Evidence, Politics, and the Class Size Debate

Eric A. Hanushek[1]

August 2000

Abstract

While the debate over class size reduction is largely a political one, some recent discussion has focused on the evidentiary base. Krueger (2000) provides two re-analyses that tend to support reduction policies: a re-interpretation of existing econometric evidence and a proposed demonstration that the small expected effects of class size reduction policies are worthwhile. Unfortunately, as shown here, neither is very convincing. First, a review of past experience from large-scale class size reduction shows no discernible effects on student achievement. Second, while Krueger's re-interpretation of the econometric evidence suggests more positive outcomes than prior reviews, his results come from placing much larger weight on low quality statistical estimates. Third, reliance on results from Project STAR does not change the policy conclusions. Finally, his benefit-cost calculations presume that more efficacious policies such as those operating on teacher quality are not feasible and that the only alternative is the current system. His demonstration that the small effects suggested by Project STAR might still be good policy relies on a series of heroic assumptions and entirely neglects any of the uncertainty in his illustrative calculations.

[1] Stanford University, University of Texas at Dallas, and National Bureau of Economic Research. Helpful comments were provided by John Kain, Alan Krueger, Steve Landsburg, Ed Lazear, Terry Moe, Paul Peterson, Macke Raymond, and Steve Rivkin. Support was provided by the Smith Richardson Foundation and the Packard Humanities Institute.

Evidence, Politics, and the Class Size Debate

Eric A. Hanushek

With the suddenness of a summer storm, politics thrust the issue of class size policy onto the national agenda. Before the political popularity to voters of reductions in class size became known, most educational researchers and policy makers had discarded such policies as both too expensive and generally ineffective, leaving only teachers unions and other with clear vested interests in the policies to support such ideas. When the political appeal of class size reductions became known – largely through the reactions to the 1996 California policies – there was a scramble to backfill evidence supporting such policies. In this current environment, the evidence about the effectiveness of class size reduction has been thoroughly spun in the political debate in order to match the preconceived policy proposals, making it difficult to conclude that the debate has been guided very much by the evidence.

This political backdrop is necessary to understand the significance of Alan Krueger's reanalysis of the existing evidence on class size (Krueger 2000). He focuses attention directly on the scientific evidence and its implications for policy, thus attempting to move the policy debate away from pure politics and toward a better basis for decision-making. While he offers no new evidence on the effects of class size on student performance, he contributes two different analyses that point toward a more aggressive policy of class size reduction: a massaging of the econometric evidence on effectiveness of class size reduction and of overall spending and a proposed demonstration that small outcome effects are still worthwhile. Upon careful inspection, however, neither is convincing. Nonetheless, policy makers should not ignore the emphasis on the importance of a solid evidentiary base.

Because supporters of class size reductions are likely to be attracted to his defense of such policies, it is important to understand the nature and substance of his analysis. First, his discussion omits mention of the long history and dismal results of class size policies. Second, his analysis of the existing econometric evidence derives its results from giving excessive weight to

low quality and biased estimates. Third, the discussion of the Tennessee STAR experiment does not make clear its limited evidence for any broad reductions and fails to indicate the uncertainty surrounding the results and their policy implications. Finally, the calculation of benefit-cost relationships takes a very narrow view of potential policies and requires a number of heroic assumptions. This set of comments discusses each of these in turn.

The issue of course is not whether there exists any evidence that class size reduction *ever* matters. Surely class size reductions are beneficial in specific circumstances – for specific groups of students, subject matters, and teachers. The policy debates, driven by the politics of the situation, do not, however, attempt to identify any such specific situations but instead advocate broad reductions in class sizes across all schools, subjects, and often grades. The missing elements are three. First, nothing in the current decision process encourages targeting class size reductions to situations where they are effective. Second, class size reductions necessarily involve hiring more teachers, and teacher quality is much more important than class size in affecting student outcomes. Third, class size reduction is very expensive, and little or no consideration is given to alternative and more productive uses of those resources.

The appeal of class size reduction is that it offers the hope of improving schools while requiring no change in the existing structure. Politicians can take credit for pursuing identifiable policies aimed at improving student outcomes. Teachers and other school personnel see added resources coming into schools without pressures to take responsibility for student performance and see these policies increasing the demand for teachers. The missing element is any reasonable expectation that these policies will significantly improve student achievement.

The history of class size reduction

Perhaps the most astounding part of the current debates on class size reduction is the almost complete disregard for the history of such policies. Pupil-teacher ratios fell dramatically

throughout the 20th century.[1] Table 1 shows that pupil-teacher ratios fell by a third between 1960 and 1995 – exceeding the magnitude of policy changes that most people are talking about today. With such substantial changes, one would expect to see their effect in student performance. Yet it is impossible to detect any overall beneficial effects that are related to these sustained increases in teacher intensity.

The longest general data series on student performance, albeit imperfect, is the Scholastic Aptitude Test (SAT). Figure 1 displays the relationship between pupil-teacher ratios and SAT scores. While there is a relationship between the two, it goes in the opposite direction expected: reductions in pupil-teacher ratios are accompanied by falls in the SAT, even when appropriately lagged for the history of schooling experience for each cohort of students. Because the SAT is a voluntary test taken by a select population, a portion of the fall undoubtedly reflects changes in the test-taking population instead of real declines in aggregate student performance, but there is general consensus that real declines also occurred (Congressional Budget Office, 1986).

A better indicator of performance is the National Assessment of Educational Progress (NAEP). While tracking a representative sample of students, scores are only available since the early 1970s (after a period of substantial decline as measured by the SAT). Figure 2 plots NAEP scores for 17-year-olds.[2] Math and reading show flat performance from earliest testing through 1996, while the comparable science and writing scores have declined significantly.[3] Thus, the consistent picture from available evidence is that the falling pupil-teacher ratios (and commensurately increasing real spending per pupil) have not had a discernible effect on student achievement.

---

[1] Pupil-teacher ratios are not the same as class size because of specialist teachers, differences between numbers of classes taken by students and numbers taught by teachers, and other reasons. Nonetheless, because class size and pupil-teacher ratios tend to move together over time (see Lewit and Baker 1997) and because Krueger disregards any such distinctions, these differences are not highlighted at this time. See also Hanushek 1999a.

[2] The NAEP has shown larger changes over time in the scores for 9- and 13-year-olds, but this has not been translated into improved scores at the end of high school; see Hanushek (1998) for further discussion.

[3] Writing scores are first available in 1984. The mid-1980s saw a narrowing of the racial gap in achievement, but this stopped by 1990 and cannot be readily attributed to overall resource patterns. Further discussion of the aggregate trends including the racial trends can be found in Hanushek 1999a.

While it is generally difficult to infer causation from aggregate trends, these data provide a strong *prima facie* case that the policies being discussed today will not have the significant outcomes that are advertised. The complication with interpreting these trend data is that other factors might work to offset an underlying beneficial effect. On this, the available evidence does not indicate that the pattern of test scores simply reflects changing student characteristics. Child poverty and the incidence of children in single parent families – factors that would be expected to depress achievement – have risen. At the same time, the increases in parental education and the fall in family sizes would be expected to produce improvements in student performance. Netting out these effects is difficult to do with any precision, but the existing analysis suggests little aggregate effect from the changing student backgrounds, and possibly a small net improvement.[4]

Table 1 also shows the significant increases in expenditure per pupil that have occurred over this period. A significant part of the increase in expenditure can be directly attributable to declines in the pupil-teacher ratio (Hanushek and Rivkin 1997), but other "improvements" such as having a more experienced and educated teacher force also contribute. Again, however, a comparison of student performance with the increases in inflation-adjusted expenditures of over 75 percent between 1970 and 1995 gives no reason to believe that more of the past resource policies will be successful.

If past declines in class size have had no discernible effect on student outcomes, why should we believe that future declines would yield any different results?

Econometric Evidence

Krueger (2000) concentrates most of his attention on the existing econometric evidence. While worrying about important issues, the analysis actually involves a set of calculations that

[4] The analysis by Grissmer et al. (1994) attempts to aggregate these changes over time based on econometric estimates of how various family backgrounds affect achievement. This analysis indicates that the overall preparation of white students (based on family background factors) seems to have improved, while that for Black students seems to have worsened. While considerable uncertainty surrounds the estimation approach, the analysis strongly suggests that changing backgrounds are not masking the effects of school resource increases. A critique of the methodology is found in Hanushek (1999a).

places the heaviest weight on lower quality estimates. By doing so, he is able to suggest that the overall conclusions about class size policies should change. If, however, more weight is placed on higher quality estimates, the overall conclusion about a lack of clear relationship between class size and student performance is strengthened.

The starting point of Krueger's work is my prior tabulations of the estimated relationship between teacher-pupil ratios on student performance, as reproduced in Table 2.[5] The 277 separate estimates of the class size relationship are found in 59 publications, all of the available analyses through 1994. (Issues about the underlying data raised by Krueger (2000) do not change any of the results, and a discussion of them is included in the appendix to these comments).

Among the statistically significant estimates – the ones for which we reasonably confident that there is truly a relationship – 14 percent indicate that raising the teacher-pupil ratio would have the "expected" positive relationship while an equal percentage indicate just the opposite. The statistically insignificant estimates – those for which we have less confidence that they indicate any real relationship – are almost evenly split between beneficial and adverse effects. Thus, the overall evidence provides little reason to believe that a general policy of class size reduction would improve student performance.

Krueger questions these conclusions by arguing that individual publications that include more separate estimates of the impact of class size on performance are lower in quality than those publications that include fewer estimates.[6] His hypothesis is that publications including more estimates will involve splitting the underlying samples of student outcomes, say by race or grade level. Statistical theory indicates that, other things being equal, smaller samples will yield less

[5] These tabulations were corrected for the previous miscoding of one article (Montmarquette and Mahseredjian 1989) that was pointed out to me by Alan Krueger. Krueger's analysis and tables of estimation results, however, do not adjust for this miscoding. A description of the criteria for inclusion is found in Hanushek (1997) and is summarized in Krueger (2000).

[6] His discussion leads to some confusion in nomenclature. For reasons sketched below, my previous analyses have referred to distinct estimates as "studies" even though more than one estimate might appear in a given publication. Krueger changed this language by instead referring to separate publications as studies. Here I will generally drop the term studies and use the nomenclature of separate estimates in each publication.

precise estimates than larger samples. He then jumps to the logically incorrect conclusion that publications with more individual estimates will tend to have fewer observations and thus will tend to produce statistically insignificant results when compared to those publications with fewer separate estimates.

There is no clear relationship between the sample sizes underlying individual estimates and the number of estimates in each publication. Table 3 shows the distribution of sample sizes for the 277 estimates of the effect of teacher-pupil ratios from Table 2. While highly variable, publications with the fewest estimates do not systematically have the largest sample sizes. The simple correlation of sample sizes and number of articles in the underlying publications is slightly positive (0.07), although insignificantly different from zero.

Before considering the precise nature of Krueger's re-analysis, it is useful to understand better the structure of the underlying estimates and publications. The explanation for varying numbers of estimates across individual publications is best made in terms of the provision of logically distinct aspects of the achievement process. For example, few people argue that the effects of class size reduction are constant across all students, grades, and subject matter. Therefore, when the data permit, researchers will typically estimate separate relationships for different students, different outcomes, and different grades. In fact, the analysis of the Tennessee class size experiment in Krueger (1999) divides the estimates by race and economic status, because Krueger himself thought it was plausible that class size has varying impacts – something that he finds and that he argues is important for policy. He further demonstrates varying effects by grade level. If there are different effects for different subsamples of students, providing a single estimate across the subsamples, as advocated by Krueger (2000) and described below, is incorrect from a statistical point of view and would lead to biased results.

Even if class size differences have similar effects across students, districts, and outcomes, it is often impossible to combine the separate samples used for the obtaining the individual estimates. For example, the publication by Burkhead et al. (1967) that Krueger holds up as an

example of multiple estimates for small samples presents a series of estimates for high school performance in different cities where outcomes are measured by entirely different instruments. There is no way in which these can be aggregated into a single estimate of the effect of class size. Of the 59 publications from Table 2 that include estimates of the effects of teacher-pupil ratio, 34 include two or more separate test measures of outcomes (e.g., reading and math) and 15 of these further include two or more separate nontest measures (e.g., college continuation, dropouts, or the like). For 14 of the 59 publications, the separate estimates of pupil-teacher effects within individual publications include students separated by more than three grade levels, implying not only different achievement tests but also the possibility of varying effects across grades. No general procedure exists for aggregating these separate effects in a single econometric estimate.

Thus, while Krueger suggests that the publication of multiple estimates is largely whimsical and misguided, the reality is that there are generally sound econometric reasons behind many of these decisions. The typical publication with several estimates actually provides more evidence than would be the case if only one estimate per publication were reported.

Krueger's hypothesis, however, is that an estimate in publications with more than one estimate provides poorer information than an estimate from a single-estimate publication. His analytical approach involves adding up the underlying estimates in alternative ways – effectively giving increased weight to some estimates and decreased weight to others. Specifically, he calculates the proportion of estimates within each publication that fits into the outcome categories (columns) in table 2 and adds them up across the 59 separate publications, i.e., weighting by individual publications instead of individual estimates of the effect of class size on student performance. Surprisingly, this procedure leads to stronger support for the existence of positive effects from class size reduction, even though the simple statistical theory outlined by Krueger suggests that only the confidence in the estimates and not the direction of the relationship should be affected. The evidence based on the estimates in Table 2 indicates an essentially identical chance of finding increased teacher-pupil ratios to be beneficial as a chance of being harmful; i.e.,

no systematic relationship between class size and student outcomes. When re-weighted, however, Krueger finds beneficial effects to be noticeably more likely.

Note, however, that still only 25 percent of the time would there be much confidence that there is a relationship between teacher-pupil ratios and achievement as indicated by their being a statistically significant and positive estimate. To reach his conclusions of different overall results, Krueger tends to emphasize the proportion of estimates that are positive (beneficial) versus negative (detrimental). This summary has a major problem. The equal weighting of statistically significant estimates (those more precisely estimated) and statistically insignificant estimates (less precisely estimated) seems to violate the basic premise of his re-weighting.[7] A more accurate picture of the impact of his weighting is seen in figure 3, which graphs the proportion of results that are statistically significant (positive or negative) and that are statistically insignificant. His re-weighting produces a somewhat higher proportion of positive and statistically significant results, but it does not reverse the overall picture of little reason to expect much if any impact from reducing class size.[8]

To deal with the apparent anomaly of finding different results when re-weighted, Krueger introduces a "theory of refereeing" for scholarly publications. He suggests that, whenever an author finds results that are statistically insignificant or that have the wrong sign, referees will insist that the author re-do the estimates by disaggregating them – in effect producing more of the insignificant or wrong-signed estimates.

While Krueger provides no evidence for his theory of refereeing, many – including Krueger himself – have argued just the opposite about the publication process. Specifically, there is a well-known publication bias toward having too many statistically significant estimates in

[7] This analysis also ignores statistically insignificant estimates for which the estimated sign is unknown, a condition making it impossible to know how to include them in the calculation. His analysis assumes that there is no information in analyses that drop further consideration of pupil-teacher ratios after an initial investigation.

[8] This graph plots the Krueger results that do not correct the coding of Montmarquette and Mahseredjian (1989).

articles that get published. Articles with insignificant estimates or incorrect signs simply do not get published with the same frequency as articles containing significant estimates of the expected sign (Hedges 1990). Krueger's own argument in discussing the literature on the minimum wages is "reviewers and editors have a natural proclivity to look favorably on studies that report statistically significant results" (Card and Krueger, 1995, p. 186).

Krueger is correct about the importance of quality of the estimates in formulating overall conclusions, and consideration of quality provides a much more natural and persuasive explanation for his altered results than does his theory of refereeing. The basic tabulation of results produced in Table 2 provided information on all available estimates of the effects of class size and of spending. The complete data are displayed not as an endorsement of uniform high quality but as a base case where there can be no possibility that selection of specific estimates and publications drives the results. At the same time, the underlying analyses clearly differ in quality, and – as discussed in Hanushek (1997) – these differences have the potential for biasing the results of the estimation. Two elements of quality are particularly important. First, education policy in the United States is made primarily by the separate 50 states, and the variations in spending, regulations, graduation requirements, testing, labor laws, and teacher certification and hiring policies are large. These important differences – which are also the locus of most current policy debates – imply that any analyses of student performance across states must include descriptions of the policy environment of schools or else they will be subject to standard statistical bias problems; i.e., they will tend to obtain estimates that are systematically different from reality. Second, education is a cumulative process going across time and grades, but a majority of estimates consider only the current resources available to students in a given grade. For example, when looking at performance at the end of secondary schooling, many analyses rely on just the current teachers and school resources and ignore the dozen or more prior years of inputs. Obviously, current school inputs will tend to be a very imperfect measure of the resources that went into producing ending achievement.

While judgments about study quality generally have a subjective element, it is possible to make an initial cut based on occurrence of these two problems. We begin with the issue of not measuring the state policy environment. If, as most people believe, states vary in important aspects of education policy and school operations, ignoring this in the econometric estimation will generally lead to biased estimates of the effect of teacher-pupil ratios or other resources.[9] The key is separating the effects of teacher-pupil ratios from other attributes of schools and families, and this generally cannot be done accurately if the other factors are not explicitly considered. Whether the estimates tend to find too large or too small effect of teacher-pupil ratios depends on the correlation of the omitted state regulatory and finance factors and class size (or spending).

The existing estimates contained in Table 2 can be used to identify the importance of biases caused by omitting consideration of differences in the state policy environment for schools. Specifically, an analysis that looks at schools entirely contained within a single state will observe a policy environment that is largely constant for all schools – and thus the econometric estimates that compare schooling entirely within a single state will not be biased. On the other hand, an analysis that considers schools in multiple states will produce biased results whenever important state differences in policy are correlated with differences across states in pupil-teacher ratios or overall resources. Moreover, the statistical bias will be largest for investigations relying on aggregate state data as opposed to observations at the classroom or school level.[10]

Thus, one clear measure of study quality is that it relies upon data entirely within a single state. For those using multistate data, estimates derived from the most aggregated data will be

[9] When important factors are omitted, estimates of the effect of varying teacher-pupil ratios will be unbiased only if there is no relationship across states between the quality of state policies and the average teacher-pupil ratio in the states. If on the other hand states with favorable education policies tend generally to have smaller classes, the estimates of teacher-pupil ratios will tend to differ systematically from the true effect of class size differences.

[10] Hanushek, Rivkin, and Taylor (1996) demonstrate that any bias in the estimated parameters will be exacerbated by aggregation of the estimation sample. For example, 11 of the 277 estimates of the effects of teacher-pupil ratios come from highly aggregated performance and resource data measured at the state level, the level of measurement where policy information is omitted from the analyses.

lower quality than those relying on observed resources and outcomes at the classroom or school level.

Table 4 provides a tabulation of the prior econometric results that is designed to illuminate the problem of ignoring the large differences in school organization and policy across states. The prior tabulation of all estimates shows that those with significant negative estimates evenly balance the percentage indicating teacher pupil ratios with significant positive estimates. But Table 4 shows that this is not true for estimates relying upon samples drawn entirely within a single state, where the overall policy environment is constant and thus where any bias from omitting overall state policies is eliminated. For single state analyses, the statistically significant effects are disproportionately negative (18 percent negative versus 11 percent positive). Yet, when the samples are drawn across states, the relative proportion positive and statistically significant rises. For those aggregated to the state level, almost two-thirds of the estimates are positive and statistically significant. The pattern of results also holds for estimates of the effects of expenditure differences (which are more likely to come from highly aggregate investigations involving multiple states).[11] Again, the vast majority of estimates are statistically insignificant or negative in sign except for those employing aggregated state-level data and neglecting differences in state policy environments. This pattern of results is consistent with expectations from considering specification biases when favorable state policies tend to be positively correlated with resource usage, i.e., when states with the best overall education policies also tend to have larger teacher-pupil ratios.

The second problem is that the cumulative nature of the educational process means that relating the level of performance at any point in time just to the current resources is likely to be misleading. The mismeasurement is strongest for any children who changed schools over their career (a sizable majority in the U.S.) but also holds for students who do not move because of

[11] Expenditure analyses virtually never direct analysis at performance across different classrooms or schools, since expenditure data are typically available only at the district level. Thus, they begin at a more aggregated level than many investigations of real resources.

variations over time in school and family factors. While there is no general theoretical prediction about the biases that arise from such mismeasurement, its importance can be understood by concentrating on estimates that do not suffer from the problem. The standard econometric approach for dealing with this is the estimation of value-added models where the statistical estimation is restricted to the growth of achievement over a limited period of time (where the flow of resources is also observed). By concentrating on achievement gains over, say, a single grade, it is possible to control for initial achievement differences (which will be determined by earlier but generally unobserved resources and other educational inputs).

Table 5 displays the results of teacher-pupil ratio estimates that consider value-added models for individual students. The top panel shows all such results, while the bottom panel follows the earlier approach of concentrating just on estimates within an individual state. With the most refined investigation of quality in the bottom panel, the number of estimates gets quite small and selective. In these, however, there is essentially no support for a conclusion that higher teacher-pupil ratios improve student performance. Only one of the available 23 estimates shows a positive and statistically significant relationship with student outcomes.

As noted previously, teacher-pupil ratios and class size and class size are not the same measure, even though they tend to move together. The general estimation in Table 2 makes no distinction between the two measures. In the case of estimation at the individual classroom (the focus of Table 5), however, teacher-pupil ratio is essentially the same as class size. Thus, those measurement issues cannot distort these results.

This direct analysis of study quality shows why Krueger gets different effects from weighting results by publication instead of by individual estimates. From Table 2, 17 of the 59 publications (29 percent) contained a single estimate of the effect of teacher-pupil ratio – but these estimates are only 6 percent of the 277 total available estimates. Krueger wants to increase the weight on these 17 estimates (publications) and commensurately decrease the weight on the remaining 260 estimates. Note, however, that over forty percent of the single-estimate

publications use state aggregate data, compared to only four percent of all estimates.[12] Relatedly, the single-estimate publications are more likely to employ multistate estimates (which consistently ignore any systematic differences in state policies) than the publications with two or more estimates. Weighting by publications rather than separate estimates heavily weights low quality estimates.

The implications are easy to see within the context of the two publications that Krueger himself contributes (Card and Krueger, 1992a 1992b). Each of these state-level analyses contributes one positive, statistically significant estimate of the effect of teacher-pupil ratios. Weighting by all of the available estimates, these estimates represent 0.7 percent of the available estimates, but, weighting by publications as Krueger desires, they represent 3.4 percent of the results. Krueger (2000) goes on to say that Card and Krueger (1992a) "presented scores of estimates for 1970 and 1980 Census samples sometimes exceeding one million observations. Nonetheless, Hanushek extracted only one estimate from this study because only one specification included family background information." This statement is quite misleading, however. While the underlying census data on earnings included over a million observations, the relevant estimate of the effects of class size in Card and Krueger (1992a) relies on *just 147 state aggregate data points* representing different time periods of schooling.[13]

Krueger's statement also implies that requiring information on family backgrounds is some sort of irrelevant technicality. There are, however, very important econometric reasons for insisting on the inclusion of family background as a minimal quality requirement. It is well known that family background has a powerful effect on student performance (see, for example, Coleman et al. (1966) or Hanushek (1992)). If this factor is omitted from the statistical analysis,

[12] In fact, using aggregate state data frequently precludes any consideration of different effects by student background, subject matter or what have you – offering an explanation for why these publications have just one estimate.

[13] While estimating some models with over a million observations, none is relevant for this analysis because each with a large sample fails to meet the eligibility criteria related to separating family background effects from correlated school resources (see below). In simple statistical terms, large samples cannot make up for estimating incorrectly specified relationships.

the estimates of pupil-teacher ratios can no longer be interpreted as the effect of class size might have on student performance. These estimates will be biased if there is any correlation across states between family backgrounds, such as income and education, and the average teacher-pupil ratio in the state. Considering estimates that do not take family backgrounds into account would be a very significant quality problem, almost certainly leading to larger distortions than considering estimates that do not consider the state policy environment.

In fact, Card and Krueger (1992b) was mistakenly included in the tabulations. Discussions with Krueger about the coding of the full set of estimates made it clear that this publication failed to take any aspect of family background into account, so it cannot adequately distinguish school effects from family effects on learning.[14]

Finally, the Card and Krueger (1992a) analysis suffers not only from the biases of aggregate, cross-state analysis discussed previously but also from another set of fundamental shortcomings. They estimate state differences in the value of additional years of schooling according to 1980 census information on labor market earnings and the state where workers were born (assumed to proxy where they were educated). They then relate the estimated value of a year of schooling to characteristics of the average school resources in the state in the years when a worker of a given age would have attended school. As critiques by Speakman and Welch (1995) and Heckman, Layne-Farrar, and Todd (1996a, 1996b) show, their estimates are very sensitive to the specific estimation procedure. Moreover, the state earnings differences cannot be interpreted in terms of school quality differences in the way that Card and Krueger interpret them. In order to obtain their estimates of school quality, Card and Krueger (1992a) must assume that the migration of people across states is random and not based on differential earnings opportunities. Heckman, Layne-Farrar, and Todd (1996a, 1996b) show that there is selective

[14] The concern is that family background merely proxies for attributes of the family, and, if family background is omitted from the analysis, the estimated effect of pupil-teacher ratio will not indicate the causal impact of differing pupil-teacher ratios. While the analysis in Card and Krueger (1992b) stratifies by race or includes a race dummy variable, the estimated effects come from variations across states and over time in class size, when race is not observed to vary.

migration and that this fundamental requirement for their interpretation is untrue.[15] Statistical shortcomings such as these can be identified in other estimates, but this example illustrates why the mechanical re-weighting proposed by Krueger (2000) can in fact push the results in a biased direction.

The alternative weighting methods of Krueger (2000) provide no better adjustments for anything that looks like quality of estimates. The two Card and Krueger articles are heavily cited in other articles, so that their combined weight increases to *17 percent of the total evidence* on a citation basis. But again this new weighting does not give an accurate estimate of the quality of the underlying estimates.[16] Similarly, the "selection-adjusted" weights place more emphasis on a positive and significant estimate if there was an estimated higher probability of getting a positive and significant estimate in an article (based solely on the number of estimates within each publication). The rationale behind this novel approach is entirely unclear and has no statistical basis.

In sum, Krueger's reanalysis of the econometric evidence achieves different results by emphasizing low quality estimates. The low quality estimates are demonstrably biased toward finding significant positive effects of class size reduction and of added spending. Remarkably, even when re-weighted, the support of overall class size reduction policies remains weak. Most of the estimates, no matter how tabulated, are not statistically different from zero at conventional levels.

---

[15] They also show that the results differ significantly across time and that they are very sensitive to the precise specification of the models. Speakman and Welch (1995) further show that virtually all of the effects of state school resources work through earnings of college attendees, even though the resource measures relate only to elementary and secondary schools.

[16] Card and Krueger (1992a) is rightfully cited for its innovative combination of labor market data with school quality data. However, because it has been controversial, it is cited in other works (such as Heckman, Layne-Farrar and Todd 1996a, 1996b) without endorsing its quality. A large number of citations are also of two different types. The first is its use in introductory material to justify a new set of estimates as in: 'while the common view is that resources do not matter, Card and Krueger find that they do.' The second use is by other researchers who are looking to justify use of expenditure data in a different kind of analysis, say of school choice or school spending patterns. Neither is a statement about quality relative to other articles.

The Tennessee Class Size Experiment (Project STAR)

A different form of evidence – that from random assignment experiments – has recently been widely circulated in the debates about class size reduction. Following the example of medicine, one large-scale experimental investigation in the State of Tennessee in the mid1980s (Project STAR) pursued the effectiveness of class size reductions. Random-assignment experiments in principle have considerable appeal. The underlying idea is that we can obtain valid evidence about the impact of a given well-defined treatment by randomly assigning subjects to treatment and control groups. This eliminates the possible contaminating effects of other factors and permits conceptually cleaner analysis of the outcomes of interest across these groups. The validity of any particular experiment nonetheless depends crucially on the implementation of the experiment. On this score, considerable uncertainty about the results is introduced. But, ignoring any issues of uncertainty, the estimated impacts of large class size reductions are small and have limited application to the current policy proposals.

Project STAR was designed to begin with kindergarten students and to follow them for four years (Word et al., 1990). Three treatments were initially included: small classes (13-17 students); regular classes (22-25 students); and regular classes (22-25 students) with a teacher's aide. Schools were solicited for participation, with the stipulation that any school participating must be large enough to have at least one class in each treatment group. The initial sample included 6,324 kindergarten students. These were split between 1,900 in small classes and 4,424 in regular classes. (After the first year, the two separate regular class treatments were effectively combined, because there were no perceived differences in student performance).[17] The initial sample included 79 schools, although this subsequently fell to 75. The initial 326 teachers grew slightly to reflect the increased sample size in subsequent grades, although of course most teachers are new to the experiment at each new grade.

[17] Surprisingly, policy discussions seldom focus on this finding about the ineffectiveness of teacher's aides.

The results of the Project STAR experiment have been widely publicized. The simplest summary is that students in small classes performed significantly better than those in regular classes or regular classes with aides in kindergarten and that the achievement advantage of small classes remained constant through the third grade.[18]

This summary reflects the typical reporting, focusing on the differences in performance at each grade and concluding that small classes are better than large (e.g., Finn and Achilles, 1990; Mosteller, 1995). But, it ignores the fact that one would expect the differences in performance to become wider through the grades because they continue to get more resources (smaller classes) and these resources should, according to the hypothesis, keep producing a growing advantage. Figure 4 shows the difference in reading performance in small classes that was observed across grades in Project STAR. (The results for math performance are virtually identical in size and pattern). It also shows how the observed outcomes diverge from what would be expected if the impact in kindergarten were also obtained in later grades. As Krueger (1999) demonstrates, the small class advantage is almost exclusively obtained in the first year of being in a small class – suggesting that the advantages of small classes are not general across all grades.

The gains in performance from the experimental reduction in class size were relatively small (less than .2 standard deviations of test performance), especially in the context of the magnitude of the class size reduction (around 8 students per class). Thus, even if Project STAR is taken at face value, it has relatively limited policy implications.

While the experimental approach has great appeal, the actual implementation in the case of Project STAR introduces uncertainty into these estimates (Hanushek 1999b). The uncertainty arises fundamentally from questions about the quality of the randomization in the experiment. In each year of the experiment, there was sizable attrition from the prior year's treatment groups, and these students were replaced with new students. Of the initial experimental group starting in

[18] Some students entered small classes in later grades, and their achievement was observed to be higher during their initial year of being in a small class than that of those in regular classes. See Hanushek (1999b) and Krueger (1999).

kindergarten, 48 percent remained in the experiment for the entire four years.[19] No information, such as pretest scores, is available to assess the quality of student randomization for the initial experimental sample or for the subsequent additions to it. (The data in Figure 4 are equally consistent with either a true small class advantage or an initial assignment of somewhat better students to small kindergartens). It is also impossible to assess adequately the impact of differential attrition of experimental subjects, particularly of those in larger classes disappointed over their placement. Substantial, nonrandom test taking occurs over the years of the experiment. But, most important, the results depend fundamentally on the choice of teachers. While the teachers were to be randomly assigned to treatment groups, there is little description of how this was done. Nor is it easy to provide any reliable analysis of the teacher assignment, because only a few descriptors of teachers are found in the data and because there is little reason to believe that they adequately measure differences in teacher quality.[20] Moreover, teachers all knew they were participating in an experiment that could potentially affect the future resources available from the state. The schools themselves were self-selected and are clearly not random. Small schools were excluded from the study, and all participating schools were willing to provide their own partial funding to cover the full costs. (This school selection issue is important, because the STAR experiment heavily oversampled urban and minority schools where the achievement response to the program is thought to be largest).[21] The net result of each of these

[19] Throughout the four years of the experiment there was also substantial and nonrandom treatment group crossover (about 10 percent of the small class treatment group in grades 1-3). That is, some students originally assigned to large classes moved to small classes later in the experiment. A smaller number also went in the opposite direction. These students were clearly not random. While this problem can be dealt with analytically, it lowers the information that can be obtained from the experiment.

[20] One measure of the importance of teachers relative to class size effects is that the average kindergarten achievement in small classes exceeds that in regular and regular with aide classes in only 40 of the 79 schools.

The teacher data include race, gender, teaching experience, highest degree, and position on the Tennessee career ladder. While there is no information about the effect of career ladder position on student performance, none of the other measures has been found to be reliable indicators of quality (Hanushek 1997). For estimates of the magnitude of variation in teacher quality, see below.

[21] Krueger (1999) identifies significantly stronger effects for disadvantaged students, which will then be overweighted in calculating program average treatment effects.

effects is difficult to ascertain, but there is prima facie evidence that the total impact is to overstate the impact of reduced class size (Hanushek 1999b).

The STAR experiment is very important from a methodological perspective. More random-assignment experimentation is desperately needed in schools. But the evidence from this specific experiment should be interpreted with caution. Moreover, the evidence as it stands speaks just to the possible small effects of major and costly reductions in class size at kindergarten or first grade. It provides no evidence about beneficial effects at later grades. Nor does it indicate what effects could be expected from reductions of a smaller magnitude than the 1/3 reductions in Project STAR.

Policy Calculations

In addition to issues of how to interpret the existing class size evidence, Krueger (2000) attempts to provide a justification for undertaking large class size reductions even if the effects are as small as currently estimated by Project STAR. His argument is simple: Small effects on achievement may have large enough impacts on subsequent earnings that the policies are justified. In order to do these calculations, Krueger takes the perspective that the proper comparison is between doing nothing and undertaking large reductions in class size. This perspective is very narrow and would lead to quite wasteful policies. Moreover, even to get to this justification, he must make a number of heroic assumptions about achievement and the labor market. These assumptions imply enormous uncertainty in the calculations, and thus in the subsequent policy recommendations.

Krueger (2000) presents a series of calculations based on chaining together a variety of uncertain estimates about key aspects of the rewards to higher achievement. In order to obtain estimates of the labor market returns to class size reductions, one must multiply the effect of the class size reduction on achievement times the impact of early achievement differences on performance throughout schooling and into the labor market. The subsequent estimates of initial

labor market advantage must be projected across a person's working life and then discounted back to kindergarten to compare to the costs of the original class size reduction. The uncertainty with each of those steps grows when they are compounded together. The relationship between early achievement and subsequent earnings, for example, relies on a single study of British labor market experiences for a group of individuals born in 1958 and recording their wages in 1981 and 1991.[22] These estimates are employed to project what expected early career labor market experiences might be in the United States around 2015, the relevant period for the policy deliberations. While it may be academically interesting to see if there is any plausibility to the kinds of class size policies being discussed, one would clearly not want to commit the billions of dollars implied by the policies on the basis of these back-of-the-envelope calculations.[23]

Surely improving achievement of students is very important and should be the focus of policy attention. The issue is not whether society should invest in quality but how it should invest. Calculations that suggest the economic justification is as close to breakeven as found by Krueger do not make a good case for the huge commitment of resources implicitly behind his calculations – particularly when the uncertainty of the calculations is recognized.

At the heart of the issue, however, Krueger ignores the fact that existing evidence points to other factors – particularly teacher quality –as being more important than class size. The extensive research on student achievement over the past 35 years has made it clear that there are very important differences among teachers. This finding, of course, does not surprise many parents who are well aware of quality differences of teachers, but it has eluded many researchers. Researchers have tended to confuse measurability of specific teacher characteristics related to quality with real differences in quality. That is, the econometric research has not identified any

[22] His discussion does consider two alternative estimates, although they appear to differ substantially from the estimates chosen for the calculations.

[23] Krueger (2000) suggests that, because of uncertainty, it might be appropriate to compare his calculated rate of return to class size reductions to a somewhat higher interest rate than the four percent he appears to favor. His suggestion of perhaps considering a six percent return, however, vastly understates the uncertainty one would calculate by the normal procedure of developing confidence intervals for the estimates that enter into his illustrative benefit-cost approximations.

teacher attributes (such as education, experience, background, type of training, certification, or the like) that are highly related to the ability of some teachers to get particularly large or particularly small gains in student learning. Nonetheless, econometric analyses have identified large and persistent differences in the effectiveness of different teachers.[24]

The magnitude of differences in teacher quality is impressive. For example, looking at the range of quality for teachers within a single large urban district, teachers near the top of the quality distribution can get an entire year=s worth of additional learning out of their students compared to those near the bottom (Hanushek 1992).[25] That is, a good teacher will get a gain of 1**2** grade level equivalents, while a bad teacher will get **2** year for a single academic year. A second set of estimates comes from recent work on students in Texas (Rivkin, Hanushek, and Kain 2000). This analysis follows several entire cohorts of students and permits multiple observations of different classes with a given teacher. We look at just the variations in student performance that arise from differences in teacher quality within a typical school and do not consider any variations across schools. The variation is large: Moving from an average teacher to one at the 85th percentile of teacher quality (i.e., moving up one standard deviation in teacher quality) implies that the teacher=s students would move up more than 7 percentile rankings in the year.[26] These differences swamp any competing factors such as measured teacher and school attributes in their impact on student performance. For example, a one standard deviation reduction in class size implies a 0.01-.03 standard deviation improvement in student achievement. The lower bound estimate on teacher quality summarized here implies a one standard deviation

[24] The econometric analysis behind these estimates involves calculating the average achievement gains across classrooms after allowing for differing student preparation, family background, and other factors. Some teachers consistently obtain high growth in student achievement, while others consistently obtain low growth. But, standard measures of teacher characteristics are not correlated with quality measured in terms of value-added to student performance.

[25] These estimates consider value-added models with family and school inputs. The sample includes only low-income minority students, whose average achievement in primary school is below the national average. The comparisons given compare teachers at the 5th percentile with those at the 95th percentile.

[26] For a variety of reasons, these are lower bounds estimates of variations in teacher quality. Any variations in quality across schools would add to this. Moreover, the estimates rely on a series of conservative assumptions which all tend to lead to understatement of the systematic teacher differences.

change in quality leads to a 0.18 standard deviation increase in achievement. Finally, quality differences in teachers in Tennessee of a similar magnitude have also been estimated (Sanders and Horn 1995).

Recognizing the importance of teacher quality is central to the discussion of class size. First, any substantial reductions in class size imply hiring additional teachers. The success or failure of a class size reduction program will depend much more on whether or not the newly hired teachers are better or worse compared to the existing teachers than it will on the impact of class size reduction per se. In fact, depending upon the structure of the enabling legislation or policy, it could have quite detrimental effects.[27] Second, the Krueger calculations never consider the possibility of much more attractive alternatives to either the current schools or to class size reductions. Employing higher quality teachers could produce major impacts on student performance that are unachievable with any realistic or feasible class size reductions.

A major difference in policies aimed at class size reduction and those aimed at changing teacher quality is their relationship to incentives in schools. There is ample reason to believe that the current incentives related to student performance are too weak (Hanushek with others, 1994). Essentially nobody within schools has much riding on whether or not students achieve at a high level. The expected pay and career of a good teacher is about the same as that for a bad teacher. Class size reduction does nothing to change this. On the other hand, if schools are to move toward attracting and retaining higher quality teachers, they will almost certainly have to build in stronger performance incentives for school personnel. The exact form that this would take is unclear, and discussion of the options is beyond the scope of this paper (see, however, Hanushek with others, 1994). The necessity of altering incentives on the other hand seems clear, at least to economists.

[27] The 1996 class size reduction program in California left inner city schools scrambling for new teachers, partly as a result of suburban districts' bidding away experienced teachers (Stecher and Bornstedt 1999). The likely net result is that disadvantaged students – the hypothesized winners from the reduction policy – actually suffered a loss in educational quality.

Reducing class size does not logically preclude doing other things, but it is almost certainly a practical deterrent. Limited political attention and constraints on public funds imply that strong moves toward class size reduction are almost certain to drive out better policies aimed at improving teacher quality.

Conclusions

Despite the political popularity of overall class size reduction, the scientific support of such policies is weak to nonexistent. The existing evidence suggests that any effects of overall class size reduction policies will be small and very expensive. A number of investigations appear to show some effect of class size on achievement for specific groups or circumstances, but the estimated effects are invariably small and insufficient to support any broad reduction policies. The flawed analysis in Krueger (2000) does little to contribute to the debate on technical grounds and, more importantly, cannot change the inherent costs and expected benefits of the basic policy. The re-analysis of econometric estimates relies on placing heavy weight on lower quality and biased econometric estimates. Even then, the efficacy of class size reduction is in doubt. The majority of his re-weighted estimates are still statistically insignificant, i.e., we have relatively little confidence that there is any effect on student outcomes. The most optimistic estimates suggest that the policy effects on student achievement would be small. The policy effects are shown by Krueger (2000) to make sense given the cost only if one makes a number of strong but uncertain assumptions and only if one believes that no other school policy is feasible.

Proposed class size reduction policies generally leave no room for localities to decide when and where reductions would be beneficial or detrimental. The existing evidence does not say that class size reductions are never worthwhile and that they should never be taken. It does

say that uniform, across-the-board policies – such as those in the current policy debate – are unlikely to be effective.[28]

A significant problem is that there are few incentives that drive decisions towards ones that improve student performance. Most economists believe that incentives are key to results – whether in education or in other aspects of life. But schools are not organized in a way that they decide to reduce class size in instances where it is beneficial for student performance and not in other instances where it would not affect performance. Without such performance incentives, simply adding more resources is unlikely to lead to improvements in student achievement. In this regard, education has made very little progress in spite of the large and continuing investment in specific programs and activities.

Class size reduction is best thought of as a political decision. Past evidence suggests that it is a very effective mechanism for gaining voter support, even if past evidence also suggests that it is a very ineffective educational policy.

[28] For example, the theoretical analysis of class size by Lazear (forthcoming) points to optimal policies when schools are trying to maximize student achievement. In this case, he shows that across-the-board reductions are never going to be the correct policy.

References

Burkhead, Jesse. *Input-output in large city high schools.* Syracuse, NY: Syracuse University Press, 1967.

Card, David, and Alan B. Krueger. "Does school quality matter? Returns to education and the characteristics of public schools in the United States." *Journal of Political Economy* 100, no. 1 (February 1992): 1-40. (a)

---. "School quality and black-white relative earnings: A direct assessment." *Quarterly Journal of Economics* 107, no. 1 (February 1992): 151-200. (b)

---. *Myth and Measurement: The new economics of the minimum wage*. Princeton, NJ: Princeton University Press, 1995.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office, 1966.

Congressional Budget Office. *Trends in educational achievement.* Washington, D.C.: Congressional Budget Office, 1986.

Finn, Jeremy D., and Charles M. Achilles. "Answers and Questions about class size: A statewide experiment." *American Educational Research Journal* 27, no. 3 (Fall 1990): 557-77.

Grissmer, David W., Sheila Nataraj Kirby, Mark Berends, and Stephanie Williamson. *Student achievement and the changing American family*. Santa Monica, CA: Rand Corporation, 1994.

Hanushek, Eric A. "The economics of schooling: Production and efficiency in public schools." *Journal of Economic Literature* 24, no. 3 (September 1986): 1141-77.

---. "The trade-off between child quantity and quality." *Journal of Political Economy* 100, no. 1 (February 1992): 84-117.

---. "Assessing the effects of school resources on student performance: An update." *Educational Evaluation and Policy Analysis* 19, no. 2 (Summer 1997): 141-64.

---. "Conclusions and controversies about the effectiveness of school resources." *FRBNY Economic Policy Review* 4 (March 1998): 11-28.

---. "The evidence on class size." In *Earning and learning: How schools matter*, edited by Susan E. Mayer and Paul Peterson, 131-68. Washington, DC: Brookings Institution, 1999. (a)

---. "Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects." *Educational Evaluation and Policy Analysis* 21, no. 2 (Summer 1999): 143-63. (b)

Hanushek, Eric A., and Steven G. Rivkin. "Understanding the twentieth-century growth in U.S. school spending." *Journal of Human Resources* 32, no. 1 (Winter 1997): 35-68.

Hanushek, Eric A., Steven G. Rivkin, and Lori L. Taylor. "Aggregation and the estimated effects of school resources." *Review of Economics and Statistics* 78, no. 4 (November 1996): 611-27.

Hanushek, Eric A., and with others. *Making schools work: Improving performance and controlling costs*. Washington, DC: Brookings Institution, 1994.

Heckman, James S., Anne Layne-Farrar, and Petra Todd. "Does measured school quality really matter? An examination of the earnings-quality relationship." In *Does money matter? The effect of school resources on student achievement and adult success*, edited by Gary Burtless, 192-289. Washington, DC: Brookings, 1996.

Heckman, James, Anne Layne-Farrar, and Petra Todd. "Human capital pricing equations with an application to estimating the effect of schooling quality on earnings." *Review of Economics and Statistics* 78, no. 4 (November 1996): 562-610.

Hedges, Larry V. "Directions for future methodology." In *The future of meta-analysis*, edited by Kenneth W. Wachter and Miron L. Straf, 11-26. New York: Russell Sage, 1990.

Kiesling, Herbert. "Measuring a local government service: a study of school districts in New York state." Ph.D. Dissertation, Harvard University, Cambridge, MA, 1965.

---. "Measuring a local government service: a study of school districts in New York state." *Review of Economics and Statistics* 49 (August 1967): 356-67.

Krueger, Alan B. "Experimental estimates of education production functions." *Quarterly Journal of Economics* 114, no. 2 (May 1999): 497-532.

---. An economist's view of class size research. mimeo, July 29, 2000.

Lazear, Edward. "Educational production." *Quarterly Journal of Economics* (forthcoming).

Lewit, Eugene M., and Linda Schuurmann Baker. "Class size." *The Future of Children* 7, no. 3 (Winter 1997): 112-21.

Link, Charles R., and James G. Mulligan. "The merits of a longer school day." *Economics of Education Review* 5, no. 4 (1986): 373-81.

Montmarquette, Claude, and Sophie Mahseredjian. "Does school matter for educational achievement? A two-way nested-error components analysis." *Journal of Applied Econometrics* 4 (1989): 181-93.

Mosteller, Frederick. "The Tennessee study of class size in the early school grades." *The Future of Children* 5, no. 2 (Summer/Fall 1995): 113-27.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. "Teachers, schools, and academic achievement". National Bureau of Economic Research, Working Paper No. 6691 (revised) 2000.

Sanders, William L., and Sandra P. Horn. "The Tennessee Value-Added Assessment System (TVAA): Mixed model methodology in educational assessement." In *Teacher evaluation: Guide to effective practice*, edited by Anthony J. Shinkfield and Daniel L. Stufflebeam, 337-76. Boston: Kluwer Academic Publishers, 1995.

Speakman, Robert, and Finis Welch. "Does school quality matter? --A Reassessment". Texas A&M University (mimeo), January 1995.

Stecher, Brian M., and George W. Bohrnstedt, eds. *Class size reduction in California: Early Evaluation Findings, 1996-98.* Palo Alto: American Institutes for Research, 1999.

Word, Elizabeth, John Johnston, Helen Pate Bain, B. DeWayne Fulton, Jayne Boyd Zaharies, Martha Nannette Lintz, Charles M. Achilles, John Folger, and Carolyn Breda. *Student/teacher achievement ratio (STAR), Tennessee's K-3 class size study: Final summary report, 1985-1990.* Nashville, TN: Tennessee State Department of Education, 1990.

Appendix: Issues with the Econometric Data

Krueger (2000) raises a number of questions about the underlying estimates included in the overall summaries. Several of them were discussed with Krueger in private correspondence but did not make it into the published version.

Three coding questions are raised. First, as mentioned above, earlier correspondence determined that I had reversed the sign on the four estimated teacher-pupil ratio effects in Montmarquette and Mahseredjian (1989) in my previous tabulations, but Krueger subsequently does not make this correction in his tables. Second, 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. In private communication with them to clarify this issue, they indicated it was included in all 24 – and this was communicated to Krueger. Third, Kiesling (1967) is a journal article that extracted results from his thesis (Kiesling 1965), and the teacher-pupil ratio results came from his thesis. While this was noted in Hanushek (1986), it was not noted in Hanushek (1997), although it also was communicated to Krueger.

Table 1. Pupil-teacher Ratio and Real Spending, 1960-1995

	1960	1970	1980	1990	1995
Pupil-teacher ratio	25.8	22.3	18.7	17.2	17.3
Current expenditure per pupil (1996/97 $'s)	$2,122	$3,645	$4,589	$6,239	$6,434

Table 2. Percentage Distribution of Estimated Effect of Teacher-Pupil Ratio and Spending on Student Performance

Resource	number of estimates	Statistically significant		Statistically insignificant		
		Positive	Negative	Positive	Negative	Unknown sign
Teacher-pupil ratio	277	14%	14%	27%	25%	20%
Expenditure per pupil	163	27	7	34	19	13

Source: Hanushek (1997), as corrected (see text).

Table 3. Sample Sizes for Estimated Effect of Teacher-Pupil Ratio by Number of Estimates per Publication

Number of Estimates per Publication	Number of Estimates (Publications)	Sample Size			
		Median	Average	Minimum	Maximum
1	17 (17)	272	1,305	48	14,882
2-3	28 (13)	649	1,095	56	5,000
4-7	109 (20)	512	2,122	38	18,684
8-24	123 (9)	266	1,308	22	10,871
1-24	277 (59)	385	1,607	22	18,684

Table 4. Percentage Distribution of Estimated Effect of Teacher-Pupil Ratio and Expenditure per Pupil by State Sampling Scheme and Aggregation

Level of aggregation of resources	number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
A. Teacher-Pupil Ratio				
Total	277	14%	14%	72%
Single state samples[a]	157	11	18	71
Multiple state samples[b]	120	18	8	74
Disaggregated within states[c]	109	14	8	78
State level aggregation[d]	11	64	0	36
B. Expenditure per pupil				
Total	163	27%	7%	66%
Single state samples[a]	89	20	11	69
Multiple state samples[b]	74	35	1	64
Disaggregated within states[c]	46	17	0	83
State level aggregation[d]	28	64	4	32

a. Estimates from samples drawn within single states.
b. Estimates from samples drawn across multiple states.
c. Resource measures at level of classroom, school, district, or county, allowing for variation within each state.
d. Resource measures aggregated to state level with no variation within each state.

Table 5. Percentage Distribution of Estimates of Teacher-pupil Ratio on Student Performance, Based on Value-added Models of Individual Student Performance

	Number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
All	78	12%	8%	80%
Estimates for single state samples	23	4%	13%	83%

Figure 1. Pupil-teacher ratios and Student Performance

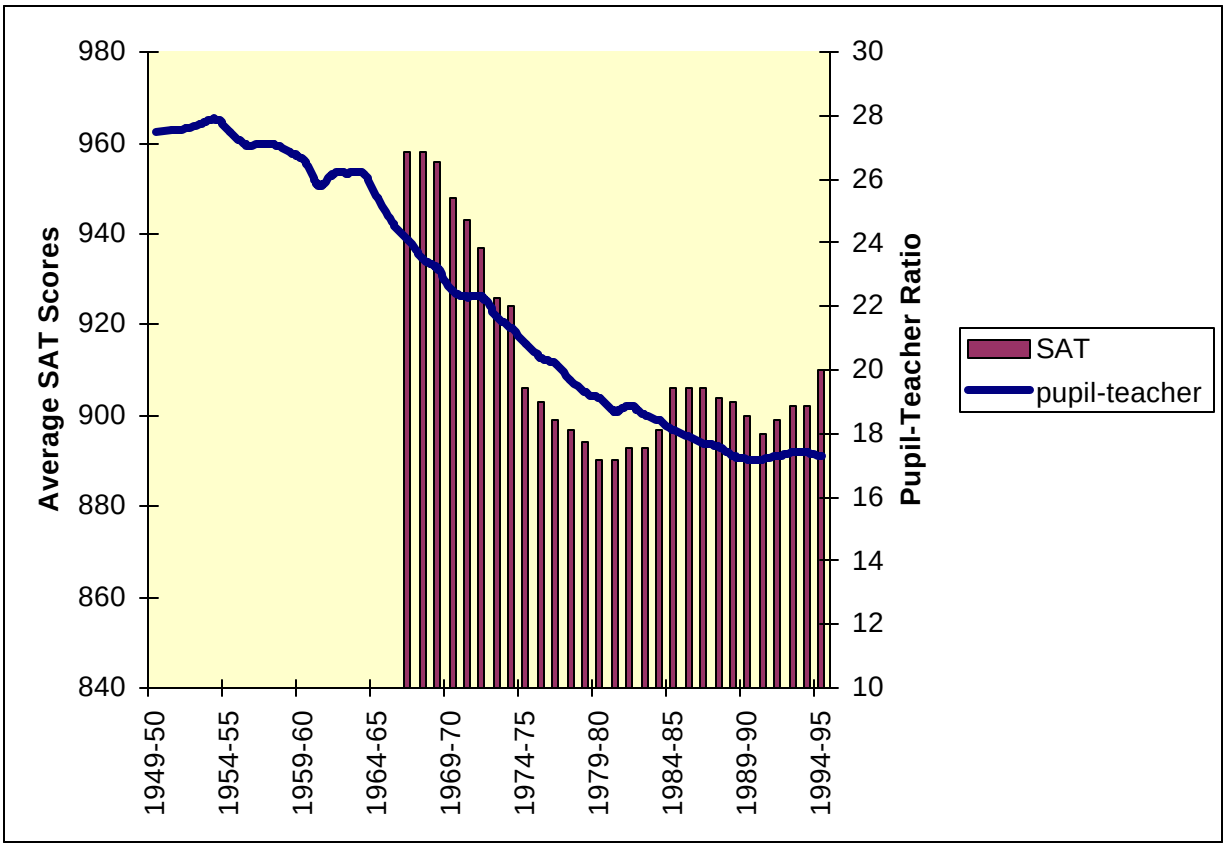


Fig. 2: National Assessment of Educational Progress -- 17 year-olds

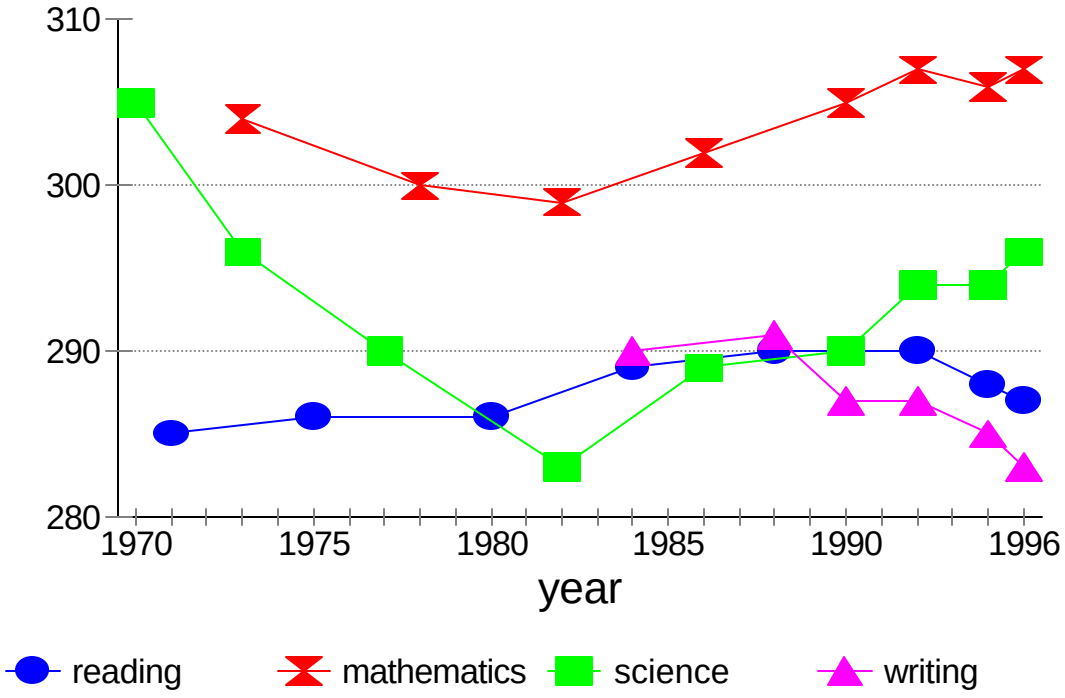


Table 3

Estimates for Teacher-Pupil Ratio
with Alternative Weighting

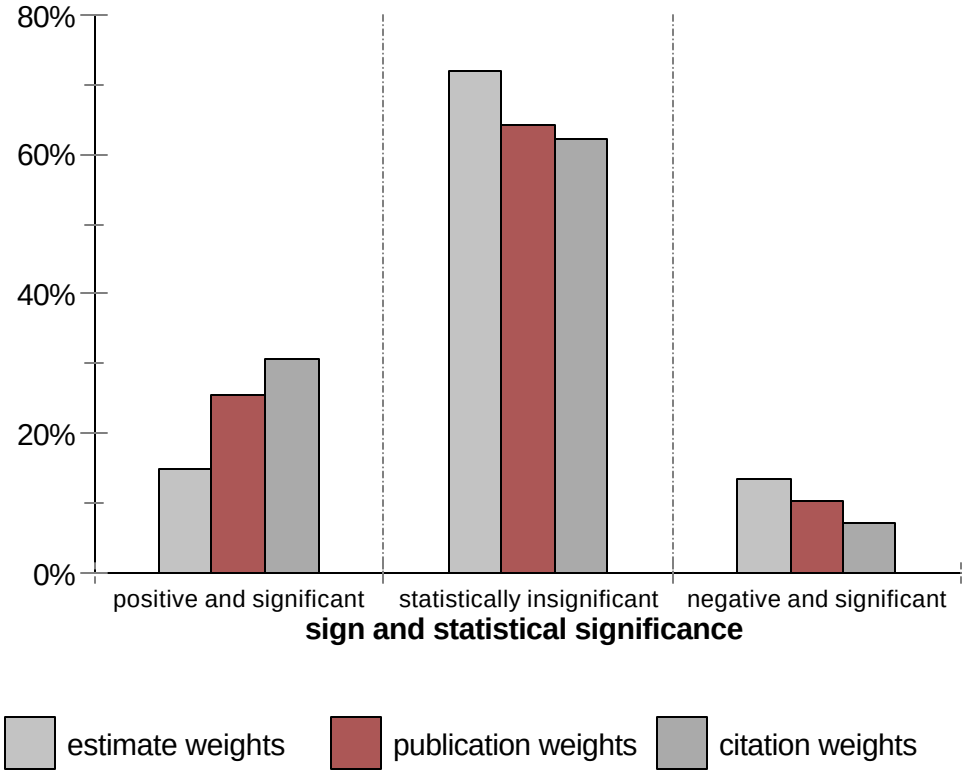
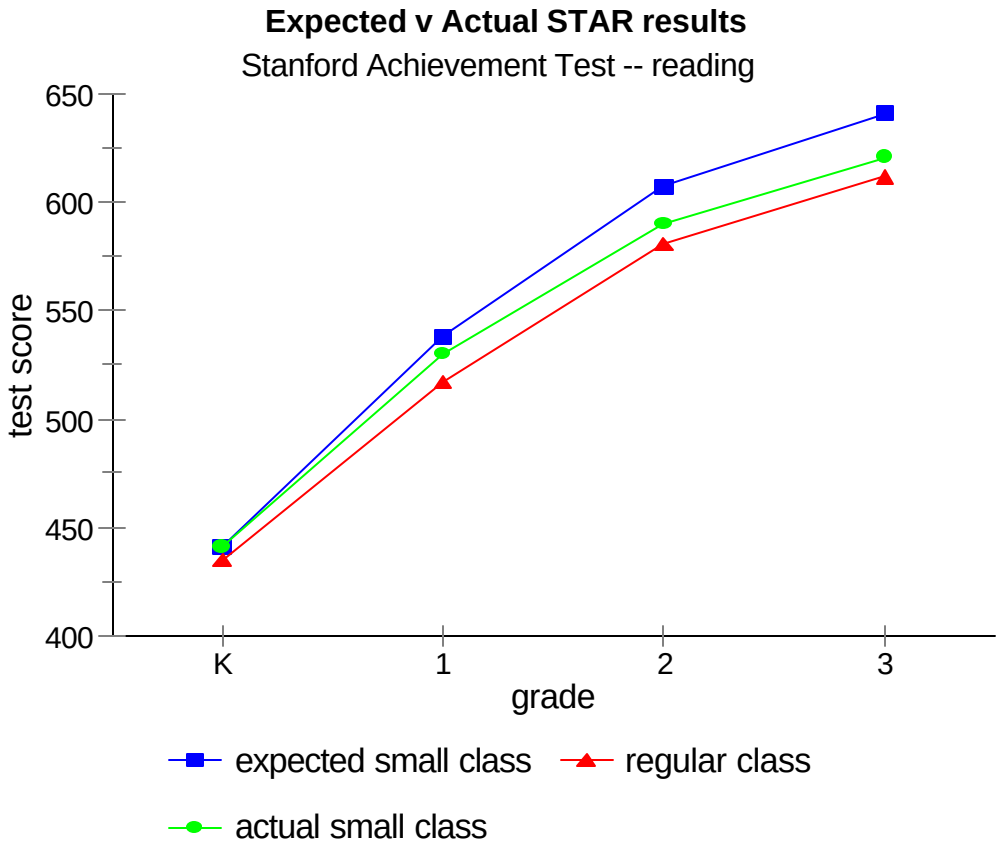


Figure 4


Expected v Actual STAR results
Stanford Achievement Test -- reading
650
600
550
500
450
400
test score
K
1
2
3
grade
expected small class
regular class
actual small class