continued

Study	Description	Hanushek Coding of Class Size Results	Comments
Sengupta and Sfeir (1986)	Sample contains 25 or 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. Mathematics, reading, and writing tests are used in two school years, yielding 12 estimates. Median sample size is 2,360 students.	9 neg & sig 3 pos & 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 variables. 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.

reasons, many of these papers provide less than compelling evidence on class-size effects. For example, Jencks and Brown (1975) analyse the effect of secondary school characteristics on students' educational attainment, but their sample is necessarily restricted to individuals who were continuously enrolled in secondary school between 9th and 12th grade. Thus, any effect of class size on secondary school dropout behaviour – 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 those in Table 4, 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).[12] In either case, the interpretation of the class-size variable in these equations is problematic. To see this, write log expenditures per student as: $EXP = TP + S$, where EXP is log expenditures per student, TP is the log teacher-pupil ratio, and S is the log of average teacher salary.[13] Assume that the 'true model' of achievement (y) is:

$$\mathrm{E}(y) = a + b\,TP + cS. \tag{1}$$

The goal is to derive unbiased estimates of b and c. The misspecified estimated model is:

$$\mathrm{E}(y) = a' + b'TP + d\,EXP. \tag{2}$$

Substituting $S = EXP - TP$ for S in (1) and rearranging terms yields:

$$\mathrm{E}(y) = a + (b - c)TP + c\,EXP. \tag{3}$$

The expected value of the estimated coefficient b', is $b - c$. If proportionate changes in the teacher-pupil ratio and teacher pay have equal effects on achievement, then b' in (2) will be zero. Clearly, this raises interpretative problems for the effect of class size if expenditures per pupil are also held constant.

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 analyse more highly aggregated data. Analyses of more highly aggregated data tend to have lower sampling variance because residual variability is averaged out. The correlation between the square root of the sample size and the number of estimates Hanushek extracted is

[12] The common use of stepwise regression in this literature is one reason why so many estimates turn up with unknown but insignificant signs – they did not make the final cut in the stepwise procedure, so they were not reported.

[13] The S variable could be interpreted more broadly as all cash outlays per classroom. The logarithmic specification simplifies the algebra and was used in many studies; interpretive problems also arise in a linear specification.

−0.24 at the school level, 0.07 at the class level, −0.10 at the grade level, −0.34 at the district level, and −0.17 at the state level.

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, for example, 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, for example, 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 and '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, for example, Sengupta and Sfeir (1986). Hanushek's selection algorithm has the effect of ignoring the effect of class size on the achievement of minority students, who often benefit the most from smaller classes.

The imprecision of the estimates in many of the papers also presents a problem. For example, the confidence interval for the change in mathematics scores associated with a reduction in class size from 22 to 15 students in Sengupta and Sfeir (1986) runs from −0.04 to 0.43 standard deviations.[14] This is wide enough to admit a large positive effect or a small negative one.

My review of the studies in Table 4 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, they 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 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 me 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.

1.3. *Value Added Studies*

In this symposium Hanushek argues that the within-state value-added studies provide the 'most refined investigation of quality'. In essence, he abandons the rest of the literature to focus on the subset of estimates he extracted from half a

[14] This is based on their equation (10), which omits the interaction between class size and the 'percent minority'. Insufficient information is reported to calculate a confidence interval for the model with interactions.

dozen studies that examine gains in students' scores. He bases this argument on the assumption that omitted state-level variables cause bias in regressions that use data at a more aggregated level. Implicitly, Hanushek's argument is that the very same states that in his view waste money on school resources like smaller classes have another set of policies that improve student achievement, creating a bias in the state-level analyses. Yet he does not specify what these policies are. He also does not report any variable that, when added to an aggregate-level regression, makes the effect of class size disappear. Indeed, several studies, including Card and Krueger (1992) and Heckman *et al.* (1995), point in just the opposite direction: when state dummy variables are added to a state-level regression, the effect of class size becomes larger. This suggests that omitted, fixed state-level variables induce a bias against finding a beneficial effect of smaller classes.

Hanushek's presumption that the value-added studies are the most informative is also challenged by Todd and Wolpin (2003) in this Feature, who show that value-added specifications are highly susceptible to bias for a number of reasons. For example, they show that even if there are no omitted variables that are correlated with class size in a value-added specification – a highly unlikely assumption – there would still be bias if test scores are serially correlated. Moreover, the value-added specification makes the untenable assumption that family inputs that affect the trajectory of student achievement are uncorrelated with class size, and that test scores are scaled in a way that makes comparison over time meaningful. Some of the value added specifications that Hanushek considers to be 'refined' also estimate highly questionable specifications. For example, Kiesling (1984), controlled for the class size and the amount of large group instruction, small group instruction, and individualised instruction in his value-added estimates. This specification allows class size to vary, but not the amount of attention students receive from the teacher.

Differencing the dependent variable could also introduce a great deal of noise, which is probably one reason why so many of the value-added studies yield imprecise estimates. The fact that the vast majority of value-added estimates – more than 80% – find statistically insignificant effects of class size does not mean that smaller classes do not help students, on average. There may be an effect, but the studies may lack power to detect it. Failing to reject the null hypothesis does not prove the null hypothesis to be true. Hanushek devotes no attention to the precision of the estimates and obscures the inference further by ignoring the sign of insignificant estimates when he tabulates the literature.

1.4. *Summing Up*

In response to work by Hedges *et al.* (1994), Hanushek (1996*b*, 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 me 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 the next Section. But the results of my reanalysis 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.

2. Economic Criterion

Hanushek (1997, p. 144) argues, 'Given the small confidence in just getting noticeable improvements [from school resources], it seems somewhat unimportant to investigate the size of any estimated effects'. The size of the effect would seem worth considering now since Hanushek's classification of *studies* in the literature does provide evidence of a systematic relationship between school inputs and student performance. Moreover, if the estimates in the literature are imprecise, they all could be statistically insignificant and unsystematic but there could nonetheless be large economic and social returns from reducing class size. The power of the estimates is relevant. Can a meaningful null hypothesis be rejected? The calculations below use the results of the Tennessee STAR experiment to illustrate the costs and benefits of reducing class size.

2.1. *Lazear's Theory of Class Size*

Before presenting benefit-cost calculations, it is useful to consider an economic model of class size. Lazear (1999) provides a model in which students who attend a smaller class learn more because they experience fewer student disruptions during class time, on average. Such a result follows naturally if the probability of a child disrupting a class is independent across children. He then quite plausibly assumes that disruptions require teachers to suspend teaching, creating a negative externality that reduces the amount of learning for everyone in the class.[15] There may be other benefits to smaller classes as well. For example, it is possible that spending time in a small class reduces a student's propensity to disrupt subsequent classes because the student learns to behave better with closer supervision or enables teachers to better tailor instruction to individual students. Nonetheless,

[15] Based on these assumptions, and the assumption that uninterrupted instruction time maps linearly into student achievement, Lazear derives a specific functional form for the education production function which is convex in class size.

Lazear's model probably captures an important feature of class size and yields a specific functional form for the education production function.

Another implication of Lazear's model is that the optimal class size is larger for groups of students who are well behaved, because these students are less likely to disrupt the class. Schools therefore have an incentive to assign weaker, more disruptive students to smaller classes. Compensatory education programmes, which automatically provide more resources to lower achieving schools, could also be viewed as targeting resources to weaker students. If schools voluntarily assign weaker students to smaller classes (as predicted by Lazear) or if compensatory funding schemes cause weaker students to have smaller classes, a spurious negative association between smaller classes and student achievement would be created. This phenomenon could explain why studies that focus on exogenous changes in class size – such as Angrist and Lavy's (1999) analysis of Maimonides' law, as well as the STAR experiment – tend to find that smaller classes have a beneficial effect on student achievement. For educational policy, it is the gain in achievement from exogenous reductions in class size from current levels that is relevant, not the relationship estimated from observed variations in class size voluntarily chosen by schools.

One final aspect of Lazear's model is worth emphasising. If schools behave optimally, they would reduce class size to the point at which the benefit of further reductions are just equal to their cost. That is, on the margin, the benefits of reducing class size should equal the cost. This implication provides a plausible economic null hypothesis. If we are starting from the optimal level, the costs and benefits of exogenous shifts in class size should be roughly equivalent.

2.2. *Benefits and Costs of Educational Resources*

Many studies suggest that education has a causal effect on earnings; see, e.g., Card (2002) for a survey. Two important benefits of improved school resources 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 school resources on achievement is most commonly measured in terms of student performance on standardised tests. This Section converts test outcome measures into dollars by using the relationship between test scores and earnings. This relationship is used to calculate the internal rate of return from reducing class size.

Three recent studies illustrate the magnitude of the relationship between students' test scores while in school and their subsequent earnings. Murnane *et al.* (1995) estimate that male high school seniors who scored one standard deviation (SD) higher on the basic mathematics achievement test in 1980 earned 7.7% higher earnings six years later, based on data from the High School and Beyond survey. The comparable figure for females was 10.9%. 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 mathematics and reading test scores at age 7 and earning at age 33.[16] They find that students who score in the upper quartile of the reading exam earn 20% more than students who score in the lower quartile of the exam, while students in the top quartile of the mathematics exam earn another 19% 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% higher earnings, while a one SD increase in the mathematics test is associated with 7.6% 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% 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 mathematics 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 mathematics or reading scores in elementary schools is associated with about 8% higher earnings.

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 realised 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 who entered kindergarten (that is, the first year of primary school) in the US in 1998. In the STAR experiment, classes were reduced from about 22 to about 15 students, so assume that funds are allocated to create $7/15 = 47\%$ more classes. Denote the cost of reducing class size in year t as C_t. The present value (PV) of the costs discounted to the initial year (1998) using a real discount rate of r is:

$$PV\ of\ Costs = \sum_{t=1}^{4} C_t/(1+r)^t. \tag{4}$$

Probably a reasonable approximation is that the cost of creating and staffing 47% more classrooms is proportional to the annual per pupil cost.[17] I assume the

[16] They estimate a multiple regression with the log of the wage as the dependent variable and indicators for the reading and mathematics scores in the bottom and lower quartile as explanatory variables. When they estimate separate regressions for men and women, they also control for father's occupation, father's education, number of children and birth order, mother's age, and birth weight. The wage gap between those who score in the top and bottom quartiles on the reading exam in these models is 13% for men and 18% for women, and on the mathematics exam it is 17% for men and 9% for women.

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

additional cost per pupil each year a pupil is in a small class equals $3,501, or 47% of $7,502, which was the nationwide total expenditures per student in 1997–98.[18] Although the experiment lasted 4 years, the average student who was assigned to a small class spent 2.3 years in a small class.[19] As a consequence, I assume the additional costs are $3,501 in years one and two, 30% of $3,501 in year three, and zero in year four. Column (2) of Table 5 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 2 illustrates the age-earnings profile for workers in 1998.[20] 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 primary school in 1998 will experience when he or she completes school and enters the labour market. Let E_t represent the average real earnings each year after age 18. Also assume that β represents the increase in earnings associated with a one standard deviation increase in either mathematics or reading test scores. The preceding discussion suggests that 8% is a reasonable estimate for the value of β. Now let δ_M and δ_R represent the increase in test scores (in SD units) due to being

Table 5

Discounted Present Value of Benefits and Costs of Reducing Class Size from 22 *to* 15 *in Grades* K-3 (1998 Dollars)

		Increase in Income ($) Assuming		
		Annual Productivity Growth Rate of:		
Discount Rate (1)	Cost $ (2)	None (3)	1% (4)	2% (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
Internal Rate of Return:		0.052	0.062	0.073

Notes: Figures assume that a 1 standard deviation increase in mathematics test scores or reading test scores in grades K-3 is associated with an 8% 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.

[18] See US Department of Education (1998), Table 169.

[19] Students spent less than four years in a small class because half of 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.

[20] 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.

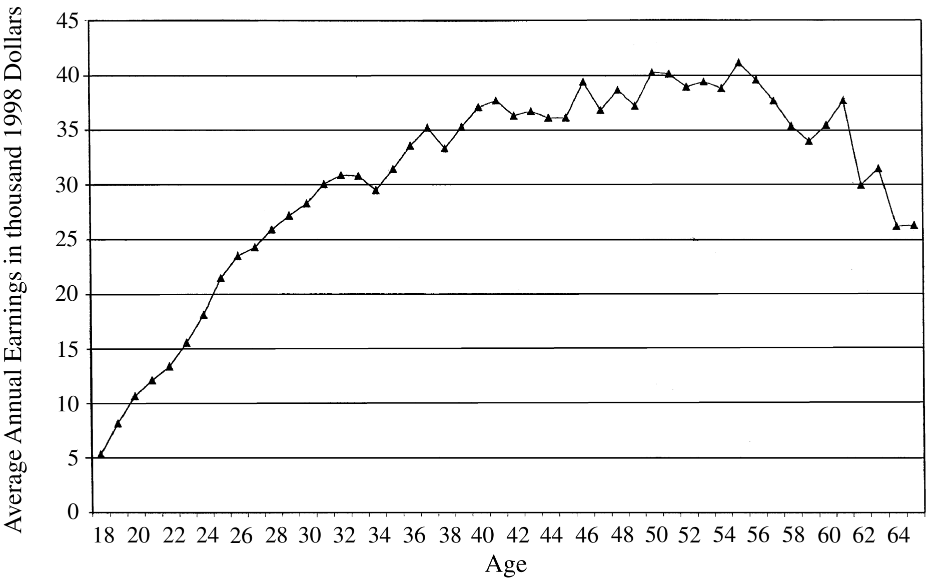

Fig. 2. *Age Earnings Profile,* 1998

assigned to a smaller class in the first four years of primary school. The STAR experiment suggests that $\delta_M = \delta_R = 0.20$ SD is a reasonable figure to use; see, e.g., Finn and Achilles (1990), Mosteller (1995) or Krueger (1999).[21] 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:

$$PV\ of\ Benefits = \sum_{t=18-4}^{61} E_t \times \beta(\delta_M + \delta_R)/(1+r)^t. \qquad (5)$$

Using these assumptions, column (3) of Table 5 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 earnings are likely to grow substantially between 1998 and when the average student who began primary school in 1998 retires. That is, when those who where in kindergarten in 1998 enter the labour market, their average real earnings will be greater than that depicted in

[21] Work by Krueger and Whitmore (1999) and Nye *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 secondary school. Although I 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 5 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.

Figure 2. Over the 20th century, real earnings and productivity have typically grown by 1% or 2% per year, roughly in step with labour productivity. The estimates of β 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% annual productivity and real wage growth after 1998.[22] The latest United States Social Security Trustees' intermediate projection is for real wages to grow by slightly less than 1% 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 US government bonds is just under 4%. If we assume an interest rate of 4% (row 3), then the benefits of reducing class size from 22 to 15 in the early grades would be 43% greater than the costs with no real wage growth, and 100% greater than the costs if real wages grow by 1% 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% and 1% annual productivity growth, the costs of reducing class size from 22 to 17 students are predicted to almost equal the benefits, in line with Lazear's prediction.

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:

$$\sum_{t=1}^{4} C_t/(1+r^*)^t = \sum_{t=14}^{61} E_t \times (0.08)(2\delta)/(1+r^*)^t. \tag{6}$$

The internal rate of return for different assumptions about productivity growth are presented in the bottom row of Table 5. If earnings grow by 1% per year, as expected by the Social Security Trustees, the internal rate of return is 6.2%.

2.3. *Caveats*

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

- The effect of test score gains on earnings in the future may turn out to be different from the value of β that was assumed. Indeed, because β was estimated from cross-section relations it could reflect the effect of omitted characteristics.[23] In addition, general equilibrium effects could affect the value of β if class size is reduced on a wide scale. It is also likely that school resources influence noncognitive abilities, which in turn influence earnings,

[22] Formally, the average real wage for a worker who reaches age A in year t, denoted Y_t, is calculated by $Y_t = E_A(1+\gamma)^t$, where E_A is the average earnings in Figure 2 for a worker of age A and γ is the rate of productivity growth.

[23] Note, however, that Jencks and Phillips (1999) find that mathematics 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.

especially for blue collar workers (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% annual growth figures probably provide a reasonable range.
- The calculations in Table 5 neglect 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 also assume that everyone works for pay, at least part year, which tends to overstate the economic benefit.
- 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.
- Inner city schools were over sampled in the STAR experiment, and as a consequence the proportion of students in the sample who are black was 32% as compared with 23% in all of Tennessee. Minority students tend to benefit more from smaller classes. However, because the sample contains few Hispanic children, the proportion of all minority students in STAR (33%) closely matches the entire US (31%).
- The cost-benefit calculations do not take into account distributional effects. Because smaller class sizes appear to generate greater benefits for economically disadvantaged students, an argument could be made that they generate positive welfare gains apart from efficiency considerations.

3. Conclusion

The method Hanushek uses to summarise the literature is often described as a 'vote counting' exercise. The results depend critically on whether the approach allows *one study, one vote*. When studies are given equal weight, the literature exihibits systematic evidence of a relationship between class size and achievement. As implemented by Hanushek, however, studies from which he extracted multiple estimates are given multiple votes. No statistical theory is presented to support this weighting scheme, and it can be misleading. For example, other things equal, studies that report a larger number of estimates for finer subsamples of a given sample will have less systematic and less significant estimates. Another reason why studies are a more natural unit of observation is that studies are accepted for publication, not estimates. The importance of a study as the unit of observation is acknowledged by the requirement that studies be published in a book or journal to assure a minimal quality check. The individual estimates that comprise a study do not pass this quality hurdle in isolation; the combined weight of evidence in a study is evaluated to decide whether to publish it. Perhaps most importantly, weighting studies equally reduces the influence of researcher discretion in selecting which estimates to include or exclude in the analysis.

A referee raised the valid question: 'Can we learn anything from meta-analysis at all?' This is a good question. At best, I think quantitative literature summarises like those relied on by Hanushek in this symposium can provide a formal representation of the research findings in a particular literature. This type of meta-analysis serves a purpose much like polling authors to assess their views of what they found. I suspect this is the reason why Hanushek (1997) reported that he tried to 'reflect the estimates that are emphasized by the authors of the underlying papers'. Thus, a meta-analysis can give a more formal accounting of what a literature has found than a qualitative literature review. If this is the goal, however, then weighting studies by the number of estimates that Hanushek extracted from them is clearly inappropriate. Equally weighting the studies would be more appropriate in this objective.

Moreover, a meta-analysis only can be as good as the underlying studies in the literature. If each of the underlying studies yields biased estimates, then their aggregate representation will be biased as well. A proper meta-analysis, however, could help to shed light more than any individual study when the statistical power of the individual studies is low. For example, if most studies use a small sample or a specification that causes estimates to be imprecise, then a proper meta-analysis could detect systematic patterns in the data even if individual studies by and large cannot. For this reason I have focused on the ratio of the number of positive to negative estimates. If estimates are insignificant, they will still tend to indicate more positive than negative effects if class size truly has a beneficial effect. By lumping all statistically insignificant estimates into one category, Hanushek obscures much of the signal that can be inferred from the studies.

I suspect the referee's concern was deeper, however. The studies are of varying quality and often examine very different outcomes for different populations. Should the amalgamation of such studies be trusted? Personally, I think one learns more about the effect of class size from understanding the specifications, data, methods and sensitivity of results in the few best studies than from summarising the entire literature; hence my earlier reference to Galileo. The quantitative summaries of the effect of class size undoubtedly impute more precision and certainty to what has been learned in the literature than is justified. Nevertheless, I think the literature as a whole certainly does not provide *prima facie* evidence that input-based school policies are a failure, as Hanushek argues in this symposium. If anything, the literature is consistent with the opposite conclusion: on average, students who attend schools with smaller classes tend have higher academic achievement. This conclusion also makes economic sense: one would not expect a free lunch. Some communities chose to spend money to reduce class size, and private schools often provide smaller classes than public schools (see Meyers *et al.* (2000) for evidence in New York City) so presumably parents feel they benefit from the extra expenditure. There must be a great deal of irrationality if smaller classes convey no benefits.

The cost-benefit calculations described in Section 2, subject to the many qualifications listed there, suggest that the internal real rate of return from a 7-student reduction in class size in the first four years of primary school is about 6%. At a 4% discount rate, every dollar invested in smaller classes yields about $2 in benefits. The 'critical effect size' – or minimum gain for the benefit of a reduction from 22 to 15 students to equal the costs – would equal 0.10 standard deviation units if productivity grows at 1% per annum and a 4% real discount rate is assumed. This would be a natural null hypothesis against which to test the findings in the literature to judge their economic significance.

One conclusion to be drawn from this reanalysis is that the literature suggests a positive effect of smaller classes on student achievement, although the effect is subtle and easily obscured if misspecified equations are estimated or small samples are used. Quantitative literature summaries can also obscure the effect of class size if some studies are given much more weight than others. 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 gains in test scores from smaller classes translate into positive benefit-cost differentials. The results of the STAR experiment suggest that the internal rate of return from lowering class size is in the neighbourhood of what would be expected from economic theory. Even subtle effects could be economically important. Anyone who expects much larger achievement gains from reducing class size is expecting extra-normal returns. Although such returns are possible, economists are usually sceptical that such large returns are available.

Princeton University and NBER

References

Angrist, J. and Lavy, V. (1999). 'Using Maimonides' rule to estimate the effect of class size on children's academic achievement', *Quarterly Journal of Economics*, vol. 114 (May), pp. 533–75.

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

Card, D. (2002). 'The causal effect of schooling on earnings', in (O. Ashenfelter and D. Card, eds.) *Handbook of Labor Economics*, North Holland: Amsterdam, ch. 30.

Card, D. and Krueger, A. B. (1992). 'Does school quality matter? Returns to education and the characteristics of public schools in the United States', *Journal of Political Economy*, vol. 100 (February), pp. 1–40.

Cawley, J., Conneely, K., Heckman, J. and Vytlacil, E. (1996). 'Measuring the effects of cognitive ability', NBER Working Paper No. 5645, July.

Chubb, J. E. and Moe, T. M. (1990). *Politics, Markets and America's Schools*, Washington, DC: The Brookings Institution.

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

Currie, J. and Thomas, D. (1999). 'Early test scores, socioeconomic status and future outcomes', NBER Working Paper No. 6943, February.

Finn, C. and Petrilli, M. (1998). 'The elixir of class size', *The Weekly Standard*, March 9, available from www.edexcellence.net/library/elixir.html.

Finn, J. D. and Achilles, C. M. (1990). 'Answers and questions about class size: a statewide experiment', *American Educational Research Journal*, vol. 27 (Fall), pp. 557–77.

Folger, J. and Parker, J. (1990). 'The cost-effectiveness of adding aides or reducing class size', Vanderbilt University, mimeo.

Fowler, W. and Walberg, H. (1981). 'School size, characteristics, and outcomes', *Educational Evaluation and Policy Analysis*, vol. 13(2), pp. 189–202.

Hanushek, E. A. (1981). 'Throwing money at schools', *Journal of Policy Analysis and Management,* vol. 1(Fall), pp. 19–41.

Hanushek, E. A. (1986). 'The economics of schooling: production and efficiency in public schools', *Journal of Economic Literature*, vol. 24 (September), pp. 1141–77.

Hanushek, E. A. (1989). 'Expenditures, efficiency, and equity in education: the federal government's role', *American Economic Review*, vol. 79(2), pp. 46–51.

Hanushek, E. A. (1996*a*). 'A more complete picture of school resource policies', *Review of Educational Research*, vol. 66, pp. 397–409.

Hanushek, E. A. (1996*b*). 'School resources and student performance', in (G. Burtless, ed.) *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, pp. 43–73, Washington DC: Brookings Institution.

Hanushek, E. A. (1997). 'Assessing the effects of school resources on student performance: an update', *Educational Evaluation and Policy Analysis*, vol. 19(2), pp. 141–64.

Hanushek, E. 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. A. (2000). 'Evidence, politics, and the class size debate', mimeo., Hoover Institute, August.

Hanushek, E. A. (2003). 'The failure of input-based schooling policies', Economic Journal, vol. 113, pp. F64–98.

Heckman, J. J., Layne-Farrar, A. and Todd, P. (1995). 'The schooling quality-earnings relationship: using economic theory to interpret functional forms consistent with the evidence', NBER Working Paper No. 5288, October.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods of Meta-Analysis*, Orlando: Academic Press.

Hedges, L. V., Laine, R. and Greenwald, R. (1994). 'Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes', *Education Researcher*, vol. 23 (April), pp. 5–14.

Jencks, C. S. and Brown, M. (1975). 'Effects of high schools on their students', *Harvard Educational Review*, vol. 45(3), pp. 273–324.

Jencks, C. S. and Phillips, M. (1999). 'Aptitude or achievement: why do test scores predict educational attainment and earnings?' in (S. Mayer and P. Peterson, eds.), *Earning and Learning: How Schools Matter*, Washington DC: Brookings Institution Press.

Kiesling, H. J. (1967). 'Measuring a local government service: a study of school districts in New York state', *Review of Economics and Statistics*, vol. 49, pp. 356–67.

Kiesling, H. J. (1984). 'Assignment practices and the relationship of instructional time to the reading performance of elementary school children', *Economics of Education Review*, vol. 3(4), pp. 341–50.

Krueger, A. B. (1999). 'Experimental estimates of educational production functions', *Quarterly Journal of Economics*, vol. 114(2), pp. 497–532.

Krueger, A. B. and Whitmore, D. (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, vol. III, pp. 1–28.

Lazear, E. P. (2001). 'Educational production', *Quarterly Journal of Economics*, vol. 116(3), pp. 777–803.

Link, C. R. and Mulligan, J. G. (1986). 'The merits of a longer school days', *Economics of Education Review*, vol. 5(4), pp. 373–81.

Link C. R. and Mulligan, J. G. (1991). 'Classmates' effects on black student achievement in public school classrooms', *Economics of Education Review*, vol. 10(4), pp. 297–310.

Maynard, R. and Crawford, D. (1976). 'School performance', *Rural Income Maintenance Experiment: Final Report*. Madison, WI: University of Wisconsin.

Meyers, D., Peterson, P., Mayer, D., Chou, J., and Hou, W. (2000). 'School choice in New York after two years: an evaluation of the school choice scholarships program', Interim Report, mimeo., Mathematica Policy Research, Princeton, NJ.

Montmarquette C. and Mahseredjian S. (1989). Does school matter for educational achievement? A two-way nested-error components analysis', *Journal of Applied Econometrics*, vol. 4(2), pp. 181–93.

Mosteller, F. (1995). 'The Tennessee study of class size in the early school grades', *The Future of Children: Critical Issues for Children and Youths*, vol. 5, (Summer/Fall), pp. 113–27.

Murnane, R., Willet, J. and Levy, F. (1995). 'The growing importance of cognitive skills in wage determination', *Review of Economics and Statistics*, vol. 77, pp. 251–66.

Neal, D. and Johnson, W. (1996). 'The role of premarket factors in black-white wage differentials', *Journal of Political Economy*, vol. 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', Seventh grade technical report, Nashville: Center of Excellence for Research in Basic Skills, Tennessee State University.

Sengupta, J. K. and Sfeir, R. E. (1986). 'Production frontier estimates of scale in public schools in California', *Economics of Education Review*, vol. 5(3), pp. 297–307.

Smith, M. (1972). 'Equality of educational opportunity: the basic findings reconsidered', in (F. Mosteller and D. P. Moynihan, eds.), *On Equality of Educational Opportunity*, New York: Random House, pp. 230–342.

Sobel, D. (1999). *Galileo's Daughter: A Historical Memoir of Science, Faith, and Love*, New York: Walker & Co.

Stanley, T. D. (2001). 'Wheat from chaff; meta-analysis as quantitative literature review', *Journal of Economic Perspectives*, vol. 15(3), pp. 131–50.

Stern, D. (1989). 'Educational cost factors and student achievement in grades 3 and 6: some new evidence', *Economics of Education Review*, vol. 5(1), pp. 41–8.

Summers, A. and Wolfe, B. (1977). 'Do schools make a difference?', *American Economic Review*, vol. 67(4), pp. 649–52.

Todd, P. E. and Wolpin, K. I. (2003). 'On the specification and estimation of the production function of cogitive achievement', Economical Journal, vol. 113, pp. F3–33.

US Department of Education (1998). *Digest of Education Statistics 1998*, National Center for Education Statistics, NCES 1999-036, Washington DC.