AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

ECONOMIC GROWTH CENTER

YALE UNIVERSITY

P.O. Box 208269
New Haven, CT 06520-8269

CENTER DISCUSSION PAPER NO. 827

THE EFFECTS OF CLASS SIZE ON THE LONG RUN GROWTH IN READING ABILITIES AND EARLY ADULT OUTCOMES IN THE CHRISTCHURCH HEALTH AND DEVELOPMENT STUDY

Michael A. Boozer
Yale University

and

Tim Maloney
University of Auckland

May 2001

Note: Center Discussion Papers are preliminary materials circulated to stimulate discussions and critical comments.

The authors are grateful to Dr. Ron Crawford for his support and comments on the issues investigated in this report, and they thank the New Zealand Treasury for financial support. In addition, Professor David Fergusson, Executive Director of the Christchurch Health and Development Study provided extensive comments on an earlier draft which greatly improved the present version, as well as comments from an anonymous referee. Correspondence: michael.boozer@yale.edu; t.maloney@auckland.ac.nz.

This paper can be downloaded without charge from the Social Science Research Network electronic library at: http://papers.ssrn.com/paper.taf?abstract_id=275297

An index to papers in the Economic Growth Center Discussion Paper Series is located at: http://www.econ.yale.edu/~egcenter/research.htm

The Effects of Class Size on the Long Run Growth in Reading Abilities and Early Adult Outcomes in the Christchurch Health and Development Study

Michael A. Boozer
Yale University

and

Tim Maloney
University of Auckland

ABSTRACT

This paper utilizes the feature of the CHDS data from New Zealand that children are sampled for extremely long individual histories of their class size experiences as well as their scholastic and early labor market outcomes. Our interest is to explore the full set of empirical implications of the recent experimental evidence on class size effects on student achievement from the United States in Project STAR for observational data. We argue that one implication of Project STAR is that only *persistent* class size reduction policies may have detectable effects, and so the longitudinal aspect of CHDS is necessary to detect class size effects. We account for the observational nature of the CHDS (in that children were not randomly assigned to different class sizes) by examining the long-run trends in test score growth, rather than levels. Consistent with the experimental evidence, we find statistically and economically significant effects of children being assigned to persistently smaller classes on both childhood test score growth as well as on early adult outcomes. Our analysis points the way towards the unification of experimental and observational evidence on class size effects, as well as highlighting several possible pitfalls in the analysis of observational data on this topic.

JEL Classification: C51, C81, I21, C23
Keywords: School Quality; Value-Added Model; Experimental Evidence

1 Introduction

The Christchurch Health and Development Study (CHDS) data have now been studied with regard to student achievement in a series of six previous reports conducted by LECG for the New Zealand Treasury. The purpose of this report is to summarize the highlights of these earlier reports as well as draw on their cumulative knowledge to further advance our understanding of school resource policy decisions and their relation to student outcomes. The resource measure we use is the class size for the student and the academic achievement measure we use is the score on the Burt Word reading test, and we have available to us both of these measures, as well as a host of control measures, for the CHDS participants from ages 8 to 13, annually. The chief novelty that these data bring to bear on these well-studied issue is that the long time-span of the individual test and schooling histories is almost unique, as well as is the annual sampling of this measure. In addition, this version of our report makes use of the recently collected age 21 data from the CHDS which allows us to study non-test score outcomes such as the quantity of completed schooling and university attendance. These additional outcomes have the advantage of being of direct policy interest as well as being free of the 'mechanical' aspects of the test score measures (we elaborate on these below) which may confound our analysis based solely on the test score outcomes.

This report reaches several conclusions utilizing this unique data, as well as the six earlier reports, which we believe will be of interest to policymakers and academics interested in basic research on schooling alike:

- The variation in Burt Word Reading Tests is almost entirely explained by individual and age differences. Together these explain about 95 percent of the variance in overall scores. This implies that great care must be used in using comparisons in tests to detect effects of policies, since, for example, a standard fixed effects estimator annihilates all but 5 percent of the variance in the tests. As a general conclusion, this outcome, because of its ability to measure 'permanent' differences, may not simply be that malleable, and so not a very powerful tool to detect policy effects if not used with care. Our means of dealing with this property of the test scores is to utilize purely the long window of time afforded by the sample frame of the CHDS and essentially discard the year-to-year fluctuations as offering little valid information.

- The distinction between a test instrument's reliability and its stability has been emphasized to us by our previous reviewers. It is not uncommon to use a measure of a test's stability properties as a measure of its reliability properties. But because social scientists (as opposed to psychometricians, for example) study not just the univariate properties of test outcomes, but the associations with policy variables, test outcomes in their 'level' form are often not

used in favor of analyses such as 'fixed-effects' analyses. These latter form of analyses make use of the *changes* in the test outcome for a given individual over time and relate these to the policy variables in an attempt to extract a more causal relationship from policy variables to outcomes from observational (i.e. non-experimental) data. The potential problem arises because the 'stable' part of the test score is discarded in such analyses, and so the reliability properties of the remaining variation must be assessed. But if the stability and reliability properties are equated (mistakenly, see Heise (1969) for a clear articulation of the distinction between these concepts as well as a framework for empirically distinguishing between them if more than two observations on the test outcome are observed for each individual) then researchers will know little of the reliability properties of the so-called 'within' (or across time) variation in the test scores for each individual. For a social scientist, it is imperative to use testing instruments with good reliability measures, but which is distinct from a high stability property, since the latter source of variation is discarded in a fixed-effects analysis so common in correlational studies in the social sciences. The same comment applies to the measure and concept of a test's 'external validity'. We find that the within person validities for the Burt tests to be substantially smaller than the reported overall validity measures, although still quite significantly related to future outcomes. Our Appendix elaborates on these methodological issues.

- Using as a point of departure the recent literature from the U.S. on experimental studies involving class size reductions and test score outcomes, we investigated alternatives to the so-called 'value-added' model. Comparing short-term fluctuations in class sizes with short-term fluctuations in test scores (in earlier work not reported here) yielded effects of the expected signs, but quite imprecisely estimated. In short, from a statistical perspective, we could not detect effects using these types of comparisons, which were suggested by the U.S. experimental literature. The point estimates were admittedly more intuitive than the estimates derived in the earlier six reports based on the standard value-added model (which were perverse in sign), but their sensitivity to alternative specifications as well as their associated confidence intervals indicate the effects may not be systematic.

- However, an alternative interpretation of the U.S. evidence, and indeed the interpretation used to critique the early work based on the experimental data, is that only sustained policies have effects. In the U.S., this meant that students initially placed in small classes only retained their initial gains in test scores if they continued to be enrolled in small classes in higher grades as well. When placed in regular classes, their initial gains were seen to fade away. Based on this interpretation of the U.S. data we investigated if a similar phenomena were at work in the CHDS data, and here our conclusions are more precise. We found that students who were always in 'large' classes throughout the sample

period had somewhat significantly lower test score growth from age 8 to 13 than other students. We also found that students who were always in 'small' classes had somewhat higher growth in test scores, but this effect is rather small and quite imprecisely estimated, so as to be indistinguishable from other students.

• Plots of the distributions in raw gain scores indicated that a mean regression may be missing what appears to be heterogeneity in the policy effects. In addition, like most test scores, the gain in Burt test scores is rather negatively correlated with the initial level (near 'topping out' becomes rather evident by age 13). Rather than massage the data to somehow 'correct' for this (via some *ad hoc* correction) we opted to simply bear it in mind in interpreting our results and remember that the high-gain students are generally the students with the initially low test score performance. For this reason we supplemented our analysis with quantile regression analysis to try to detect the heterogeneous policy responses.

• In particular, we looked at quantile regressions for the effects of permanent class size categories at the 9 deciles. While the quantile regression coefficients are rather imprecise owing to the small samples (particularly for these permanent class size categories), the point estimates suggest a pattern of effects that are larger in magnitude at the higher quantiles (above 0.5) as compared to the lower quantiles. Since the higher *unconditional* quantiles of the gain distribution correspond to the *lower* quantiles of the test score level distribution, then if the conditioning does not lead to too great a reversal of the unconditional quantiles, this suggests slightly larger effects of class size reduction policies for children with lower initial Burt test scores. While this conclusion may seem rather intuitive, we offer this conclusion cautiously owing to the high imprecision of the quantile estimates. In addition, this result may largely be a feature purely of the Burt scores and the 'topping out' problem - i.e. Burt scores may simply be more malleable for the children who initially score low as compared to those children who initially score high. If this is true, then were we to use another, less persistent outcome measure, we might then find more uniformity in class size effects for that outcome.

• Less distinct class size groupings, such as classifying children by their Average as opposed to Permanent class size experiences, lead to point estimates which are less sharp. However, owing to the greater use of the full sample, the sampling variances are accordingly smaller, and for the most part, many of our estimates of the effects of what we call 'persistent' class size reduction policies are statistically significant in the 5 to 10 percent range. It is really the overall pattern of results (signs, sizes, and statistical significance) that leads us to make forceful conclusions. In addition, given the marked persistence in Burt scores, it is rather amazing we find the effects we do over a six year sampling window

for the CHDS cohort.

• Compared with the U.S. literature on school quality and class size effects, our estimated effect size is large, varying from about 0.03 to 0.06 of a standard deviation of the gain scores for a one student reduction in average class sizes. Owing to the estimated diminishing returns to lowering class sizes for smaller initial class sizes, this effect size declines (to the bottom estimate of 0.03) as the initial class size is smaller. One apparent reason our effect size is large is we are considering 'persistent' (ages 8 to 13) class size reduction policies, as opposed to reductions for just a single year. As for two examples of the economic (or contextual) significance of our estimates, roughly speaking a reduction of 3 to 4 students in average class sizes over this age range would close the Maori / Non-Maori test score gap by about 70 percent. This same policy idea would close the gap in test scores between the children from families with the lowest decile of family income and the highest decile of family income (one of the largest test score gaps by demographic classifications we could find) by about 16 percent. We found that indeed, given the within (or change) Burt score correlations with external measures such as teacher ratings of the child, the closing of these Burt score gaps would be associated with greater uniformity of overall academic progress.[4]

• Finally, we were able to recently acquire the age 21 follow-up data to the CHDS to examine some of the early adult outcomes. While the young age of the individuals as of this follow-up precluded us from meaningfully analyzing outcomes such as wages or income, we were able to analyze education completed as of age 21, as well as the incidence and duration of unemployment. We utilized a methodology similar to that used to study the Burt scores in that we used a 'long time window' value-added model, holding constant the Burt score at age 8 as well as family background, while using the class size at age 13 to proxy for the high school class size exposure. We found that lower class sizes are moderately related to more completed education as of age 21, and while the effects are not overwhelming, a more complete cost-benefit analysis may show them to be worth public financing of class size reductions. When we looked at unemployment incidence, and especially duration (conditional on having *some* incidence), we found stronger effects. A class size reduction by 5 students at age 13 was associated with a shorter time unemployed from age 18 by almost 1.7 months for those individuals having experienced some unemployment (for this *conditional* sample, the average unemployment time was 9 months from age 18 to 21). Thus, we tended to find larger effects of class size reductions on these outcomes as of age 21 for those who appear to be less well off (i.e. among the

[4] Although the estimated validities for the *change* of Burt scores with the *change* in classroom performance (both reading and overall) is on the order of 0.23. As we discuss in the text (page 5), the validity measures using the Burt score *levels* is about 0.7 to 0.8.

group of individuals experiencing some unemployment). In addition to their own substantive interest, the analysis of these early adult outcomes does not suffer from the 'mechanical' problems potentially confounding our analysis of the Burt gain scores (such as the 'topping out' problem, discussed below). As such, we take these findings to corroborate our test score analysis, as well as lending some extra confidence to our methodology which was driven by trying to mimic the findings from the experimental literature on class size effects from the U.S.

2 Burt Word Reading Tests

Early in the work done by LECG on the CHDS data, the academic outcome of interest settled on the Burt Word Reading Test scores. The principal reason for doing so was that unlike some of the other aptitude/achievement tests given over time in the CHDS, the Burt tests were given at each age from age 8 to age 13, then again at age 18 (indeed, the most any other particular testing instrument was administered was twice, which is similar to the testing frequencies found in U.S. data on this topic). Since the most prominent and unique feature of the CHDS data is its time span for each individual, using the large time coverage of the Burt scores was a highly sensible choice. But it is important to bear in mind other test scores are included in the CHDS, although since they were given at most 2 or 3 times, using them as the outcome variable would reduce the CHDS to have a time span comparable to similar datasets elsewhere in the field. As the CHDS cohort ages and leaves school, enters the workforce, make family decisions, etc. the scope for studying long term outcomes of direct policy interest will make the CHDS a highly desirable source of data to return to in the years to come.

The Burt Word Reading Test is literally just that - a student reads from a list of 110 words and her score is the number of words read correctly. Generally, as the children age, they naturally tend to become better at this exercise, and so one of the most prominent features of the test scores over time is this growth in scores by age. By age 13 a number of children have nearly 'topped out' by nearing the ceiling score of 110, although both technical material on the Burt test and inspection by the authors indicate that using the scores through age 13 is not flawed due to the ceiling effects. While only 2 children in our sample score a perfect 110 at age 13, about 10 percent score in the range from 105 to 110. We discuss this property further at the end of this section.

Testing instruments are generally characterized by two standard quantities: 'reliability' and 'external validity' (or, more generally, simply 'validity'). However, as we discuss in our Appendix, and as was impressed upon us by one of our earlier reviewers, a third property of a test instrument is 'stability', which is often ignored (by assuming it is 1). An optimal test would have high reliability as well as good external validity. When stability is ignored, a test's

reliability is measured by the correlation in the test scores for a reference group of children for two test administrations given only a few weeks apart. However, as we show in the Appendix, drawing on the work of Heise (1969), when tests are assumed to not be perfectly 'stable' (the precise definition is found in the Appendix), then the simple test-retest correlation does not measure the test's reliability, and so alternative means or measures must be made to determine the test instrument's reliability. The technical material we have on the Burt test reports reliabilities in excess of 0.95 at all age levels, and ranging up to 0.99 (Gilmore, Croft, and Reid, *Burt Word Reading Test, New Zealand Revision, Teachers Manual*, page 9) - in other words, the Burt testing instrument is highly reliable. External validity properties of tests often conflict slightly with the test's reliability, and thus serves as a brake for simply creating tests with extremely high reliability. External validity is basically a measure of the correlation of Burt scores with either other test score outcomes, or, what we mean here by 'external' validity is correlation with child outcomes of more direct policy interest than test scores *per se*. Unlike the reliability measure, which is basically a context-free notion, assessing validity depends on what policy question is being asked. Psychometricians generally only report validity measures with other common testing instruments. As one example of such a measure, the Burt test has an estimated validity with the Test of Scholastic Abilities (TOSCA) on the order of 0.7 to 0.8 (Gilmore, Croft, and Reid, *Burt Word Reading Test, New Zealand Revision, Teachers Manual*, page 10). Assessing the external validity of Burt scores with longer term outcomes such as school completion, college attendance, employment outcomes, etc. would be a highly useful complement to the present study, once these data are able to be collected from the CHDS cohort. In some initial work in this vein reported in the next-to-last section of this paper, we provide some qualitative evidence in this regard.

These somewhat disparate characteristics of reliability, stability, and external validity will become important when we utilize the Burt scores as an outcome of interest with which to detect the influences of policies such as the altering of schooling resources. Simply put, a test which has very good stability properties (as distinct from reliability properties) may actually be a very poor choice to detect policy effects if, as a consequence of being so highly stable, it is essentially not mutable or changeable for a given individual. The same caveat can be ascribed to the test's validity as well if, for example, the test is very good at indicating *between* children, who will score higher on say a TOSCA exam, or go on to college. But it is another question altogether if the movements in the Burt scores for a *given* child, over time, translate to *improvements* in either TOSCA scores or enhancing the probability of the child attending college.[5] Thus, when we bring the interplay of policy effects on test outcomes in to the picture, a

[5] Using teacher ratings of student performance at ages 8 and 12, we can do this exercise within our sample. Using either teacher ratings on just reading or on 4 subject areas combined leads to an estimated validity of the *change* in Burt scores with the *change* in teacher ratings of about 0.23 and this is significant at better than a 1 percent level.

clear tension begins to emerge between choosing a measure to detect effects of policies versus choosing tests with good reliability and/or validity measures between individuals and 'within' individuals (i.e. for a particular individual) over time. It is when the stability of a test instrument is ignored, as we discuss in the Appendix, that reliability and mutability of test scores come into direct conflict.

The tension arises because it is quite common in the social sciences to be wary of simple comparisons of policies across people and their associated outcomes and draw causal conclusions from such comparisons. The clearest example of this problem of inference can be depicted by looking at the CHDS as an aggregate. As we discussed above, one of the most prominent features that emerge from a first blush examination of the Burt word reading scores is the growth in scores as the children age, for the cohort as a whole. One possible interpretation of that 'aggregate' growth in scores as the children age is that some fraction is due to the *entire* cohort getting better schooling resources, improvements in family background, etc. In fact, we do not do this - because we lack any comparison group for the CHDS *as a whole* we simply throw away the growth in test scores by age, and in doing so assume there is no *aggregate* improvement in policy variables for the CHDS cohort. 'Throwing away' the information in average scores by age simply means including indicator variables for the age of each child in our analysis - we discuss our precise strategies below.

At a more micro level of the data, of course, we *can* make comparisons across people, although a similar logic to that just given makes us unwilling to do so. If we see differences in the *average* scores of individuals, is that attributable to differences in the policy environments they face? Or differences due to family background, etc.? While at least here we do have comparisons we *could* make, putting more than a grain of faith in them is hard to do, since it is difficult to untangle what portion of tests is attributable to policy, family background, etc. Instead, a much more common practice is to again throw away information which we do not think will generate useful comparisons - in this case the between (or average) information across people in the test scores. What is left after obliterating both the cohort level secular age effects and the individual level average differences is the *growth* in test scores for each individual. To the extent policies make a difference in these scores, this residual variation should represent a relatively clean lens with which to view such effects.

While from a conceptual perspective we can make the argument this aspect of the outcome variable is a relatively clean lens to detect policy effects, it begs questions of the baby - bathtub variety, especially in light of the reliability / validity concerns above. After all, if tests are designed to have good reliability properties, *defined* by correlations across people (a point we take up in the following paragraph), and validity is also measured by asking how well the Burt scores characterize different students' outcomes, it is not at all clear that *any* variation should be left after throwing away person and age specific averages. And even if such variation does exist, it is not clear it will have systematic

properties (relating to reliability for a given individual) or good validity properties (relating to whether *growth* in test scores relate to other *better* outcomes such as *growth* in TOSCA scores). In other words, if we opt to use the types of comparisons described above to detect policy effects, then we want to think carefully *which* types of those comparisons will most preserve the good reliability and validity properties the test has across people. Failure to carefully consider the properties of the residual variation may imply that not only has the baby been thrown out with the bathwater, but that we have opted to keep the soapy residue left on the tub as a suitable replacement.

Of course, as our reviewers have emphasized, work by Heise (1969) informs us that this conundrum need not be so vexing: it is quite possible to construct tests with good reliability properties but without the confining high stability properties by recognizing these are distinct concepts. As social scientists, we should play closer attention to all *three* of a test instruments properties. And if we find ourselves needing to utilize fixed-effects, or time differencing methods in analyzing test score outcomes, then we should seek out test instruments with good reliability, but low stability measures so as to at least *allow* the possibility of finding policy effects on such test score outcomes. We elaborate on the formal argument by Heise in the Appendix to this paper.

This leads then to the question of what are the properties of the Burt scores in light of the discussion just given? Table 1 does an elementary decomposition of the Burt scores, from ages 8 to 13, for the full 873 individuals in the sample into the between person variation and the within person (across ages) variation in the sample.[6] We see that simply accounting for average differences *between* children accounts for over half (about 60 percent) of the total variation in Burt scores. In the bottom panel of Table 1 we also account for the other source of secular variation in Burt scores, and that is the age effects. Together, we see that the person and age effects account for about 95 percent of the total variation in Burt scores! Thus, net of the sources of variation with which we cannot credibly identify policy effects, we are left with only about 5 percent of the original variation.

It is probably no accident that the R-squared of the above relationship is 0.95, which corresponds to a correlation of about 0.975. That correlation is approximately the reliability measured for the Burt test holding constant age, and this is roughly the conceptual idea of the regression defining the measure above. The distinction here, however, is that we are not measuring a given child *weeks* apart, but years apart, and hoping that the remaining variation will at least *potentially* be mutable by policy. The similarity of these two measures, however, indicates this may be a grim hope - we need to use comparisons in the

[6] In the work we do below we make substantial restrictions on the sample so that the number of individuals represented will be substantially less than 873. Here, the panel of test scores is allowed to be unbalanced and variables other than just the test score data can be missing, thus allowing many more observations. Since the conclusions here are qualitative only, restricting the samples to be exactly the same is unnecessary and not especially desirable.

remaining data which give us the most 'signal' of what is going on for a given person over time (but net of age). But before coming to that, a parenthetical remark is useful here in noting that overemphasis on test instruments with high reliability, such as the Burt tests, may have the undesirable feature that policy effects are simply undetectable with such tests, even though policy itself may have significant effects on other outcomes of interest, but which remain unmeasured. The very heart of the reliability concept in the data is typically thrown away by researchers in looking for policy effects (of which 'fixed effects' analysis is but one version), so that what is left over may be quite noisy and with no known validity properties, in contrast to the between person variation.

With this in mind, we then ask "What comparisons in the data give us the best chance of detecting effects, given this feature of the data process?" Borrowing a page from the measurement error in panel data literature, we know that the so-called 'long difference' - the Age 13 score minus the Age 8 score - will potentially contain the most signal, under certain assumptions (which we do not go into here - see Griliches and Hausman (1986) for a detailed discussion of these issues). The basic idea is that while the year to year differences observed in Burt scores may be a rather noisy reflection of the child's environment, the difference using the full time span of the data will reflect the broad trends in the environment. In addition, as pointed out by Ron Crawford, when we bring the policy variables into the picture, the exact timing of *when* we should see those policy variables have an effect is not at all clear *a priori*. In the U.S. literature, researchers are forced to take a stand on that issue only because of the short time span of the data for each individual. Here, we can be more agnostic about say, at what age should we begin to see effects on test scores if class sizes at age 9 are reduced?[7] Of course other comparisons could be made, and dependent on the time series properties of the tests and their signal and noise components, these may well be superior. But this simple measure has intuitive appeal as well as serving the dual roles of retaining signal re: policy effects as well as being as vague as possible on the timing of the policy effects. Simply put, given the strong persistence in the Burt scores, it seems implausible to us that the 'higher frequency' (i.e. year to year, etc.) variations in the test scores exhibit much signal. The best bet, if we restrict ourselves to use variation net of person specific differences, are the long run changes by individual.

Having settled on these (long run) gain scores, we next turn to examining their properties. Notice that when we come to examining the relationship of these gain scores with policy variables, if we have a constant in that regression, this is completely analogous to including age effects for secular Age 8 and Age 13 differences, since the constant will absorb the secular long run trend in scores. However, another issue arises, apart from the mean differences in scores by age. Figure 1 returns for a moment to the densities of the *level* of the Burt scores

[7] Another reason to be agnostic on the timing stems from the administration of the Burt tests, which were given near the birthday of each individual, and so the relevant timing of the class size measure would be different for each individual for that reason.

by age. By age 13 a significant fraction of children are near the ceiling score of 110, and so the 'topping out' problem may be especially pronounced for our long difference measure.[8] To check this directly, we plotted the gain in the test scores against the initial test score level at age 8. This is shown in Figure 2, and it reveals a substantial negative dependence (as expected) between the initial level and the subsequent gain.[9] To try to account for this ceiling effect, we experimented with the standard transformation of the log gain (defined as log(Age 13 Score) - log(Age 8 Score)) as a potential alternative. The idea here is that the log gain score measures (approximately) the percentage gain as opposed to the absolute gain. The log gain versus initial (un-logged) level is shown in Figure 3, and it is apparent this does little to alter the story. We decided, therefore, rather than adopt some *ad hoc* 'correction' for the negative dependence between the level and the gain (and the upper censoring in the age 13 score) that we instead understand and appreciate its presence and interpret our results accordingly. In particular, it is important that we understand that the 'high gain' individuals tend to be the 'low level' individuals, and so if we talk about the upper quantiles of the gain distribution, then, in the unconditional distribution at least, we tend to be speaking of the initially low test score level individuals, who tend to come from poorer family backgrounds. The negative dependence between gain and levels in test scores, by the way, is generally a feature of all test scores. It is rarely emphasized because it seems to be a point researchers are uncomfortable with and feel the need to 'correct' for once it is uncovered.[10]

Returning to Figure 2 for a moment, we also looked at the extreme observations, defined as lower than the first percentile and higher than the 99th percentile on the gain scores. These are indicated by the horizontal bars on Figure 2, and these indicate that for the most part, these are observations with extremely low test scores in both waves, or extremely low scores at age 8 only to have extremely high scores at age 13 (and vice-versa in 2 cases). The high-to-low (and the 2 vice-versa cases) observations were found to have largely incomplete data, and so discarded on low overall data quality grounds. The extremely low score observations in both waves were also discarded with the thinking being that these are special/handicapped type students, and so not likely amenable

[8] The key word in this sentence is 'near'. In fact, only 2 students scored a perfect 110 at age 13, so less than 0.5 percent of our analysis sample. However, 51 students, or close to 10 percent of our analysis sample, scored in the range of 105 to 110 at age 13.

[9] The points on the negative diagonal represent the children who nearly 'topped out' at age 13 by scoring in the range from 105 to 110. Though not apparent to the eye, the actual age 13 score declines as the age 8 score declines, even though the negative diagonal indicates no perceptible curvature.

[10] Given that a test score measure is already essentially a latent variable to begin with, it seems especially undesirable to use a method such as a Tobit to correct for the censoring at age 13, in contrast to say using a Tobit to correct for censoring of income. The simple reasoning is that a test score, being a latent variable, can be re-scaled which would then lead to different Tobit corrections. The Tobit is simply not identified. To discuss this issue in this space, however, is outside the scope of this report.

to the standard policy variables we are examining here. From a statistical perspective these extreme observations added a significant component of noise to our regressions, and indeed, the results are more focused when they are not included. In Figure 4 we display the smoothed density for the gain score measures in our sample, with the trimming points again noted (again, integrating up to about 2 percent of the sample). Certainly symmetry about the mean and to a degree normality summarize the resultant density well, especially when trimmed of the long right tail. In Figure 5 we show the density for the log-gain score measure for just the trimmed sample. Even with the trimming, the log-gains show a significant right skew, and this re-emphasizes our choice to opt for the unadulterated gain scores (but trimmed of outliers) as our dependent variable, so as to have the mean and the mass of data coincide.

3 The Nature of the Possible Policy Effects

The earlier reports by LECG and the accompanying referee reports have illuminated well the assumptions underlying the standard techniques used in the school quality - student achievement literature. As such we will only briefly repeat those issues here. By far the most common tool to examine policy effects on test scores and other achievement outcomes with non-experimental (or observational) data is the so-called 'value-added model'. In its simplest form, this is given by the regression:

$$\Delta TS_{it} = \alpha + \beta CS_{i,t-1} + x'_{i,t-1}\gamma + u_{it} \tag{1}$$

and a slightly less restrictive version by:

$$TS_{it} = \alpha + \delta TS_{i,t-1} + \beta CS_{i,t-1} + x'_{i,t-1}\gamma + u_{it} \tag{2}$$

Both regressions try to account for the fact that the samples used to estimate such relationships typically initiate somewhere in the middle of a child's schooling career, and so both regressions 'hold constant' (or take as given) the initial test score performance (denoted as $TS_{i,t-1}$) as a summary of the past inputs from both the home and the school. Then given some schooling inputs (denoted as $CS_{i,t-1}$) and family inputs and control variables (denoted as the row vector $x'_{i,t-1}$) after the initial test score performance (together with a stochastic component), the resultant test score is given as TS_{it}. The assumption on timing is that tests are given at the beginning of a 'period' and the inputs occurring at the middle and the end. Implicit in the regression is also the notion that the effects of schooling inputs have 'high frequency' effects in that they manifest themselves the very next period. Of course, even if that is true, a longer difference measure would still detect the high frequency effects, albeit with lower power. In addition, a longer run relationship, as we discussed above, has the virtue of also detecting more lower frequency effects. If the timing is correctly

specified, then the high frequency model is clearly preferred, but that is a rather big if.

The value-added model does have the virtue that by comparing the growth (or change) in test scores with the level of school resources, it throws away the permanent (or between person) component of test scores. While it does so ostensibly for a lack-of-data reasoning, as we discussed above, it appears undesirable outside of an experimental setting to utilize that variation to draw inferences about policy effects. However, the value-added model also simply *assumes* the nature of the production function of student achievement, albeit in a rather intuitive way. To put it as briefly as possible, the production function assumes, for what started with reasons owing to incomplete data, that school resources affect the subsequent *growth* of student achievement. While that certainly does not appear *un*reasonable, it is possible to imagine alternatives which may also appear reasonable. Consider, for example, that the key to academic success is really just 'learning how to go to school' (or just a 'clue-in' effect). In that case, we might expect to see an influx of resources generate a once and for all jump in performance, with no further enhancement in performance possible from more resources being directed at the student later on. In that case, a high frequency value-added model would miss this jump in the intercept unless the sample window of time contained the intercept shift. A low frequency model, however, would still pick up an effect (albeit a muted one) if at least one endpoint straddled the intercept shift. The problem is that if indeed effects are 'one off' after which the students in the low resource environment and the high resource environment continue on parallel trajectories, the value-added model would only compare the slope of the two trajectories, and so, as a consequence, fail to detect an effect.

This discussion is highly relevant because in a true experimental study of pure class size effects conducted in the United States in Tennessee (called Project STAR), precisely this type of intercept-effect of class size was found. Whether or not this represents an effect of interest to policy makers has been an issue of some debate - see the papers by Alan Krueger and Eric Hanushek for pro and con views on this point. But both authors agree that the effects which *were* found in the STAR project are largely undetectable with the value-added model. However, because the project involved students at the very beginning of their schooling careers, as well as randomly assigning students to different class sizes, the value-added model was not needed to account for a missing data problem. Thus more transparent methods could be employed. In particular, Krueger estimated a 'levels on levels' model of test scores on the exogenous portions of class sizes:

$$TS_{it} = \alpha + \beta CS_{i,t-1} + x'_{i,t-1}\gamma + u_{it} \tag{3}$$

In this case the stochastic error u_{it} captures the other omitted pre-histories and unobserved factors, but the estimates of β can be taken as unbiased due

to the randomization of the students into varying class sizes, and so by design orthogonal to any omitted variables.

A referee's comments on earlier work by LECG noted that the specification used by Krueger to uncover 'intercept' effects (i.e. 'level on level' estimates) could also be applied to observational (i.e. non-randomized) data *if* the unobserved pre-histories could plausibly be controlled for via individual fixed-effects. From the perspective of what is missing (early schooling resources, etc.), this would seem unlikely to be true. But given the time-series properties of the Burt scores, with a large cross-sectional (and so fixed) component, this is probably not a bad approximation. In this case, the above model used by Krueger can be amended by:

$$TS_{it} = \alpha + \beta(CS_{i,t-1}) + x'_{i,t-1}\gamma + f_i + u_{it} \tag{4}$$

where the term f_i is the purely cross-sectional component of Burt scores that may be correlated with class sizes, but which is unobserved to the researcher, and so captures the observational nature of the data process. Then *any* transformation of this model which differences the model across time will eliminate the fixed-effect and allow us to obtain consistent estimates of β, even with non-randomized data, if our assumptions are true. One such example is simply first-differencing the data:

$$\Delta TS_{it} = \pi + \beta\Delta(CS_{i,t-1}) + \Delta x'_{i,t-1}\gamma + \Delta u_{it} \tag{5}$$

It was essentially this idea that led to this current project. This last equation has the same dependent variable as the value-added model, namely the growth in test scores for an individual. But now the policy variable of interest is the *change* in class sizes, since it is the 'switchers' in class size that identify 'intercept' effects with observational data when the endogeneity in class sizes is constrained to run entirely through the fixed effect. But as we said above, *any* time-differencing operation will eliminate the fixed effect, including say, the longest difference allowed by a panel of time-length T for each individual:

$$\Delta_T TS_{it} \equiv TS_{i,T} - TS_{i,1} \tag{6}$$

where 1 and T represent the first and last observation for each person in the sample. We know from the literature on misspecfication in panel data, such as measurement error (see Griliches and Hausman (1986)), that this so-called 'long difference' may have desirable properties *vis a vis* the first difference, under certain assumptions. In our setting here, it is clear that given the high persistence in the Burt scores for each individual, expecting that the 'high frequency' movements in the Burt scores from year to year to contain much true signal is likely asking a lot of the within variation of the Burt scores. In addition, the low frequency or long run (from age 8 to age 13) growth for an individual may extract the most signal allowed by the sample.

As we alluded to above, the long difference also has the attractive feature that if we are uncertain *when* we should expect to see a movement in test scores from a change in class sizes at age 8, say, the long difference allows us to be as agnostic as possible on this issue. This brings us to a discussion on *how* we should model the right hand side of the above equation as regards the policy inputs. Here again the experimental literature from the U.S. is useful, but here it is the *interpretation* of the experimental results (indeed, it is the critique offered by Eric Hanushek) that we draw on. In the Project STAR experiment, it was found that not only was there a once and for all jump in test scores the first time a student enrolled in a small class, but that furthermore, those gains eroded or faded away if the child subsequently returned to a small class (implying that our first difference model suggested above is not correctly specified as regards the class size variable so as to mimic the dynamics exhibited by the experiment). Hanushek interpreted this as evidence that only *persistent* class size reduction policies 'matter', or at least allow the small-class children to retain their initial gains.

Because the dual issues of correctly specifying the dynamics of the class size inputs (class size reductions have only a one time effect, effects are not necessarily symmetric for a given magnitude *gain* in class size as for a *reduction*, fade-out effect, etc.), as well as the appropriate timing of the class size effects, we opted to first set our sights on a lower target. While Hanushek took the Project STAR evidence as confirming the conclusion he has held for almost twenty years of research - that class size reductions have no *systematic* effect - all parties examining the STAR evidence agreed that *persistent* class size reduction policies had persistent effects. So our (less heroic) question became: "Can we detect effects of persistent policies with observational data such as the CHDS, analogous to those of Project STAR?" This also made sense as a point of departure because of the persistence in the Burt measures themselves. Viewing the model purely from an analytical and statistical perspective does imply we should able to estimate the class size reduction effects off the high-frequency movements in test scores. Intuitively, however, once one looks at the large persistence in Burt scores, it seems highly unlikely that short-run movements in policy could have detectable impacts on the year to year fluctuations in test scores. Putting this all together implies we should look at the longest possible trend allowed by the data for each individual, and at least start our investigation of policy effects by looking at persistent policies, and then move to higher-frequency investigations.

Finally, one other point on the *type* of effects we will be looking for. Returning to the plot of the gain in test scores versus the level (Figure 2), it may well be that the effects of class size reduction policies are not constant for all students. In particular, we may anticipate that, for those students for whom their grade 8 score is high (and so, on average their gain score will be in the lower part of the distribution, all else equal) it will be hard for class size policies to have much *additional* effect. For this reason, as well as just examining plots of the raw data below, we examine quantile regressions of the class size effects,

which allow the coefficients to vary at different quantiles of the conditional gain score distribution. Because of the topping out problem, intuitively we would expect to see larger effects in the higher conditional quantiles than in the lower conditional quantiles.

3.1 Measuring the Impact of Schooling Inputs on Early Adult Outcomes

The methodology to detect class size effects just outlined was motivated by one concern - to mimic the experimental results found in the U.S. via Project STAR. The large mass of the school quality literature almost always takes a different tack in first considering a type of policy effect, then seeing if it can be found. But in having access to a dataset as unique at the CHDS in its longitudinal design, we were able to abstract from many of the assumptions inherent in that approach and ask the somewhat weaker question of whether, with rich enough *observational* data (i.e. no randomized design to the survey process), we could detect effects of the kind found using *experimental* data, and that answer appears to be yes. But we acknowledge that the type of effects we have found do not map into a policy question of natural interest (such as what one reviewer suggested to us as "for every x years a student spends in a small class as compared to a large class, their performance is enhanced by y percent"). The methodology was completely driven by asking, if we are given only test scores as outcome measures, how might we expect class size inputs to affect them in a detectable way?

However, since the initial draft of this report was written, the age 21 data for the CHDS data has been released and made available to us (it was collected in 1998). These data contain information on a number of outcomes that are of direct interest to policy makers, and so we need no longer rely on the external validity properties of a given test instrument in analyzing test score outcomes. In addition, since these outcomes such as years of completed education, unemployment, arrest information, and so on are not subject to the severe persistence properties of the Burt Word Reading scores, we need not confine the methodology to mimicing the Project STAR analysis. Instead, we can ask if, conditional on initial performance (or early childhood and parental inputs - summarized by the Burt score at age 8) class size inputs sampled as late as possible impact these age 21 outcomes in a systematic way. In short, our methodology used to analyze the age 21 outcomes is, loosely speaking, like a 'large time window' value-added model of sorts. It isn't quite a value-added model (as was discussed in the previous subsection) because the lagged outcome variable (the Burt score at age 8) is different from the age 21 outcome which we study (one of the early adult outcomes just listed). In contrast to our methods used to analyze the Burt scores, our approach now we think is much more intuitive as well as being more directly policy relevant.

We are still left with the issue of the appropriate *timing* for the class size

measure(s). We have chosen to use the class size as of age 13 for a variety of reasons. One, as of the current data extract, this is the highest age for which we have class size. We have done some work outside the scope of this report on investigating the time series properties of the class sizes across ages, as well as investigating which children get assigned to which class size. To summarize some of that work, the class sizes before age 12 are much more collinear and driven by the initial Burt performance than the age 12 or age 13 class sizes. Furthermore, as the child has changed schools as of age 13, the age 13 class size is likely to be the best indicator of the class sizes from ages 14 to 17, for which we lack class size data. Thus, for the dual reasons that the age 13 class size is the best proxy we have for high school class sizes as well as it being less due to the compensatory class size assignment mechanisms (whereby students with lower initial performance get assigned to small classes), our measure of class size in analyzing the age 21 outcomes will be the age 13 class size.

To distill this discussion to its conceptual points (and to compare it to the value-added type regressions above) let us denote the age 8 time-period as '$t-1$', the age 13 time-period as 't', and the age 21 time-period as '$t+1$', and introduce the notation $L_{i,t+1}$ as a labor market outcome as of age 21. Our empirical specification is then:

$$L_{i,t+1} = \alpha + \beta(CS_{it}) + x'_{it}\gamma + \delta(TS_{i,t-1}) + u_{i,t+1} \tag{7}$$

where the lagged test score is included to proxy for early childhood inputs and child-specific tendencies in the labor market outcomes which would be unrelated to policy impacts. The timing on the class size variable is chosen at as late an age as allowed by our data (age 13) so as to proxy the best for the class size environments in junior and senior high faced by the student. In this sense, we have to be careful that even though the *measure* we use is just class size at age 13, because of the serial correlation in class size over the child's school history, it may be proxying for class sizes at the later ages, and therefore its coefficient in the regressions should be interpreted as such.

3.2 Relation of Our Methodology to Hierarchical Linear Modelling (HLM)

The empirical method we used to analyze the possible class size policy effects on both the childhood outcomes of Burt test scores (to age 13) and the early adult outcomes as of age 21, while perhaps intuitive, may appear to be *ad hoc*. Of course, we spent considerable effort in our discussion above to indicate why (i) we wished to use *changes* in outcomes (i.e. either the growth in Burt scores from age 8 to 13 or the age 21 outcome holding constant the age 13 Burt score) in order to abstract from the possibly confounding factors (such as family background, etc.) affecting the outcome *levels*. (ii) We used the broadest time

(or age) window allowed by the CHDS data due to the persistence in the Burt test scores to construct these changes. And (iii) we relied on measures of the class sizes the child was exposed to which captured the *permanent* or average component of the class sizes since we saw this as one of the primary conclusions of the U.S. literature which used experiments to deduce these effects. While we feel this econometric specification is reasonable given both the properties of the observed variates in the CHDS, as well as building upon what we have learned in the school quality literature, notably the experimental kind, we agree it is useful to give a discussion that ties our methodology to more orthodox methodologies found in the educational, sociological, and the psychometric literature. One of the most widely used of these methodologies outside of economics is Hierarchical Linear Modelling (HLM) articulated by Bryk and Raudenbush (1992).

Econometricians will recognize HLM as a two-stage variant of a fixed-effects regression for an '*i, t*' type of panel. In the first stage, no regressors are used, but the dependent variable of interest (in this case the level of the Burt scores for each individual from age 8 to age 13) is regressed on a person specific intercept and a person specific trend (the trend being denoted as T_t):

$$y_{it} = \alpha_i + \beta_i(T_t) + u_{it} \tag{8}$$

In the second stage, the cross-sectional collection of these N intercept and slope (or trend) estimated coefficients are then regressed on a set of k cross-sectional regressors, denoted as x_i' to separately explain the person-specific estimated intercepts and linear trends:

$$\hat{\alpha}_i = \pi + x_i'\delta + e_{it} \tag{9}$$

$$\hat{\beta}_i = \kappa + x_i'\eta + v_{it} \tag{10}$$

The objects of interest from this procedure are the fitted values of δ and η obtained by some suitable form of weighted or generalized least squares applied to equations (9) and (10).

Since the person-specific intercepts α_i are estimated (or conditioned on) in the first stage regression, the estimates of the influence of the regressors x_i' in the second stage, η, are estimated net of the purely cross-sectional variation. Therefore, whether or not in the second stage the researcher wishes to model the influence the effect of the regressors on the intercepts as well as on the slopes, or focus on the slopes only, the N estimated slope parameters will not be biased if we simply wish to condition on the N intercept parameters α_i, and so effectively throw away the purely cross-sectional information. As we discussed in Section 3.1 above, whether or not the researcher is using HLM or a variant of the value-added model as we are, due to the observational nature of the CHDS data as well as the timing of the class-size (policy) variables, we cannot plausibly argue the association of the regressors with the fitted intercepts has anything approaching

a causal interpretation, and so by either methodology, we opt to simply throw away this information.

So the question of comparison of HLM to our methodology comes down to comparing the regression of the fitted person-specific trends on a set of purely cross-sectional regressors versus our utilizing the 'long difference' of the Burt scores regressed on some reduced dimensional sequence of the class size measures for each age. In fact, our method as described in the text can, in this light, be viewed as an inefficient variant of HLM. The reason for the inefficiency stems from the fact that our methodology uses only the first and last (age 8 and 13) Burt scores for the age range for which they are valid, whereas HLM assumes a *linear* trend, and then uses all six Burt scores to estimate the trend. HLM also has the advantage that if the Burt score is missing for any of the ages, including these endpoints, a fitted trend for that individual may still be obtained. The only way in which our method might be superior is if the underlying trend is in fact non-linear over these ages, although then *what* slope (since there is a multiplicity of them for a non-linear profile) to use is no longer clear, and both methods will suffer from that uncertainty.

However, even in the case of linearity, due to the strong persistence in the Burt scores, then as an empirical matter, the degree to which the endpoints determine the linear trend might be quite good. Therefore, the efficiency gain obtained by using all six data points for each person in HLM might as a practical matter be quite small relative to our method. As economists, we have a slight preference for our method as the discussion given above delivers (we hope) a reasoned discourse of how we modify the value-added model used so frequently by economists working in this area. However, we are also encouraged that our methodology, designed to fit the realities of the data as well as capture the findings from the experimental literature on class size effects, meshes so well with the HLM methodology that is so widely received (and used) in sociology and related fields. We hope this discussion makes clear the close relation between these two methods.

In the next section, we take up the issue that arises both for our method as well for HLM, and that is how to reduce the T dimensional vector of observations on the class sizes observed for each child to some summary measure of the class size inputs each child experiences over the ages of 8 to 13. We do not think our approach is exhaustive in this regard (many other methods of reduction of the multiplicity of inputs could have been tried on our part) but our attempts here, in an effort to take up the *way* in which the Project STAR evidence appears to tell us the class size effects are working, are to use a reduction of the class sizes that get at the *permanent* component in the class size measures. This is quite different from a standard value-added approach which uses the year-to-year fluctuations in class size to generate short-term fluctuations in academic outcomes.

4 The Measures of Class Size Inputs

As alluded to above, we organize this section by the *power* of the statistical comparisons. What this means is that we start with the comparisons by class size that give us the best possibility of detecting effects, based on the U.S. experimental literature, and then work backwards to utilize less 'distinct' comparisons based on what we learn from the data using these 'extreme' comparisons.

In a preliminary report leading to this project, LECG produced some plots of average test scores by age broken out by 'large', 'medium', and 'small' class size categories to accord to some similar plots done by Hanushek and Krueger with the Project STAR data. The class size designations corresponded to students who were *always*, for the time period of the sample, in the same class type (large, medium, and small), and not surprisingly, only a small fraction of all of the sample are in the same class type. But the small sample issue aside, the plots from the CHDS data looked strikingly similar to the analogous plots from the Project STAR data. Given the age effects, and also for the multitude of reasons discussed above, as we move towards putting that graphical analysis into a regression format, we simply use the Age 13 Score - Age 8 Score as the dependent variable of interest.

We altered the designations of 'large', 'medium' and 'small' slightly from the preliminary report of LECG in order to enhance the number of observations in each category to enable us to look at the distributions of test scores by each category, as well as just the mean effect. The magnitudes of the point estimates are smaller with this re-categorization (not surprisingly, as the comparisons are less distinct), but owing to the growth in the cell sizes, the associated t-statistics remain roughly the same. We should mention the qualitative nature of the results given below are quite robust to alternative definitions of the class size categories, and not some artifact of their construction. We used the 10th percentile (which is 24.5) and the 90th percentile (which is 32.8) of the overall class size distribution as the cutpoints.[11] To create the *permanent* class size categories, we then require that the child be in the same class type for all 6 years of the sample. For this reason, the dummy variables for permanent small, medium and large classes do not add to one, and in our regressions we use the rest of the sample (who switch at least once in a class size category at some point during the time frame of the sample) as the comparison group.

However, before we even move to the regressions to get precise measures of the magnitudes of the effects, it is useful to look at raw plots of the densities of the gain scores broken out by the permanent class size categories. The results of this are shown in Figure 6. Again, the vertical lines here represent the outlier points of 13 and 68 in the gain distribution. These kernel density plots are instructive because not only do they indicate the gain scores rank according

[11] These are for the full sample - in reality, due to missing data in the analysis sample, the percentages end up being less than 10 percent in each of the two extreme cells, but there is no need to be precisely at 10 percent since this is an *ad hoc* definition to begin with.

to our intuition from low to high as the class size categories go from large to small, but they also inform us as to the distribution of these effects. Owing to the fact that the low gain students are the high initial test score level students, visual inspection of this figure indicates that the effects in the left tails of these distributions are much less dramatic than when viewing the right tail. This is especially true when contrasting the large class density to the small class density. Basically, this graph summarizes all of the results which will follow below, but the regression based results have the virtue of quantifying these effects, as well as clarifying if the visual differences we see here are statistically significant or not. But this graph has the virtue that the effects we report below are apparent to the naked eye, and not some statistical artifact.

Since our dependent variable is the *difference* in the test scores, there are not many demographic variables in our data which are needed as controls. Thus, for parsimony as well as for ease of interpretation, we keep the number of control variables to a minimum. The one notable exception that we detected to this observation is that female students tend to have significantly lower growth in scores than do males. We also included indicators for whether the mother and father are Maori - when included jointly, these ethnicity variables tend to be insignificant and roughly of equal magnitude. The point estimates indicate that, all else equal, children with Maori-identified parents tended to have slightly higher gain scores, but not significantly higher than 0. Finally, we also included the change in family income during the sample. Not surprisingly, unlike the correlation of the level of income with the level of test scores, the correlation in the changes of both measures is quite small and insignificant.

The left column of Table 2 reports the results of a simple OLS regression of the gain scores for each individual on the control variables just discussed as well as the 3 dummy variables for the permanent class size categories. From a statistical significance perspective, the effects of always being in a large or small class during the sample have significantly negative and positive effects respectively at about a significance level of 20 percent - the t-values are about 1.3 in magnitude. The point estimates are about equal and opposite for the small and large class size categories at about 3.2 in magnitude, and we devote a separate section to the interpretation of these estimates below. In the brackets below each class size category, we indicate how many observations are in each permanent class size categorization. Only 70 of the 569 observations are in any of these 3 cells, and the resulting small cell sizes, with about 20 observations in each of the two extreme cells, may be partly behind the lack of statistical significance.

While the regression just presented has appeal from the power of the *conceptual* idea behind it, it lacks power because of these small cell sizes. The conceptual idea is to track students who are in large, medium, and small classes for the 6 years at the beginning of the sample, and then look at their resultant growth in test scores. The problem is that the conceptual idea uses only about 12 percent of the available sample. To counter this, we next think of a

conceptual idea that uses the full sample. Consider now categorizing children by the *average* class size they face from ages 8 to 13, and again break this categorization into Small, Medium, and Large categories.[12] Now, because the three categorizations exhaust the full sample, we must omit one of the categories as a reference group. The results of this regression are in the second column of Table 2. Not surprisingly, as the conceptual experiment is not as sharp as that defining the regression in the first column, the point estimates decline in magnitude somewhat.[13] Indeed, in a plot analogous to the kernel densities by the Permanent class size categories in Figure 6, Figure 7 displays the kernel densities for the three class size categories defined by the Average class sizes. It is visually apparent that the differences based on this conceptual experiment are not nearly as striking as in Figure 6. However, since the *statistical* exercise utilizes more of the sample, the standard errors decline by about the square root of the ratio of the cell sizes, and so the qualitative conclusions are not greatly changed when looking at the two columns.

It is instructive to do a bit of interpretation of the comparison of the magnitude of the coefficients from the two columns at this point, and then below discuss more of what the estimates mean for policy. The average class sizes for the Permanent Small, Middle, and Large class size categories are 19.0, 29.9, and 33.8. For the Average Small, Middle, and Large class size categories, the analogous averages are 21.2, 29.7, and 33.2. In both regressions, the drop in average class size from being categorized from large to medium is on the order of 3.7, and the reduction in class size of going from a medium to small class is about 8.5 students on average. The fact that the marginal reduction in class sizes in going from large to medium size classes is smaller than going from medium to small classes is an important point. It implies that while the point estimates on the large and small class size dummies in Table 2 are nearly equal, when they are converted to a *per student* reduction in class size, the marginal effects (on the gain in test scores) are greater for the large classes than for the small classes.

But it is clear both of these two ways of measuring the *persistent* aspect of class size policies have their relative merits. In our next table we examine the heterogeneity in the class size effects by running quantile regressions using the same specification as was used for the conditional mean regression in Table 2. The quantiles here refer to the conditional quantiles of the distribution of the gain (i.e. Age 8 to 13) scores. Table 3 presents only the coefficients for the class size categories at the 9 deciles from 0.1 to 0.9. The use of quantile

[12] The cutpoints for these categories were taken as 24.5 and 32.2, which differ slightly from the cutpoints used for the permanent class size categories, which were taken as 25 and 31.5 respectively. This small difference does not affect the qualitative results at all, and the failure in the consistency of the two definitions only derives from how the data were arrayed when the deciles were examined.

[13] As noted in the Note accompanying Table 2, the full set of results are virtually completely unaltered in the first column if, analogous to the second column, we restrict the coefficient on the Medium Class category to be 0, and so include it as part of the reference group.

regressions is motivated by several factors - foremost among them was the idea that gain scores are, by construction, smaller for those near the top of the initial levels distribution than for those near the bottom. Rather than 'correct' for the ceiling effects through some *ad hoc* method, we reasoned this should be visible via quantile regressions, with lower effects of class size policies in the smaller quantiles, and larger effects in the larger quantiles (since the gain distribution is negatively related to the levels distribution). Not surprisingly, the quantile estimates are not very precisely determined. But if we examine the point estimates in the top panel, we see for the Large Class size category, the reductions in test scores gain are on the order of 5 points in the upper quantiles, but on the order of 2 points in the lower quantiles. Similarly, for the Small Class size category, we see the increase in the gain scores is about 6 points in the upper quantiles, but on the order of 2.5 in the lower quantiles. This is not precise, but it does indicate that there may be some degree of heterogeneity in the class size effects by decile that is possibly arising from the ceiling effects of the Burt scores, and/or that Burt scores are more 'mutable' for students who are initially performing at lower levels. Distinguishing between these two explanations does not appear possible with this dataset.

In the bottom panel of Table 3 we also report the quantile regression coefficient estimates for the conceptual experiment based on the average class size categories. Here the heterogeneity is much less apparent, perhaps owing to the extreme lack of precision in the estimates. We should mention, at this point, that this dulling of the sharpness of the results as the conceptual experiment is blurred generalizes to work we do not report on here. In the initial class size categorizations used by LECG, the effects at the upper quantiles far dominated those for the lower quantiles. But the categorizations were so small that even though the point estimates were significant, they did not inspire much confidence. The results reported here are much more broadly representative of the overall picture of the possible range of results a researcher would find with alternative definitions of categories and/or thought experiments.

However, from a statistical perspective, the contrasts presented in Tables 2 and 3 are not the sharpest available. In some sense, the results are presented there in a format most amenable to the idea of comparing a student in a Small or Large class to the 'Average' student. However, since power is a key problem in detecting policy impacts on Burt scores, we now present much the same information, but with standard errors and coefficients pertaining to comparing students in Small Classes to students in Large Classes. The results for the simple regressions are displayed in Table 4. To make the Small to Large comparison meaningful for the Permanent regression in column 1, we restrict the sample to only those students who were in one of the 3 Permanent class size categories in Table 2 (for which N=70). The second column contains the same information as in Table 2, but uses the Large Class category as the reference group. The Small to Large comparison in the first column represents a drop in the average number of students in the class by about 15, and in the second column by about 12. The

Small to Large coefficient estimate in column 1 (7.9) is significant at better than a 5 percent level, and the coefficient in the second column (3.5) is significant at better than the 10 percent level, which is also true for the Medium class size coefficient of 2.2. This latter estimate corresponds to an average reduction in class size of about 3.5, reflecting again the concavity in benefits to class size reduction policies.

For Table 5, we repeat the quantile exercise of Table 3, but this time only for the Average Class Size categories. In keeping with the last table, the Large Class size category is our reference group here. For the Small Class effect, we see little in the way of systematic heterogeneity across the quantiles. For the Medium Class size effects, there is slight evidence of greater effects at the higher quantiles than for the lower quantiles, but a precise conclusion on this front cannot be made. The conditional mean effects for these categorizations reported in Table 4 appear to be adequate representations of the effects across the deciles. The net conclusion we draw from all of the quantile investigations is that the topping out aspect of the Burt gain scores does not appear to be affecting our class size estimates in ways that we can detect. To the extent their is a bias introduced by the gain score constructions, it would appear it would have to be fairly uniform across the deciles, which seems unlikely, since the topping out is heavily related to the deciles of the gain distribution.

In addition to shedding some evidence on the importance of the topping out for our results (as well as simply speaking to heterogeneity in the class size effects more generally), we also sought to use the quantile regressions for evidence against our identifying (or exogeneity) assumptions for these regressions. Recall from our discussion of just the test scores that we did not wish to use the pure cross-sectional variation in test scores (the fixed effect), because it seems plausible (indeed probable) that the correlation of the fixed component of test scores with class sizes is almost surely not a causal relationship, but likely more just *reflective* of how students are assigned to different class sizes on the basis of their test score performance. In order for the above regressions to have a causal interpretation, we have to assume the average or permanent class size categories are orthogonal to the long run *growth* in residual test scores.

On the face of it, this seems reasonable, especially since the residual variation in test scores is so small, it is difficult to imagine the behavioral process whereby this residual variation plays a significant role in average class sizes. The reality of the gain scores, however, again contaminates such a simple explanation, since we have already shown the gain to be correlated with the level of test scores. Since the 8th grade test scores *are* likely to be used in determining subsequent average class sizes, even though our analysis mutes the presence of the level of test scores substantially (as opposed to using the levels of the test scores themselves, for example), this could invalidate our identification assumptions. However, since this correlation of the gain with the level of scores varies by quantile of the gain distribution (in particular it *declines* as the gain increases - this may not be quite evident to the naked eye from Figure 2), a quantile

regression analysis serves as a useful *check* to see if the use of the raw gain scores is invalidating our mean regression results.[14]

We should emphasize two points: 1. Quantile regressions are certainly not a *correction* for the endogeneity of class sizes *vis a vis* the gain scores, and 2. This "test" is rather heuristic and not of great power. It is entirely possible to have endogenous class sizes, and yet not detect it by looking at quantile regression estimates. We are simply exploiting the fact that in this case, we anticipate the endogeneity to be different at different quantiles of the gain distribution (and more pronounced for the lower quantiles) if the endogeneity is working the way we think it might be. To the extent, however, we see the largest effects at the *upper* quantiles, then, if anything, the endogeneity appears to be biasing down our results at the lower quantiles. Thus, if we could somehow 'correct' for the endogenous class size assignments (a rather dodgy business in itself), we might expect to see *large* mean regression estimates, and a growth in the coefficients in the lower quantiles relative to the upper quantiles. For the purposes of this project, however, we will interpret our mean regression estimates as, if anything, *lower bound* estimates of the true effects, with no apparent evidence of upward bias in our estimated effects.[15]

4.1 Comparison of Our Results to HLM

As we discussed in Section 3, owing to the strong persistence properties in the Burt test scores, using the 'long difference' (age 8 to 13) test score gain as our measure of achievement may well yield empirically similar results to using the person-specific trend for the second stage of HLM. Here we briefly discuss our results with this approach. Our suspicions were borne out, as the long-difference gains were highly correlated with the person specific gains. Figure 8 shows the plot of the trend coefficients against our gain measure, and it is visually apparent that there is little difference between the two measures (the coefficients of variation for the two measures were virtually equal).

[14] A number of assumptions will be needed to actually *sign* the bias or the bound implied by our estimates, as opposed to just detecting heterogeneity. Since we do the bounding exercise below, we should spell out our necessary assumptions: 1. The true effect of class size is negative (and so more negative estimates in magnitude are *overstating* the effect of class size) and 2. The covariance of the fixed, or cross-sectional component of Burt scores with class size is positive. Since the fixed effect correlates positively well with family income, this may seem counterintuitive. But work from the U.S. indicates strong redistributive and compensatory elements in setting class sizes, indicating this covariance being positive may be reasonable. Indeed, our work with the CHDS discussed in section 7 on the gender differences in achievement and class sizes assignments indicates this assumption is reasonable for these data. But frankly, this covariance could go either way.

[15] Of course, this is all predicated on our assumptions about the underlying processes being correct. Lacking any truly exogenous variation in class sizes, say by some fantastic instrument or a genuine randomized experiment, we cannot assess the validity of our identifying assumptions. They derive from our prior work in this area and seem reasonable in that context, but in reality, they are simply a tautology needed to make the analysis proceed.

We computed the person-specific trends by running 787 person-specific regressions for the sub-sample with at least 3 observations on Burt scores out of the possible six scores from age 8 to 13. Owing to the low degrees of freedom, and the number of near-exact fits, we used the number of non-missing observations for each individual rather than the inverse standard error as the weight for the fitted trend coefficient in the second stage regressions on the class size measures. We used the same sample restrictions as for Table 4, and the results for our analogous HLM exercise are reported in Table 6. Ignoring the fact that we delete the few outliers with a gain score from age 8 to 13 of 13 points (and so create a non-zero intercept in our working extract) the results in Table 6 may be compared to those in Table 4 by using the conversion factor of 5 (since we have 6 time periods and so 5 changes between them). When this is done, the conclusions from Table 4 are almost completely unaltered, in that the coefficients and their statistical significance (and confidence intervals) are essentially the same.

While HLM might be thought to be more efficient, owing to its use of *all* of the test scores recorded from age 8 to 13, as opposed to just the endpoints in our methodology, the reader can see the degree of (converted) imprecision is essentially the same in the two tables. The 'long difference' methodology discussed in this paper has the virtue of being adapted from the value-added model in a way that accounts for the high degree of persistence in the Burt scores, and consequently attempts to extract the maximal degree of signal from such data. As the value-added model has a strong foothold in the economics of education, it is worthwhile exploring the mesh between that model and HLM in a way that accounts for the time-series properties of the test score measures.

5 Interpretation of the Class Size Estimates

The usual problem with utilizing test scores as an outcome variable is that the coefficients themselves mean basically nothing. They require some metric to yield to interpretation. A naive approach might take the coefficient from Table 4 on the Medium Average class coefficient (column 2) of 2.2, and compare that to the average growth in test scores from Age 8 to Age 13 in the sample, which is about 40 points. However, since growth in scores by age is completely secular, this comparison is completely meaningless, and gives us no gauge of whether the effect is large or small.

We turn first to the usual practice in the school quality literature, which is to divide the coefficient estimate by the standard deviation of the dependent variable, so that the scale of the (essentially latent variable) test score divides out. In this way we can compare the effects estimated here to other estimates in the literature, even if it does not allow comparison with something of direct interest to policymakers. For our analysis sample, the standard deviation of the gain scores is 10.6. Now as we noted in the previous section, the extra gain

in going from a Large Class to a Medium Class is larger than the extra gain in going from a Medium Class to a Small Class *per student reduction*. So, for example, for both the Permanent and Average categorizations of the Large to Medium Categories, this reductions is roughly 3.5 students. Using the Average (and so the more conservative) estimates, this implies a gain of about 2.2 points. Going from Large to Small, however, represents a drop in average class size of about 12 students and yields a gain of about 3.5 points. The effects appear to exhibit diminishing returns to lowering class size somewhere in the range from 21 to 29 students.

Keeping then with the Large to Medium average class size reduction, the reduction of 3.5 students leading to a 2.2 gain, implies a 1 student reduction leads to about a 0.63 gain. Relative to the standard deviation, this then implies an 'effect size' of about 0.06σ. Relative to the U.S. literature, this is large: the Tennessee experiment, for example, yielded effect sizes on the order of 0.04 to 0.03, although the context there was a reduction from 23 students to 15 students, and if the diminishing returns to class size reductions were of the same nature as for CHDS (a debatable point, certainly), this might account for the slightly larger CHDS effects relative to the experimental effects. In fact, the Large to Small class size reduction of 12 students in the CHDS leads to a computed effect size of 0.03σ, entirely in accord with the Project STAR evidence. Drawing on the work of Hanushek and others, these effect sizes are also on the order of some of the more recent and novel instrumental variables estimates of effect sizes deriving from the U.S. literature. Thus, even the Large to Small class size reductions, while less 'bang for the buck' than the Large to Medium reductions, still appear to yield large results by comparison with the norms from this literature.

5.1 Maori / Non-Maori Test Score Gaps

While these computations allow comparison with the rest of the literature, they do not help much in the way of policy discussions, since they do not relate to policy outcomes. To enable this, we turn to two comparisons which can be made while still utilizing test score outcomes, and these are simply gaps in test scores by ethnicity (Maori / Non-Maori) as well as by family income (top 10 percent versus bottom 10 percent of the family income distribution). Defining a child as Maori with either birth parent as Maori yields us 59 children, with the remaining 510 being classified as Non-Maori. The respective average Age 8 test scores for the two groups are 43.3 and 46.3. For Age 13, the corresponding averages are 83.9 and 86.2. The initial gap in test scores is about 3 points, and not surprisingly, this gap has closed somewhat by Age 13 to 2.3. The average class sizes for Maori children is roughly 0.7 students larger.

We can now ask two questions, the first relating to a real policy, the second relating more to a counterfactual policy experiment. First, if we eliminated the Maori / Non-Maori class size gap for the *average* class sizes from Age 8 to

Age 13, we might expect to see the Maori / Non-Maori test score gap close by about an additional 0.5 point, or about 16 percent of the initial gap in scores. Second, we could imagine a more radical policy which would lower class sizes by say 3.5 students for Maori children in large classes relative to Non-Maori children. This would lead to an estimated 2.2 point gain in Maori test scores by 13 (extrapolating the point estimate outside of its 'proper' range), which would then close the Maori / Non-Maori test score gap by about 70 percent.

Whether or not such a reduction is 'worth it' depends on the associated costs of the necessary class size reductions, as well as the external validity properties of the Burt test score *gains*. We would then compare the cost of this form of social policy to other possible social policies of the same dollar amount. The estimates here serve as a key ingredient in making that informed policy analysis. In addition, they provide some sense that class size reductions are efficacious, at least in terms of differences in the test score outcomes by group. They also illuminate the diminishing efficacy of class size reductions and that they are no 'magic bullet' - from what we can see, more persistent class size reductions yield greater benefits. But in contrast to much of this literature, they do imply that class sizes are a useful policy instrument for altering academic outcomes.

5.2 Top to Bottom Income Decile Test Score Gaps

The levels of Burt scores are highly correlated with family income. Indeed, one of the apparent strengths of Burt scores, in their level form, is that they serve as a decent summary statistic of family background information. For this reason, test score gaps by income are one of the most striking, compared to using any other demographic measure alone. Comparing the average test scores for the portion of the sample with the lowest decile in family income to the highest decile in family income, we obtain average Age 8 scores of 39.9 and 53.4, respectively. By Age 13, the respective averages are 79.2 and 92.9. Thus the initial test score gap of 13.5 has remained essentially unchanged at 13.7. Average class sizes are much the same for the two groups.

If, as for the Maori / Non-Maori comparisons above, we use the Large to Medium class size effect from Table 4, column 2 again, a 3.5 student reduction in average class sizes from Age 8 to Age 13 would lead to about a 2.2 gain in scores for the very poor students relative to the very rich. This would close the initially large gap of 13.5 points (or about a full standard deviation of the Burt score levels) by about 16 percent. While that may not seem large, it is important to recognize that is one of the largest differences in average scores between two groups seen in our sample. Again, since the type of class size reduction policies we are utilizing here are not cheap (an average of 3.5 students for 6 years of class sizes!), it is important to assess what the Burt scores tell us what the benefits should be.

Apart from giving us a sense of the magnitude of the class size effects, this discussion also allows us to speak of the *relative* desirability of class size reduc-

tion policies. First, reducing class sizes where class sizes are initially at their largest appears to yield greatest benefits for a given reduction in class size. Secondly, if the comparisons between columns 1 and 2 of Table 4 are to be believed, policies that lead to children *always* being placed in smaller classes (at each age) rather than just on average appear to yield larger effects. But of course, due to limited resources for schooling, such *persistent* class size reduction policies may simply not be possible. Third, if the quantile estimates are to be believed as they stand (which, we acknowledge, there are good reasons to doubt this), then children who are at the upper quantiles of the *gain* distribution, and so at the lower quantiles of the *levels* distribution appear to be the most responsive to class size reduction policies. While this speaks to the *conditional* quantiles, in terms of the unconditional quantiles, this indicates the poorest children benefit the most from class size reductions. A very important caveat to this last conclusion, however, is that this may be more a function of the outcome we study of Burt scores and their topping out, than of social policy outcomes more generally. Poor children may respond more simply because, as far as Burt scores are concerned, they have more 'room' to respond.

6 The Distinction Between Class Size Effects in Public and Private Schools

In earlier reports produced by LECG using the CHDS data, they noted the importance of an indicator for the child predominantly attending a private school on the growth of the child's Burt scores. We put this in a separate section in our report here, because of the question of the interpretation of what this statistical finding means. Indeed, for our 'long difference' approach in this report, we find that private school students have roughly 2 point larger growth over the ages from 8 to 13 (even though they have higher age 8 scores to begin with), and this effect is significant at about a 15 percent level of significance. However, what is most interesting about the private school effect is not its role as just an additional dummy variable regressor we can hold constant (and then just how to interpret that *vis a vis* policy questions is not entirely clear, since it is clearly a choice which is constrained by family income, etc.), but in the overall sample we use above that so many of the classes designated as small or (to a lesser degree) medium are in private schools. The flipside of this is that virtually none of the average large class size observations we have come from private schools, whereas almost 40 percent of the small average class observations are for students in private schools (roughly 13 percent of the CHDS students we classify as attending a private school, which we code as a binary variable - 0 or 1.) [16]

[16] It turns out that over the ages from age 8 to 13, the students tend to be either mostly 1's or mostly 0's, so whether we use say 0.5 or 0.7 as the cut-point yields the same classifications,

This basic feature of the intertwining of our class size designations with our private school categorization is displayed well in Figure 9. This shows the densities for the average class size broken out by our Public school and Private school designations. The two vertical bars mark the cutpoints that break our sample into Average Small Class, Average Medium Class, and Average Large Class. There are 495 observations in the Public School density, and 74 in the Private School density. Thus, even though visually it is clear that as a percentage, there are many more observations in the Small Average Class category for the Private school students, as an actual number, there are fewer students in the Private school left tail of the density than in the Public school left tail. But that caveat aside, Figure 9 makes clear the tendency of Private school class sizes to be smaller, and thus be somewhat collinear with our class size categories. In addition, Figure 9 shows that the modal average class size (about 31.5) for Public school students is well to the right of most of the distribution for the class sizes for the Private school students.

This implies that we may have difficulty in trying to detect a Private school effect separate from a Small/Medium class size effects. Furthermore, since that regression somewhat mixes apples and kiwifruit, as one measure is a resource measure and the other a choice on behalf of parents, we decided instead to avoid the whole collinearity/interpretation issue by just running separate regressions for the Public/Private school sectors in order to understand the resource effects *within* each of these sectors. The collinearity issue is further displayed in Table 7, which is just a cross-tabulation of the observation frequencies and associated percentages conditional on not being in a large class (as this is our base, or reference, group in the regressions). The cross-tab makes clear that only a total of 15 percent of the observations lie in the off-diagonal cells, and so allow for separate identification of the Private school effect from the Small/Medium class size effects in a pooled regression.[17]

Analogous to Table 4, we then consider only mean regressions for the Public and Private school effects, and these are given as the two columns in Table 8. The results display a substantial heterogeneity in class size effects for the two sectors. Relative to our results for the pooled sample in the second column of Table 4, the class size effects for just the Public school students are now somewhat larger. Whereas the Small Class effect in Table 4 (relative to the base group of Large Average class size) is 3.5, now this effect is 5.5 with a standard error of 2.3 and a significance level of about 2 percent. The Medium class size effect is largely unchanged, rising from 2.2 to 2.4 and the significance level of still about 10 percent. The sample size was 569 for the overall (pooled) sample, but 495 for the Public school students. Converting these into the *per student* effect sizes, as defined above, the Large to Small class reduction yields an effect size of 0.04σ, and the Large to Medium reduction in class size still

for this reason.

[17]Strictly speaking, this isn't quite true as there are other covariates in the model, and so the *conditional* variation may imply that this number is not really 15 percent.

rounds out to just over 0.06σ. Thus, there still appear to be declining marginal benefits to class size reductions, but this is not as pronounced as compared to the results based on the pooled sample of the Public and Private sectors. Thus, the evidence on the declining marginal benefits of class size reductions may be largely (although perhaps not entirely) an artifact of pooling the data across sectors.

For the 74 students in the Private school sector, however, there appear to be no detectable class size effects. If anything, the point estimate on the Small class effect suggests a *negative* association with the growth in test scores, but the magnitude of the point estimate is well within a standard error of 0. The point estimate for the Medium class size effect is positive, but highly insignificant. Thus, Table 8 indicates that the apparent declining marginal benefit of lowering class sizes from Large to Medium to Small is evidently due to going from Public to Private sectors. Looking purely *within* the Public sector, the returns to lowering class size appear to be much more uniform, regardless of where the class size reductions are implemented. Table 8 also indicates that class size reductions in the Private sector appear rather ineffective, at least insofar as Burt Reading test scores are concerned. This may be because, to some extent (as displayed in Figure 9), class sizes are already somewhat smaller in the Private school sector. While these two sets of results are perfectly valid *conditional* on the Public/Private sector, to extract more meaning would require modeling the Public/Private school decision more, and it is not clear to us what extra benefit (or further questions answered) this would yield. The results for the Public sector seem particularly useful, since it is likely a majority of those children are there by constraint as opposed to by choice.

7 Gender Differences in Reading Achievement and Class Size Assignments

Finally, we would be remiss if we did not comment on and further analyze the differences in achievement by gender in our sample. It has likely not escaped the reader's attention that the control for the student's gender (Female) in our regressions is almost always significant and when it is, it is always negative. Of course, part of the explanation for this is simply that females do better on the Age 8 Burt Word Reading test, and due to the inverse correlation of the subsequent gains with the levels, this lower growth in scores for girls might be expected. But what we did not expect was that girls are also more likely to be placed in larger classes, particularly at younger ages, and that it may be *because* of this feature of class size assignments that (at least in part) girls have lower growth in scores than do boys from ages 8 to 13. The net result of this is that by age 13, the densities of test score performance for girls and boys are nearly identical.

What makes this story a little complicated, however, is that it is wrapped up with the Public/Private school differences. Looking purely within the Public school sector, there is still a gender gap in achievement as well as a gender gap in class size assignments, but it is substantially muted relative to the overall gender gap in class sizes. Conditional on being in the Private school sector, girls are in substantially larger class sizes. Why this is, is not entirely clear to us, and we invite feedback on whether this is perhaps secular to the Christchurch area private schools, or to this cohort. But it seems that either boys and girls are being sent to different private schools with perhaps differing emphases, or, if they are sent to the same private schools, girls are placed in different tracks than boys. We found this unusual, because we started this project figuring we would see differences of this nature along demographic lines such as ethnicity or family income, but the demographic differences that are most systematic are those by gender.

At age 8, female students on average score 4.7 points better on the Burt tests than their male counterparts, and this has a p-value of less than 0.1 percent. By age 13, this mean difference has shrunk to 2.5 and is statistically significant at only a 6 percent level or greater. Figures 10 and 11 show the Age 8 and Age 13 densities respectively, and the reader can see that at age 8 the female density is nearly a uniform rightward shift for much of the range of the scores (except near the top). At age 13, the two densities are nearly identical except for a bit of mass just to the left and right of the median. Again, this would mean little owing to the mechanical negative correlation between test score levels and gains were it not for the patterns of class sizes by gender displayed in Figure 12.

Figure 12 shows that the density of class sizes for boys stochastically dominates the class size density for girls. In terms of just the means of these densities, girls are in class sizes about 0.65 students larger on average than boys, and this is statistically significant at a 5 percent level of significance. In fact, a Wald-like estimator using the gender differences in class size (0.65) and the lower test score growth (-2.2) to compute a per-student effect size of 0.03σ which is on the order of the lower effect sizes computed above. While this is not, in and of itself evidence of causality (i.e. that it is the larger class sizes *causing* girls to achieve less from ages 8 to 13 than boys), it does raise the disturbing question of *why* girls tend to be placed in slightly larger classes?

Part of the answer lies in how girls and boys are assigned to classes of different sizes in the Public and Private schools. Figure 13 repeats the plots of Figure 12, but now just for the students in Public schools. While the mode of the density for boys still lies to the left of the mode of the density for girls, and to the eye Figure 13 may not appear much different than Figure 12, the meshing of the left tails especially implies much of the difference is lost. The difference in the means has been cut by about a third, to 0.44, and is now significant at only a 17 percent level of significance (unlike the 5 percent level for Figure 12). Figure 14 then does the same exercise for the Private schools, and here the gender differences in class sizes are visually apparent. The difference in the means is

now at 1.34, but owning to the small sample size ($N = 62$) the p-value is still only about 20 percent.

Unfortunately, we leave this analysis for now, perhaps creating more questions than answering them. But it does seem clear that whatever gender differences do exist in the Public school sector in Christchurch, they are much smaller (by about a factor of 3) than those found in the Private schools, insofar as class size resources are concerned. Part of what we think is going on here is a tendency of Public schools to be somewhat compensatory in their assignments of students to different size classes. We have found in our data a clear pattern of students who do well on the Burt tests subsequently being placed in larger classes (a similar pattern has been documented in U.S. schools by Boozer and Rouse (1999)). As a result of initially performing better on Burt scores, females then get placed in larger classes as a result of their higher initial aptitude. But while this seems to explain the behavior of the Public schools, something very different appears to be going on in Private schools, where aptitude differences cannot explain the differential treatment of boys and girls. We are currently seeking more institutional background to fully understand the differential treatment of boys and girls in this sample in the Private school sector (i.e. did they attend different Private schools, or the same ones? Was the emphasis of the Private schools perhaps different for those attended by boys as opposed to those attended by girls?)

8 Do The Test Score Conclusions Hold for the Age 21 Outcomes?

This section has been added after the analysis of the initial report. It takes advantage of the age 21 data for the CHDS cohort, collected in 1998, and as such, allows us to look at outcomes apart from just test scores which are also of more direct policy interest. These include analyses of completed educational levels to date, unemployment experiences, and criminal justice violations. However, even these data will have their own set of limitations *vis a vis* the test score outcomes. A prime example of this is illustrated by the weekly earnings data - they have the *opposite* relation one would have expected based on the test score analysis. However, it is well known (see eg. Mincer (1974) for U.S. data) that at age 21, those individuals with higher education levels tend to have *lower* earnings, in contrast to later ages, because to a certain extent, their educational experiences are not yet completed. Thus, analyzing the correlation between class sizes and weekly earnings, say, will not provide a good indication of the relation between class sizes and *life time earnings*, for this reason.

The chief motivation to examine these age 21 outcomes is that, unlike test scores which may or may not be relevant, these outcomes are of direct (or ultimate) policy interest. Test score outcomes serve as a means of providing

faster gratification in terms of ascertaining the value of various educational and child health interventions. But as we discussed at length above, test scores have many limitations, and so one of many such concerns might be that due to the 'topping-out' phenomenon in the Burt scores, our estimated class size effects in the earlier part of this report are driven by this 'mechanical' aspect of the test scores, and perhaps not indicative of any real phenomenon. Thus, the age 21 outcomes will also serve to 'validate' our test score analysis. In this sense, the age 21 analysis will likely not be of as much interest to policy makers, but will be quite valuable to educational researchers who are constrained to rely on test score analyses. Here again the unique sampling design of the CHDS becomes apparent, in that it is rather infrequent that researchers have access to early childhood cognitive measures combined with early adulthood social and labor market outcomes. Thus our chief conclusion from the academic research perspective is, the choice of econometric specification for analyzing the Burt scores which was based on experimental analyses from the U.S. is corroborated by finding similar effects for the age 21 outcomes, thus lending somewhat more credence to our choice of 'long difference' in the test score growth over the childhood years.

As we discussed in Section 3, we analyze the Early Adult outcomes in a fashion analogous to how we studied the Burt score outcomes. The difficulty here is that we do not observe in the CHDS the class size inputs after age 13. However, since children often are in their final secondary school as of age 13, there is good reason to suspect the age 13 class size measure is a decent proxy for the 'permanent component' of the class size inputs faced by the individual from ages 13 to 18. For that reason, we use the age 13 class size measure as our proxy measure for the relevant inputs. And analogous with the Burt score analysis, we again hold constant the age 8 Burt score to net out the influence of omitted factors such as family inputs, etc. which may influence the age 21 outcomes, but which are not due to any policy inputs. It is worth noting that the Burt scores at age 8 have a strong relationship with outcomes such as education completed as of age 21 - this speaks directly to the external validity of the Burt scores in their level form for outcomes of direct policy interest.

To keep with our goals in the above discussion, here again in analyzing the Age 21 outcomes we may, at times, forgo utilizing an econometric technique which some might advocate as more 'correct' in favor of methods which are more transparent and easier to interpret. The reader should not worry too much at this choice, however, as we have made sure the results we present here are not sensitive to alternative methodologies.

To start this discussion, one of the first outcomes we analyzed was education completed as of age 21. Of course, some fraction of the sample will not have *completed* their education as of age 21, although we suspect many of these individuals will be in the midst of their schooling because they may seek higher degrees, etc. Nonetheless, due to how our measure of schooling is coded, this is likely not a problem. In particular, due to potential measurement problems,

we used the CHDS re-code of the underlying education responses called TERTIARY. This is an ordered variate, taking on the values 1 (no post secondary schooling/training), 2 (Basic Skills training), 3 (Intermediate Skills training), 4 (Enrolled for Bachelor Degree). We actually use OLS (as opposed to an ordered probit or logit, for example) to analyze this variable, arguing that the jump to go from 1 to 2, for example, is the same hurdle as the jump from 3 to 4. Some *ad hoc* tests we do not report on here indicate this may not be such a bad approximation, and the interpretability of the OLS results as opposed to ordered logit results (which ignore the scale of the dependent variable ranks) led us to choose to report the OLS results in Table 9.

The t-statistic on the class size variable is marginally significant with an absolute value of about 1.7. The coefficient estimate, however, seems rather small. To give it some interpretation, consider reducing class size by 10 students (almost the margin needed to change from the Large to Small class categorizations we used above): this would yield an effect on the dependent variable of 0.15, or a little less than 1/6 of the distance needed to move from one category to the next. Of course, we should not expect that a class size intervention at age 13 is a 'magic bullet' in terms of pushing students up one full category in terms of our dependent variable. It is also possible, with full data on the class size inputs between the ages of 13 and 18, that 'permanent' class size reduction policies might again be associated with stronger effects on this educational attainment outcome. To try to convert this into dollars, if we assume 1 extra year of education yields a 10 percent rise in lifetime income, and if we make the (perhaps very) rough approximation that the dependent variable codes align with years of schooling, then the 10 student reduction in class size at age 13 would roughly increase lifetime earnings for those students by about 1.6 percent.

As we mentioned briefly above, we cannot examine the earnings of these individuals directly, simply because age 21 is too young. At this age, it is well documented by labor economists that schooling and earnings are *negatively* correlated (as indeed, we find in the CHDS) and age 21 earnings are not well correlated with lifetime earnings. Thus examination of the age 21 earnings directly would produce misleading results, and so we leave that exercise until the CHDS cohort ages sufficiently to revisit that question. We can, however, meaningfully examine outcomes such as unemployment experiences for those individuals seeking work. These results are presented in Table 10, and again, we have opted for econometrically simple methods to keep the focus on interpretation. In the two columns of Table 10 we have split the unemployment variable (time unemployed since age 18) into the binomial incidence variable (the linear probability results which are reported in Column 1), and conditional on positive incidence, the duration of the *total* time unemployed in Column 2.

The coefficient on the class size variable in Column 1 is positive with a t-statistic of 1.5. As we saw in Table 9, the statistical precision for this early-adult outcome is again not overwhelming by conventional standards (both are signifi-

cant at a 10 percent level of significance).[18] However, the coefficient is positive and indicates that a 10 student reduction in class size at age 13 would be associated with a 6 percent rise in the probability of experiencing an unemployment spell at some time by age 21. By comparing the number of observations in Column 2 and Column 1, we can see that just over 41 percent of the CHDS sample experiences *some* unemployment incidence between the ages of 18 and 21, as we have defined it. Thus, the 6 percent reduction in the unemployment incidence associated with the 10 student lowering of the class size at age 13, implies a reduction in the unemployment "rate" for this sample of about 15 percent. This is a sizable reduction, but again the class size intervention we are considering (a 10 student drop in class size) is significant.

Conditional on experiencing *any* unemployment incidence, in Column 2 we examine the total duration of all spells between the ages of 18 and 21, for the 41 percent of the sample with non-zero spells.[19] For this conditional sample, the mean duration time is 9 months (the dependent variable here is measured in months) and the mean duration for the unconditional sample is 3.6 months.[20] Here we see the effect of class size on duration is quite significant (a t-statistic of about 2.8) and sizeable. A 10 student reduction in class size would be associated with about a 3.3 month reduction in time unemployed between the ages of 18 and 21, which is more than a 35 percent reduction in the mean unemployment duration for this conditional sample.

Thus, on the whole, class size reductions, even holding constant family background measures as well as Burt test performance (which, it is worth noting, is *strongly* negatively associated with the duration for the conditional sample, and so again speaks to the Burt test's external validity in levels) are associated moderately with reducing the likelihood of experiencing an unemployment spell at all between the ages of 18 and 21. However, conditional on experiencing a spell, class size reductions are strongly associated with shorter durations. This again corroborates some of our Burt score analysis discussed in earlier sections that the class size reductions appear to be most efficacious for those individuals worst off. To that end, we also examined the relationship of class size with arrest incidence, but perhaps owing to the low incidence of arrests in our sample (about 9 percent) we did not have enough power to detect any systematic effects. Lacking a clear pattern of results, we do not tabulate these results here, but suffice it to say, we were unable to detect an association between class size at age 13 and arrest incidence given our econometric methods and standard

[18] Since this is a linear probability model, the standard errors in Column 1 have been corrected for heteroskedasticity.

[19] If we simply pool the zero spell and non-zero spell durations together and estimate the model in either column of Table 10 via OLS, the coefficient on class size is now highly significant (a p-value of less than 1 percent). That regression, however, mixes apples and oranges and so we present the regressions 'split out' as described in the text and Table 10.

[20] Specifically, we use variable x134 from the CHDS data as our measure of unemployment duration. The results are not much changed if we instead use the duration of unemployment benefit *receipt* (variable x136) as our measure instead.

model. In terms of the "economic" outcomes, this analysis is rather exhaustive of the outcomes measured as of the age 21 followup to the CHDS.

9 Conclusions

A conundrum has existed in the academic literature in the U.S. for quite some time over the observational and experimental evidence on the effects of class size reduction policies on student outcomes. Unlike many academic debates, this one not only has some conceptual nuances embedded in it, but more importantly, it has profound implications for how we should structure social policies for children to further their intellectual development. The majority of published studies on class size effects time and again tend to reveal no, or possibly even perverse, impacts of reducing class sizes on subsequent academic growth.

The CHDS data are radically different in one principal respect as compared to almost all other data on this topic, and that is the long individual time span covered by the early years of the sample: six years of test taking and school resource information collected so as to be comparable across time is unheard of. The closest data approaching that in the U.S. that is known to be available is the National Education Longitudinal Survey (NELS) data, which has similar coverage at three points in time for grades 8, 10, and 12. But the CHDS has the additional advantage of initializing the sample early in the child's schooling history, and due to the cumulative nature of schooling, is less susceptible to missing data problems arising for that reason as compared to the NELS, say.

Typically, because of the nature of the U.S. data, studies have had to look for effects of policies in the latter portion of the child's public schooling experience, and even then look at only a year or two after the dating of the policy variable, such as class size (to use the best case scenario of the NELS, a researcher could look at subsequent growth compared to the initial test score in the 8th grade of a class size reduction in the 8th grade, and then try to detect effects on tests taken in 10th and 12th grade.) Because of both the short time window of the data as well as the sampling occurring in the latter part of the schooling experience, researchers have made extensive use of the 'value-added' model to account for these shortcomings of the data. Given its intuitive appeal, as well as lacking any good alternative method (and frankly, no good reason to look for other methods), this statistical method has been taken as an article of faith and a given.

While expressing doubt over econometric assumptions is pervasive, evidence in the past five years has shown this concern to be far more than an academic quibble. In particular, some widely cited evidence from a true experimental study done in Tennessee in the U.S. revealed that a value-added model, unless applied to a time window that exactly covers when a child first enrolls in a small class, will *entirely* miss the effects of the class size reduction. That indicates the thirty year history of research on this question is fundamentally flawed in

the statistical tool that is used to detect *any* effects of class sizes. For technical reasons, because of its long time window of sampling each child, the CHDS was ideally suited to testing the assumptions of the value-added model, and this was done earlier in work by LECG.

But this is only half the picture of the Tennessee results, and this report draws further on the full picture that emerges from that experiment, to design the correct statistical methodology. Eric Hanushek in a series of papers critiquing the *interpretation* of the Project STAR results also notes that a child who enrolls in a small class only to later enroll in a regular sized class sees the one-off gains he initially enjoyed erode or fade away. In other words, to *keep* the initial boost in test scores obtained by enrolling in a small class, as measured by test scores, the child had to remain enrolled in a small class. Hanushek interpreted this finding to imply short term class size reduction policies have only temporary effects on the academic outcomes of the students, before reverting back to the average. We, however, take no ideological stand on the issue, but instead a methodological one: not only should enhanced schooling resources in the guise of smaller classes affect more the intercept than the slope of an indicator of academic success as the child ages, but also we may expect to find such effects only if those reductions are somewhat persistent.

This indicates a radically different statistical framework than the simple value-added model.[21] To add to this complexity, when then have to ask: "The effects of class size on *what* outcome?" In our case, with the CHDS data for early childhood outcomes comparable over the years, the answer was clearly the scores on the Burt Word Reading tests. These are not a bad measure to use in that they are documented to have good discriminatory properties *across* children.

But Burt test scores are also rock hard. In this report we showed that removing variation due to secular increases in scores as the children age, as well as variation which is fixed relative to each child, and so not likely due to any policy inputs, but factors more intrinsic to the child, eliminates all but 5 percent of the variance. And it is only in that 5 percent do we as researchers allow policy effects, evolution of family background, random factors which lead to higher or lower scores on the test day, etc. It would seem intuitively apparent that something so 'rock hard' is not going to be changeable in any systematic way in a period of one or two years. It was really this aspect of our outcome measure that invalidated one of our initial plans for this project that was to look for evidence of the short-term 'intercept' effects noted by Alan Krueger's work on the Project STAR data. Furthermore, the experiment provided clear evidence on the timing of when students were assigned to small classes, and when the exams were given to monitor their subsequent performance. The CHDS, by simply sampling what resources the student is exposed to, is far murkier on the

[21] Indeed, with due respect to Hanushek, it is interesting that the value-added model is embraced so warmly, that it is clear in his papers he *defines* the value-added coefficients to be the objects of policy interest.

appropriate choice of *when* in the child's subsequent tests researchers should expect to see those effects. And if the effects are in intercepts as opposed to slopes, then timing, as we say, is everything. The realities of the data make this conceptual exercise simply not feasible.

But the methodology inspired by Hanushek's critique of the experimental data meshes extremely well with the nature of our outcome variable, the Burt test scores. For a persistent outcome, it is likely, if at all, that only persistent policies will have an effect, and even then over a suitable length of time. This is where using the unique aspect of the CHDS data is of key importance: we can set the time window at the maximum allowed by the CHDS data - six years - and then look to see if persistent policies have effects. If the test scores in the U.S. data were as persistent as the Burt scores, there is simply no possible way a researcher could detect effects with a time window of 2 or 3 periods.[22]

While our finding of significant class size effects is, in both a statistical and interpretative sense, rather novel, and not common in the bulk of the literature on this topic, it also indicates class size policies are not a magic bullet. As far as we can tell, indeed only persistent class size policies have effects we can detect. The caveat to this criticism offered by Hanushek and others is that it depends critically on the outcome being studied. Unfortunately, little to no analysis of the type given above applied to the Burt scores has been done on the tests used to measure the Tennessee effects. Perhaps there too only persistent policies yield effects because those tests exhibit similar persistence properties. This has been a topic that is sorely missing from much of this literature, namely that, in order to speak sensibly of the nature of the effects (or lack thereof) of a policy on an outcome, we need to know the properties of that outcome.

Do our results here, or the Project STAR results, imply that short-term class size reductions will have *no* persistent effects on *any* outcome? Of course not - but lacking those other outcome measures, we must maintain the conclusion offered here and its accompanying critique given by Hanushek. But this observation indicates another challenge to researchers who want to look at policies affecting early childhood development. Many such measures are quite good at discriminating *between* children, but we need to discover more measures that correlate well with the subsequent *evolution* of children. It is quite likely that a measure that suits one purpose well will serve the other one poorly.

One way out of this box is to wait: as the CHDS cohort ages, we can return to this data and relate both short and long term class size reduction policies to outcomes of direct policy interest, such as school completion, employment prospects, family formation, etc., as well as study the intermediate linkages of how the evolution of Burt scores helps predict this array of outcomes later in life. Once again, the extraordinarily long time window allowed by the CHDS will be highly important in refining our knowledge of the relative efficacy of short run

[22]The tests in the NELS, by virtue of attempting to measure higher order skills than just word recognition, have nowhere near the reliabilites of the Burt scores, and so are (at least potentially) much more mutable.

versus long run policy revisions. In this paper, we were able to report on a few of the outcomes of the CHDS as of their age 21. These include: completed education to date, unemployment incidence and duration from ages 18 to 21, and arrest incidence as well as wage outcomes, conditional on employment. The wage analysis highlights the need to apply some care in the analysis, as wages are *negatively* correlated with education as of age 21, owing to this being an age of cross-over from education and training into the labor market. But for the education and unemployment measures, we found class size effects completely in accord with our analysis of the Burt scores: lower class sizes were associated with more completed education as of age 21, lower incidence of unemployment spells, and conditional on experiencing an unemployment spell, substantially shorter durations. These findings are firstly of a pure research interest, since they seem to corroborate some of the methodological innovations we introduced in this report with regards to how observational data on student outcomes and policy inputs such as class size are analyzed. They are also of public policy interest, since it indicates that class size reductions are not just efficacious in raising academic outcomes, but also outcomes pertaining directly to the utility and well-being of the individuals themselves.

Our analysis has tried to build upon the highly unique aspect of the CHDS data, and that is its extraordinarily long time span over which it follows these children, who are now young adults. In so doing we have been able to re-examine some of the assumptions conventionally made in analyzing the effects of class size policies on both test score and labor market outcomes. We concluded that many of the assumptions made by the conventional value-added model were not supported by recent results from experimental studies of class size effects conducted in the U.S. We instead developed methods which took advantage of this unique aspect of the CHDS data, the long time span of observation, which allowed us to dispense with many of the *ad hoc* assumptions used in those earlier studies relying on the value-added model. When viewed in a way consistent with the experimental studies findings, we found effects remarkably consistent with those from the experimental studies, but here such effects were extracted from the observational data of the CHDS. We found that *persistent* class size reduction policies were associated with significant increases in Burt Word Reading performance from age 8 to 13. This is perhaps unexpected, if only because the Burt score is itself something which is highly stable and rather immutable. Furthermore, using the newly released age 21 data, we corroborated these findings for the Burt test scores by directly examining the effects on early adult outcomes using a similar methodology.

This paper, put together with the U.S. experimental evidence, should help draw the observational (or correlational) and experimental evidence closer together. A natural next step, both for the research community as well as the public policy community, would be to conduct further experiments which build upon and enhance our knowledge base. A highly useful direction would be to expand the time window of observation and even the experimentation period rel-

ative to the experiment conducted in Tennessee in the U.S. (the program ended after third grade). We hope that this report, together with the experimental literature and its subsequent critiques and corroborations, provides a basis for more careful experiments and analysis of observational data in the future.

10 Appendix - The Distinction Between Test Reliability and Stability

The paper by Heise (1969) discussed at the beginning of this report points out that a test instrument's reliability and stability properties need not be identical. Furthermore, under assumptions of temporal stationarity and the absence of serially correlated testing errors, he develops a framework to empirically identify distinct reliability and stability measures if a minimum of 3 re-tests are available to the researcher. However, his paper is written in the language of path analysis, and so perhaps for this reason, is largely unknown to econometricians. The point of this appendix is to couch his argument in the notation of measurement error models familiar to econometricians.

As mentioned above, all of these derivations assume temporal stability as well as the absence of serial correlation in the measurement errors. To begin with, assume we have only a test and re-test on individuals available to us:

$$x_{i1} = x_{i1}^* + e_{i1} \tag{11}$$

$$x_{i2} = x_{i2}^* + e_{i2} \tag{12}$$

and we make the conventional assumption that the measurement errors e_{i1} and e_{i2} are uncorrelated with any of the true values. Assuming stationarity, the test instrument's reliability may be measured by either:

$$\frac{Var(x_{i1}^*)}{Var(x_{i1})} \tag{13}$$

or

$$\frac{Var(x_{i2}^*)}{Var(x_{i2})} \tag{14}$$

Since neither x_{i1}^* nor x_{i2}^* is observed given the data, the conventional assumption is to assume

$$Cov(x_{i1}, x_{i2}) = Cov(x_{i1}^*, x_{i2}^*) = Var(x_{i1}^*) = Var(x_{i2}^*) \tag{15}$$

However, the possible failure of this equality is precisely the point of the Heise paper. If the test instrument is not perfectly stable, then this covariance of the 2 observed outcomes need not identify the (temporally stable) variance in the 'true' test score measure. Lacking any further data or information on

this structure, the empirical identification of the reliability measure remains unsolvable. Of course the conventional path taken by econometricians is to assume perfect stability (thus leading to the last two equalities in the previous equation), in which case the empirical counterpart to the covariance given in the previous equation allows for identification of the variance in the true test score, and so estimation of the reliability measure by either of the 2 formulae above. If we let λ denote the reliability ratio, then under conventional assumptions we have:

$$Cov(x_{i1}, x_{i2}) = \lambda Var(x_{i1}) = \lambda Var(x_{i2}) \tag{16}$$

or, dispensing with the (possibly empirically invalid) last equation, and simply writing everything in terms of correlation coefficients rather than covariances, we have:

$$Corr(x_{i1}, x_{i2}) = \lambda \tag{17}$$

If, however, a third test is available, taken 1 period after the period 2 test and denoted as x_{i3}, then the stability and reliability measures can be separately identified. Let the stability measure between periods 1 and 2 be given by s_{12}, then in terms of the algebra above:

$$Cov(x_{i1}, x_{i2}) = Cov(x^*_{i1}, x^*_{i2}) = s_{12} Var(x^*_{i1}) = s_{12} \lambda Var(x_{i1}) \tag{18}$$

Again, rewriting the subsequent two equations in terms of correlation coefficients as above, we have that:

$$Corr(x_{i2}, x_{i3}) = s_{23} \lambda \tag{19}$$

and

$$Corr(x_{i1}, x_{i3}) = s_{13} \lambda \tag{20}$$

By squaring this last equation, and noting that stationarity delivers the equality $s_{13} = s_{12} s_{23}$ (i.e. the stability between the first and third test is simply the products of the stabilities between the first and second test and the second and third test) and abbreviating the correlations by c_{12}, etc. we have the relation:

$$\lambda = \frac{c_{12} c_{23}}{c_{13}} \tag{21}$$

Since the right hand side of this equation has an empirical counterpart, this equation serves as a basis for the estimation of the test reliability, even when the simple test/re-test correlations do not suffice, due to less than perfect stability. Similarly, we can estimate the 3 (of which only 2 are uniquely determined) stability coefficients by:

$$s_{12} = \frac{c_{13}}{c_{23}} \tag{22}$$

$$s_{23} = \frac{c_{13}}{c_{12}} \tag{23}$$

and

$$s_{13} = \frac{c_{13}^2}{c_{12}c_{23}} = s_{12}s_{23} \tag{24}$$

For our purposes, the fundamental point of the Heise paper is that test stability and test reliability are conceptually distinct concepts. Indeed, they can be empirically distinct concepts as well, with enough repeated observations. But the important implication for our analysis, as well as for the study of testing outcomes in observational studies, is that a test instrument can have high reliability properties even though it may have low stability properties. This is important, since it is common for economists and other social scientists to rely upon fixed-effects or similar strategies when dealing with observational data to control for certain types of endogeneity concerning the policy variables of interest. The problem is that if test instruments are constructed so as to have good reliability properties, *and if* those reliability properties are computed via what is really a stability measure as Heise discussed, then very little variation will be left in the dependent variable when these high reliability / high stability test scores are used as the analysis variable of interest. Thus, we should not be too surprised if, as researchers interested in the influence of public policy variables, we find that after differencing such test score measures over time we typically find little correlation with the policy variable of interest. The simple reason may be that little signal is left in the data as compared to measurement error after the data are differenced.

Of course it is true generically in panel data methods that differencing error-ridden measures when the measurement error is uncorrelated will exacerbate the measurement error problem. But what makes this problem particularly acute in the case of test scores is that test instruments may well be *designed* to have high stability properties, if the stability properties are used to proxy for test reliability. The point of the discussion here is to emphasize that social scientists relying on fixed effects (and analogous methods) should investigate not just the reliability properties of the test instruments, but also the stability properties whenever possible so as to mitigate the problems that can arise when a highly stable test instrument is being analyzed.

References

[1] Boozer, Michael A. and Cecilia Rouse, (2001), 'Intraschool Variation in Class Size: Patterns and Implications,' *Journal of Urban Economics*, forthcoming.

[2] Bryk, Anthony S. and Stephen W. Raudenbush, (1992) *Hierarchical Linear Models*, Sage Publications: California.

[3] Gilmore, Alison, and Cedric Croft and Neil Reid, (1981), 'Burt Word Reading Test, New Zealand Revision: Teachers Manual,' New Zealand Council for Educational Research, Wellington.

[4] Griliches, Zvi and Jerry Hausman, (1986), 'Errors in Variables in Panel Data,' *Journal of Econometrics*, 93, 93-118.

[5] Hanushek, Eric A., (1999), 'The Evidence on Class Size,' in *Earning and Learning: How Schools Matter*, ed. by Susan E. Mayer and Paul Peterson, Brookings: Washington, D.C.

[6] Hanushek, Eric A., (1999), '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(2), 143-168.

[7] Heise, David R., (1969), 'Separating Reliability and Stability in Test-Retest Correlation,' *American Sociological Review*, 34(1), 93-101.

[8] Krueger, Alan, (1999), 'Experimental Estimates of Education Production Functions,' *Quarterly Journal of Economics*, 114(2), 497-532.

[9] Mincer, Jacob, (1974) *Schooling, Experience, and Earnings*, NBER: New York.

Table 1

Decomposition of the Variance For the Burt Word Reading Tests

---

Within and Between Person Decomposition

Total Sum of Squares	626.99	N=873, Total Obs. 4347, Avg. Obs. Per Person 5
Between Sum of Sqs.	378.64	(60%)
Within Sum of Sqs.	248.34	(40%)

Decomposition of Burt Scores Allowing for Secular Age Effects

(Net of Age Effects)

Between Sum of Sqs.	342.14	(55%)
Within Sum of Sqs.	34.65	(5.5%)

Age effects account for 86% of the within variance, 10% of the between variance and 34.5% of the overall variance of Burt scores.

Fraction of the Variance in Burt Scores Explained by Person and Age Effects: 94.5%

Coefficients and Standard Errors on the Age Effects From the Fixed Effects Regression

Constant	45.2
	(0.23)
Age 9	9.0
	(0.31)
Age 10	19.1
	(0.32)
Age 11	27.2
	(0.32)
Age 12	33.7
	(0.32)
Age 13	39.7
	(0.32)

Table 2
Regressions for the Change in Burt Scores (Age 13 - Age 8) By Permanent and Average Class Size Categories

Permanent:		Average:	
Small Class [N=17]	3.22 (2.61)	Small Class [N=48]	1.31 (1.62)
Medium Class [N=32]	-1.25 (1.95)	Medium Class [N=447]	[Omitted Category]
Large Class [N=21]	-3.24 (2.37)	Large Class [N=74]	-2.19 (1.33)
Female	-1.84 (0.89)	Female	-1.86 (0.89)
Mother Maori	0.86 (2.41)	Mother Maori	0.90 (2.41)
Father Maori	0.95 (1.70)	Father Maori	0.83 (1.68)
Change in Family Income	0.002 (0.21)	Change in Family Income	0.014 (0.21)
Constant	40.79 (0.70)	Constant	40.88 (0.73)
R-squared	0.02	R-squared	0.02
Number of Observations	569	Number of Observations	569

Note: Cell Sizes for the Class Size Categories Given in Brackets. Reference Category for the regression in the left column is the group of student who are not always in a class size of a given category for the duration of the sample period, for which N=499. The results are unchanged if we restrict the Medium Class Size coefficient to 0, and so include these as additional members of the reference group, analogous to the regression on the right.

Table 3
Quantile Regressions for the Change in Burt Scores (Age13 - Age8)
By Permanent and Average Class Size Categories
(Standard Errors in Parentheses)

	Conditional Decile								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Permanent:									
Small	4.23	1.95	2.46	3.00	3.00	5.42	8.82	7.00	3.23
	(4.0)	(3.5)	(3.8)	(3.3)	(3.4)	(3.0)	(4.5)	(4.1)	(3.6)
Medium	1.47	-0.68	-2.68	-4.00	-1.00	-0.08	1.10	-1.00	-1.11
	(3.7)	(2.7)	(2.7)	(2.5)	(2.6)	(2.3)	(3.2)	(3.1)	(3.3)
Large	-2.40	-0.40	-2.23	-3.00	-6.00	-4.53	-5.02	-5.00	-3.97
	(3.4)	(3.2)	(3.2)	(2.9)	(2.9)	(2.8)	(4.0)	(3.8)	(4.0)

	Conditional Decile								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Average:									
Small	3.97	2.67	1.52	2.00	-0.12	1.08	1.09	2.00	2.51
	(3.3)	(2.2)	(2.1)	(2.0)	(2.5)	(1.9)	(2.3)	(3.4)	(2.4)
Medium	----	----	----	----	----	----	----	----	----
Large	-2.18	-1.31	-0.88	-3.00	-2.32	-1.99	-3.01	-3.00	-0.57
	(2.9)	(1.8)	(1.7)	(1.6)	(2.0)	(1.9)	(1.9)	(2.6)	(2.0)

Notes: Additional covariates included, but not reported, are the same as for Table 2: gender, mother and father ethnicity, and the change in family income. The number of observations for both sets of regressions is 569, and standard errors are reported in parentheses. Each column in each panel represents a separate quantile regression.

Table 4
Assessing the Significance of the Small vs. Large Class Size Effects from Table 2

Permanent:		Average:	
Small Class [N=17]	7.91 (3.24)	Small Class [N=48]	3.51 (1.97)
Medium Class [N=32]	3.20 (2.92)	Medium Class [N=447]	2.19 (1.33)
Large Class [N=21]	----	Large Class [N=74]	----
Female	4.18 (2.41)	Female	-1.86 (0.89)
Mother Maori	13.02 (7.85)	Mother Maori	0.90 (2.41)
Father Maori	4.06 (4.10)	Father Maori	0.83 (1.68)
Change in Family Income	0.87 (0.46)	Change in Family Income	0.014 (0.21)
Constant	32.07 (2.76)	Constant	38.68 (1.35)
R-squared	0.21	R-squared	0.02
Number of Observations	70 (Sample Restricted Only to Those in a Permanent Class Category)	Number of Observations	569

Note: Cell Sizes for the Class Size Categories Given in Brackets. The regression in the leftmost column is purely to indicate the statistical significance in the differences in the small and large class size effects displayed in Figure 6 and reported in Figure 2. The substantial change in the sample definition accounts for the difference in the control estimates. The regression on the right, however, presents no new information from what was presented in the right column of Table 2 - here it is simply in a format that allows easier comparison of the large and small class size effects.

Table 5
Quantile Regressions for the Change in Burt Scores (Age13 - Age8)
Contrasting the Small vs. Large Class Size Effects
Average Class Size Categories Only
(Standard Errors in Parentheses)

	Conditional Decile								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Average:									
Small	6.16	3.99	2.40	5.00	2.20	3.08	4.10	5.00	3.07
	(4.1)	(1.3)	(2.5)	(2.4)	(3.0)	(2.3)	(2.8)	(4.0)	(2.9)
Medium	2.18	1.31	0.88	3.00	2.32	1.99	3.01	3.00	0.57
	(2.9)	(1.8)	(1.7)	(1.6)	(2.0)	(1.9)	(1.9)	(2.6)	(2.0)
Large	----	----	----	----	----	----	----	----	----

Notes: Additional covariates included, but not reported, are the same as for Table 2: gender, mother and father ethnicity, and the change in family income. The number of observations is 569, and standard errors are reported in parentheses. Each column in each panel represents a separate quantile regression.

Table 6
Comparison of the Table 4 Results With HLM:
Regression of the Individual Specific Trends on Class Size

Permanent:		Average:	
Small Class [N=17]	1.48 (0.70)	Small Class [N=48]	0.77 (0.40)
Medium Class [N=32]	0.49 (0.63)	Medium Class [N=447]	0.44 (0.27)
Large Class [N=21]	----	Large Class [N=74]	----
Female	0.75 (0.52)	Female	-0.30 (0.19)
Mother Maori	2.69 (1.70)	Mother Maori	0.14 (0.50)
Father Maori	1.07 (0.89)	Father Maori	0.22 (0.35)
Change in Family Income	0.21 (0.10)	Change in Family Income	0.00 (0.05)
Constant	6.51 (0.60)	Constant	7.76 (0.28)
R-squared	0.21	R-squared	0.01
Number of Observations	70 (Sample Restricted Only to Those in a Permanent Class Category)	Number of Observations	569

Note: Cell Sizes for the Class Size Categories Given in Brackets. To compare these results to those in Table 4, the approximate scaling factor is roughly 5 (= 6 time periods minus 1). The regressions are weighted by the number of per-person observations used in the regressions to obtain the per-person trends. We do not use the inverse standard errors in this case owing to the extremely small degrees of freedom (ranging from 1 to a maximum of 4 for each person), and the associated problems of a near-exact fit due to the small number of degrees of freedom. In this case, the number of valid observations on test scores for the individual is a more appropriate weight.

Table 7
Cross Tabulation of Small and Medium Class Size Categories by Public and Private School Sectors (Conditional on Not Being in a Large Class)

	Private School	Public School
Small Class	20 (18.4) [2.5]	33 (4.8) [4.1]
Medium Class	89 (81.6) [11.1]	656 (95.2) [82.2]

Notes: The numbers for each cell correspond to, the cell frequencies, the column percentages (in parentheses), and the overall table cell percentages (in brackets). Since the off-diagonal cells contain 15.2 percent of the data, this indicates that roughly 15 percent of the data seperately identify a private school effect from a small class size effect, relative to the base (or omitted) class size category of Large Class. As in Table 1, here again we utilize the full 873 observation sample, of which 798 observations are represented in this cross-tab.

Table 8
Effects of Average Class Size by
Public and Private School Sectors

Public Schools		Private Schools	
Average:		Average:	
Small Class [N=30]	5.52 (2.34)	Small Class [N=18]	-3.85 (4.25)
Medium Class [N=400]	2.39 (1.43)	Medium Class [N=47]	1.56 (3.77)
Large Class [N=65]	----	Large Class [N=9]	----
Female	-1.76 (0.96)	Female	-2.63 (2.49)
Mother Maori	-0.45 (2.61)	Mother Maori	12.24 (6.30)
Father Maori	0.98 (1.79)	Father Maori	0.11 (4.89)
Change in Family Income	0.04 (0.24)	Change in Family Income	-0.10 (0.51)
Constant	38.11 (1.46)	Constant	42.67 (3.55)
R-squared	0.02	R-squared	0.09
Number of Observations	495	Number of Observations	74

Note: Cell Sizes for the Class Size Categories Given in Brackets.

Table 9

Regressions for the Degree of Completed Schooling by Age 21

Class Size at Age 13	-0.015 (0.009)
Burt Score at Age 8	0.026 (0.003)
Female	-0.01 (0.10)
Mother Maori	-0.32 (0.26)
Father Maori	-0.12 (0.18)
Family Income at Age 8	0.06 (0.04)
Family Income at Age 13	0.11 (0.02)
Constant	0.87 (0.32)
R-squared	0.25
Number of Observations	549

Notes: The dependent variable is coded 1,2,3,4 corresponding to a recode of the TERTIARY variable provided by CHDS. The category 1 corresponds to no post secondary schooling/training. Category 2 corresponds to basic skills training, 3 to intermediate skills training, and 4 to enrollment for a bachelor degree. Bear in mind, this variable does not indicate completed schooling, but schooling as of age 21.

Table 10

Regressions for the Incidence and Duration of Unemployment Spells From Age 18 to Age 21

	Incidence (Robust Standard Errors) (Linear Probability Model)	Duration (Incidence = 1 Sample)
Class Size at Age 13	0.006 (0.004)	0.33 (0.12)
Burt Score at Age 8	-0.0005 (0.001)	-0.11 (0.03)
Female	-0.15 (0.04)	-0.78 (1.16)
Mother Maori	-0.12 (0.10)	3.87 (3.49)
Father Maori	0.13 (0.07)	1.64 (1.77)
Family Income at Age 8	-0.02 (0.02)	-0.05 (0.42)
Family Income at Age 13	-0.04 (0.01)	-0.74 (0.31)
Constant	0.61 (0.14)	7.91 (4.07)
R-squared	0.09	0.16
Number of Observations	578	213

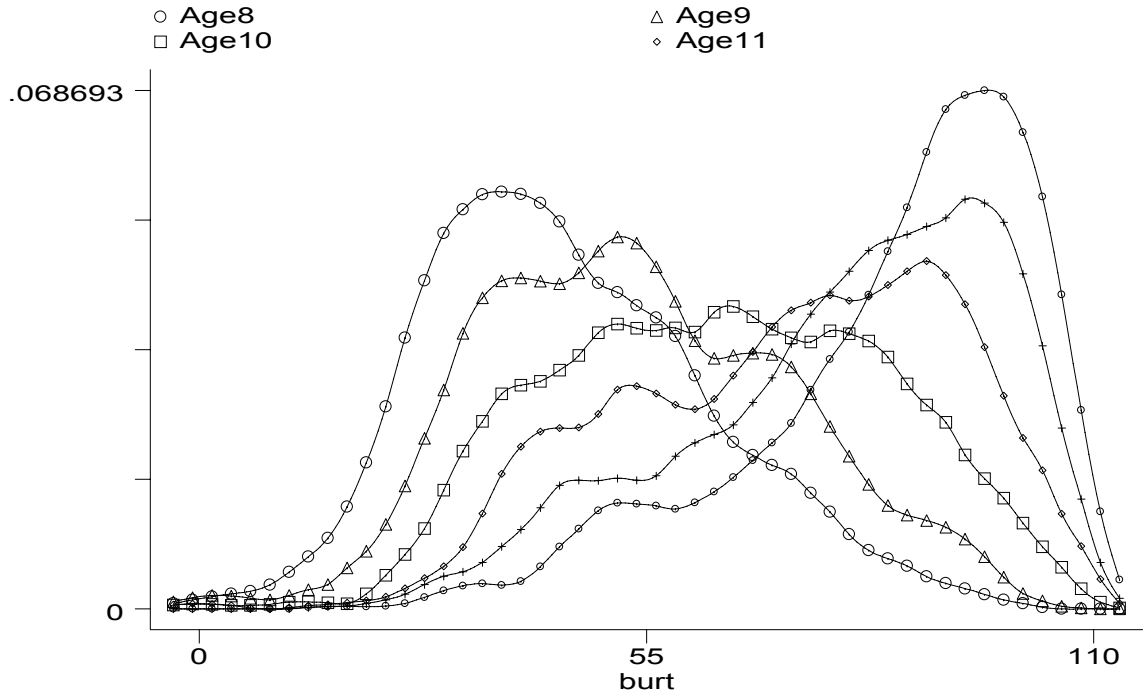

Figure 1: Kernel Densities of Burt Scores By Age

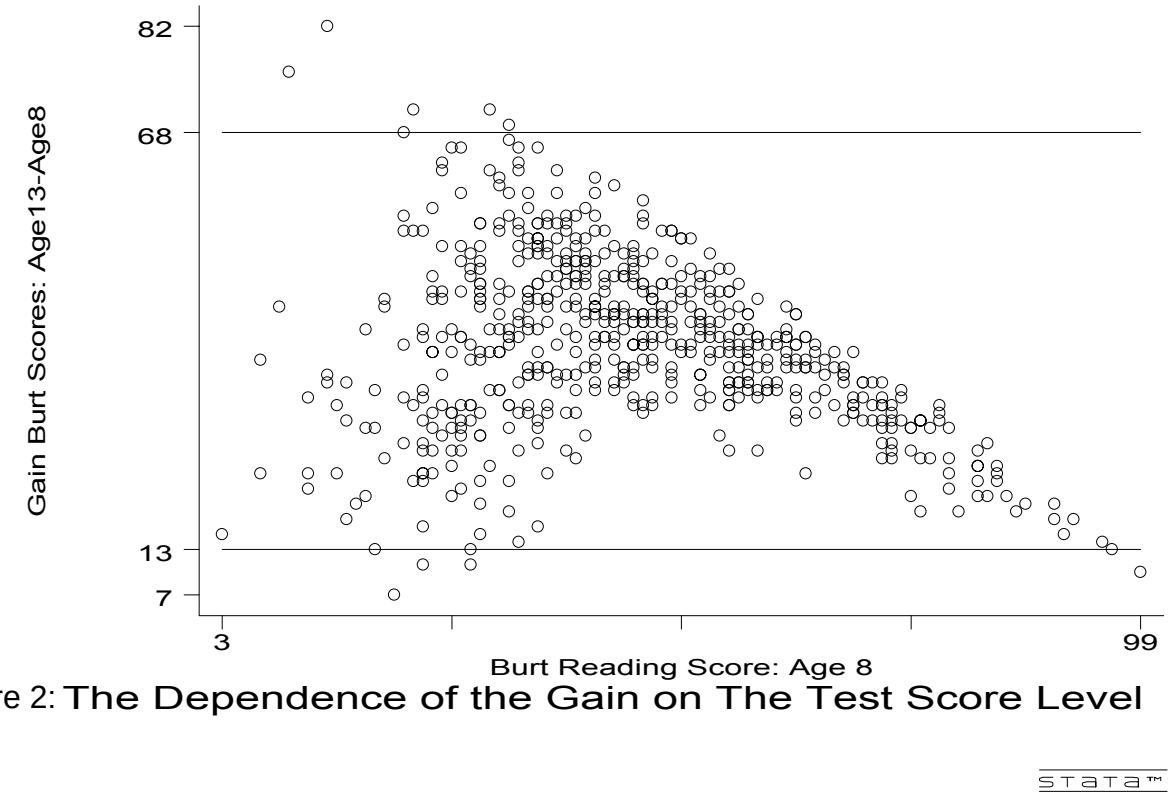

Figure 2: The Dependence of the Gain on The Test Score Level

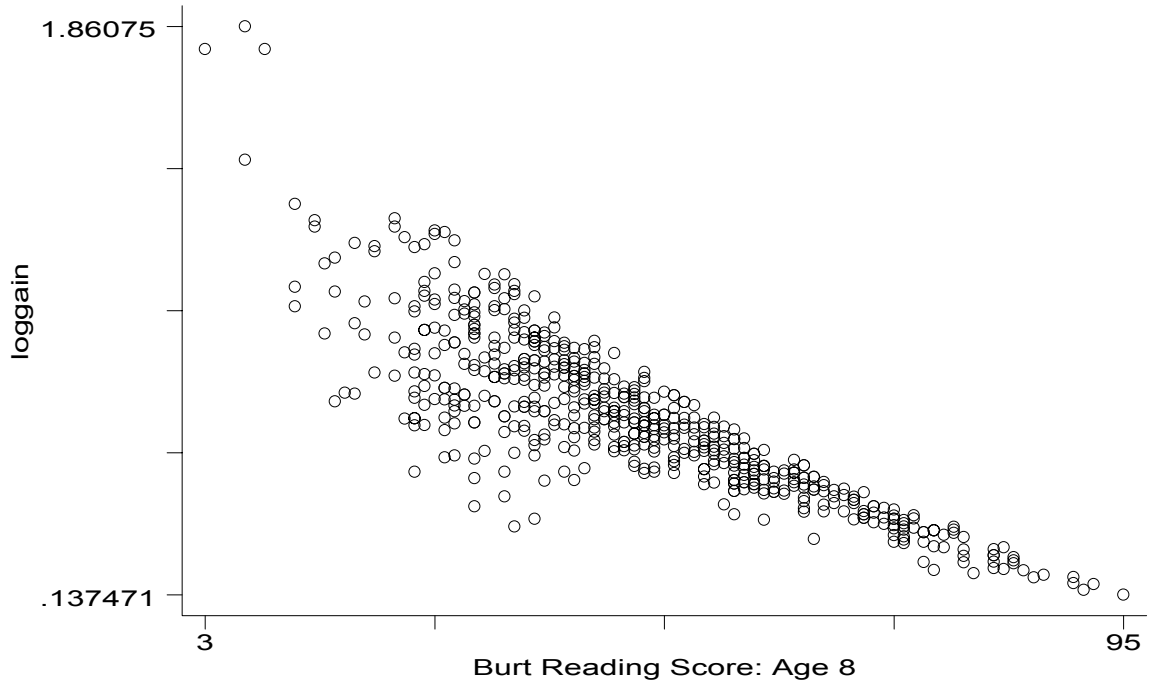

Figure 3: Log Gains vs. Initial Level - Trimmed Sample

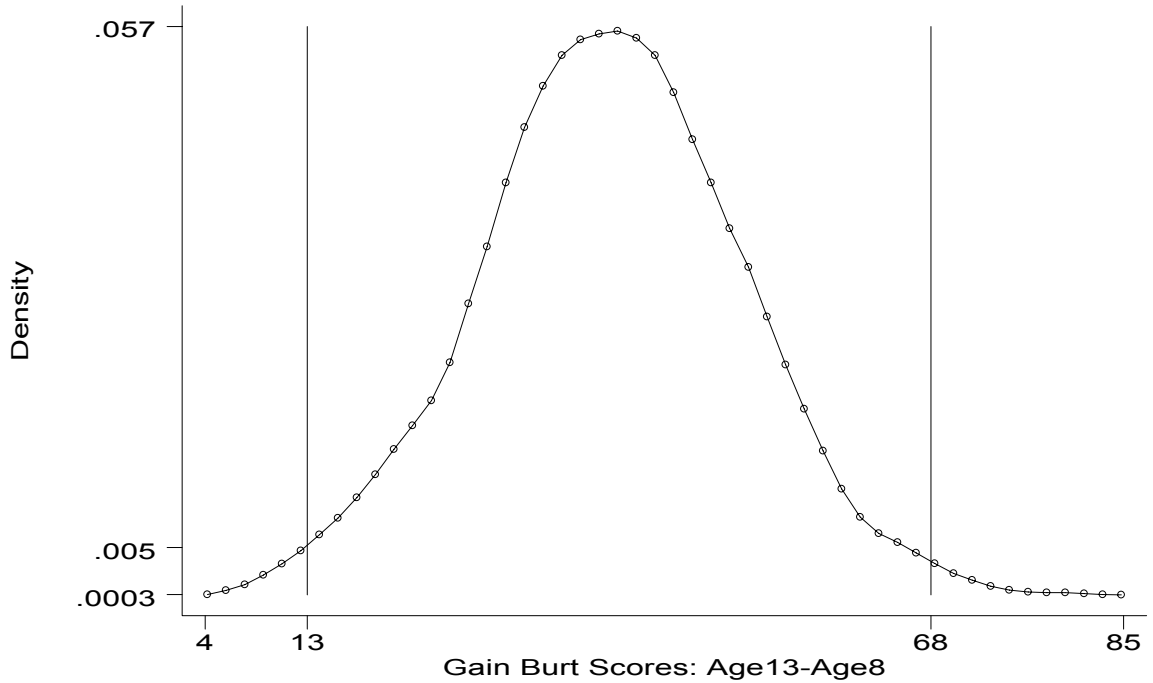

Figure 4: Kernel Density for Raw Gain Scores: Trim Points Noted

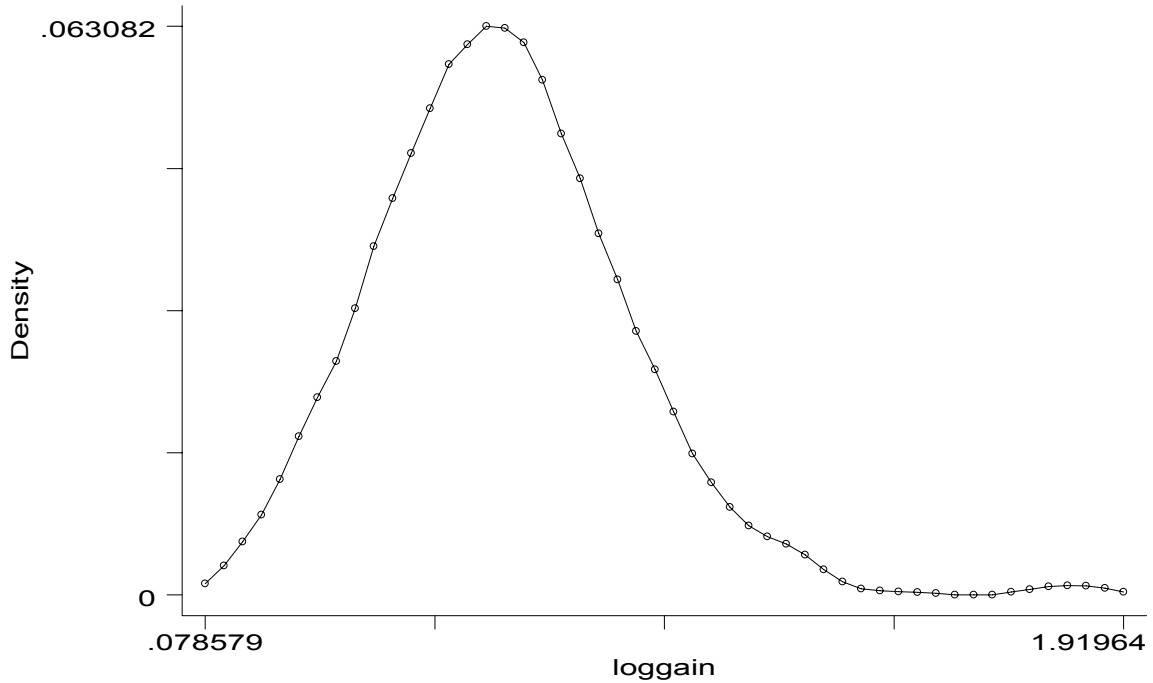

Figure 5: Kernel Density for Log Gain Scores - Trimmed Sample

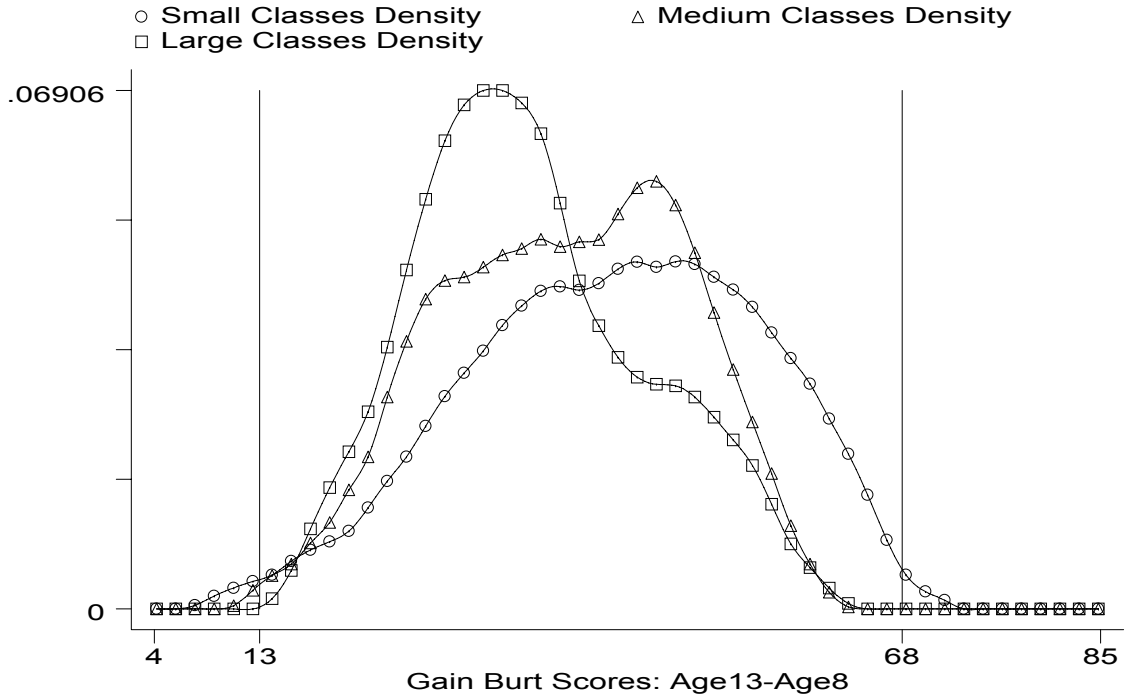

Figure 6: Kernel Density of Test Score Gains By Perm. Class Cats.

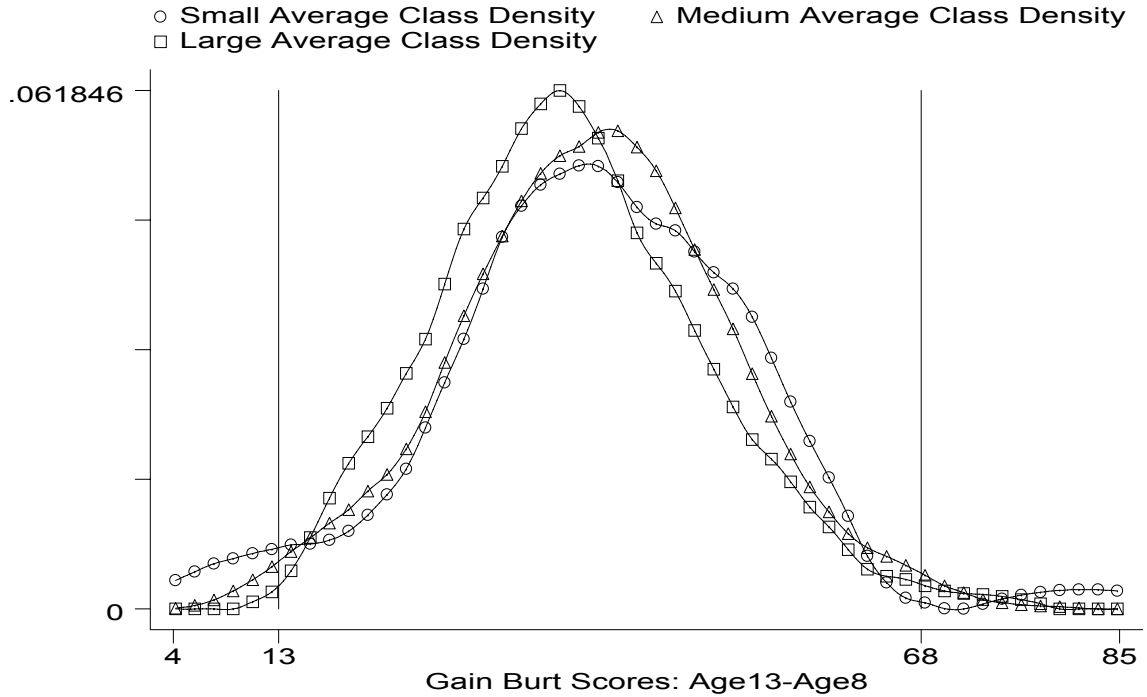

Figure 7: Kernel Density of Test Score Gains By Avg. Class Cats.

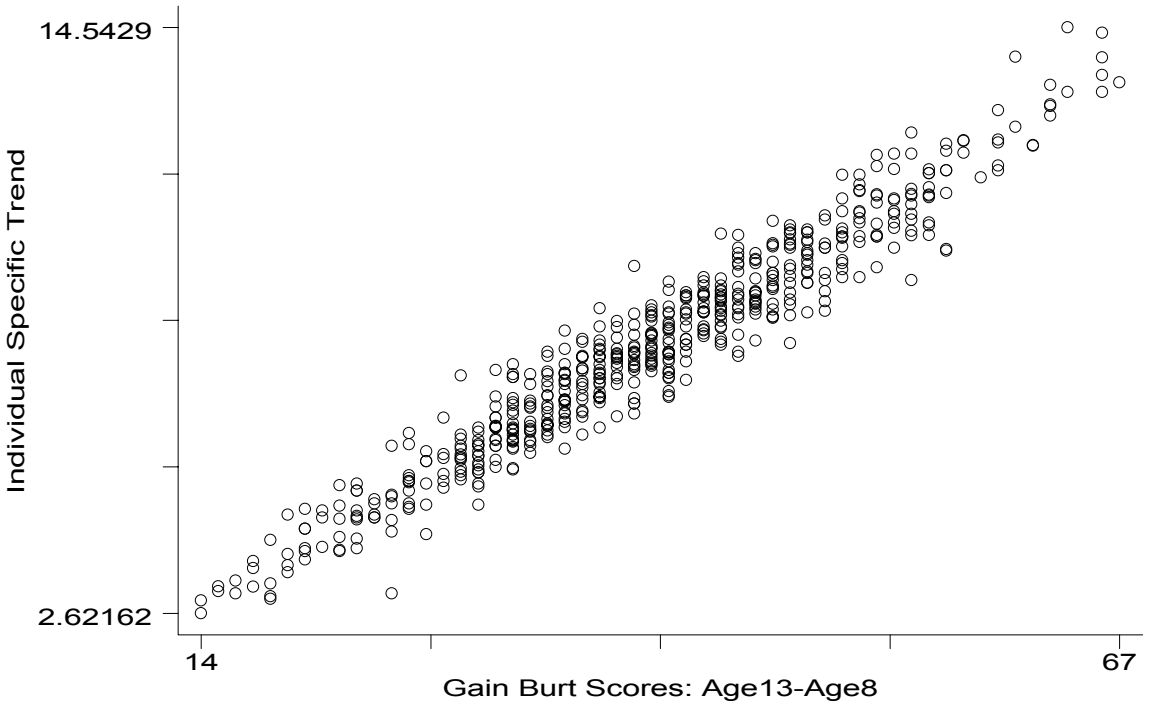

Figure 8: HLM Indiv. Spec. Trend vs. Age 8 to 13 Gain

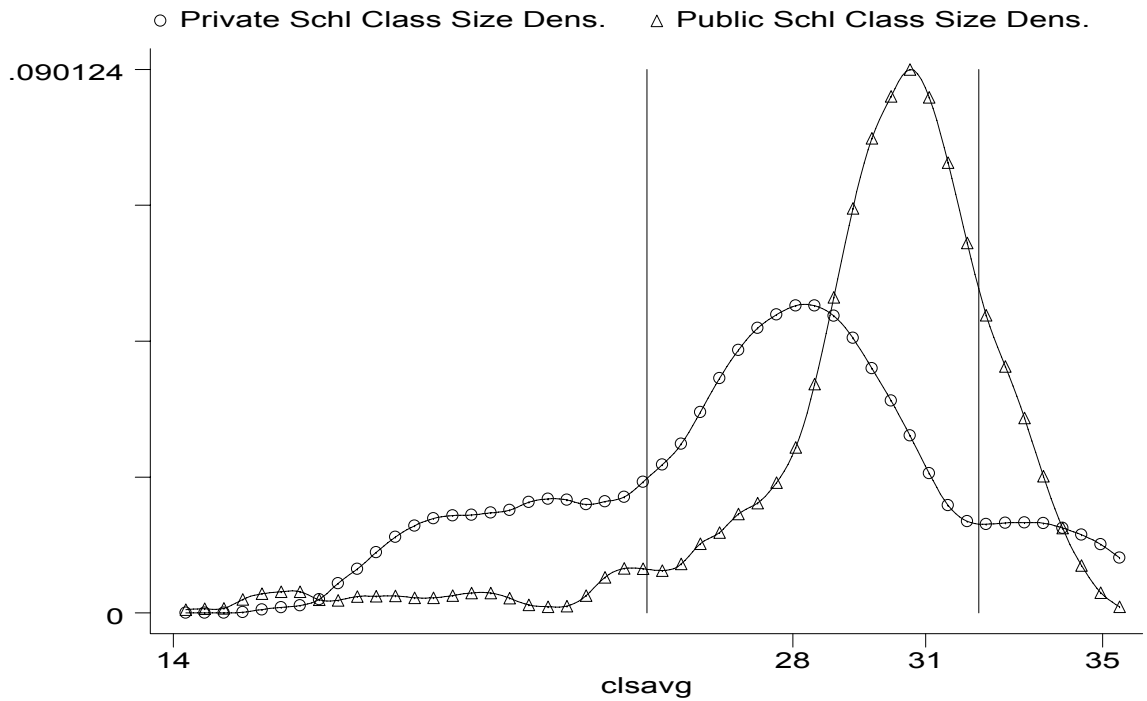

Fig. 9: Densities for Class Size by Public and Private

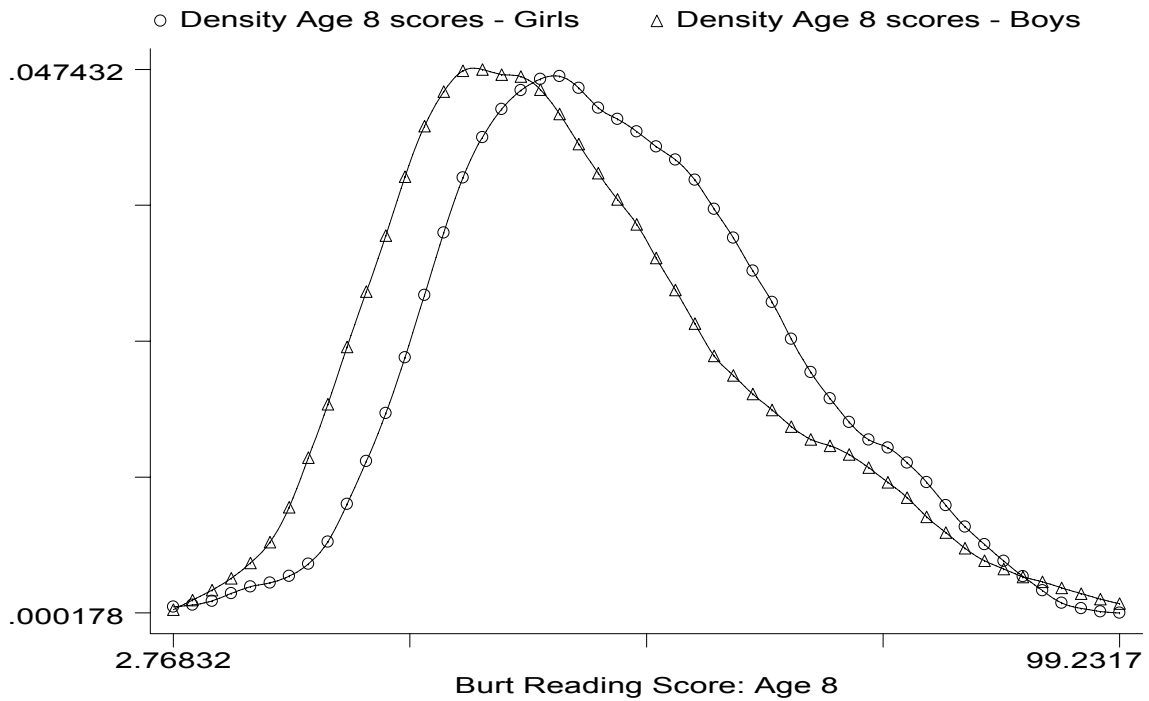

Figure 10: Densities By Gender of Age 8 Burt Scores

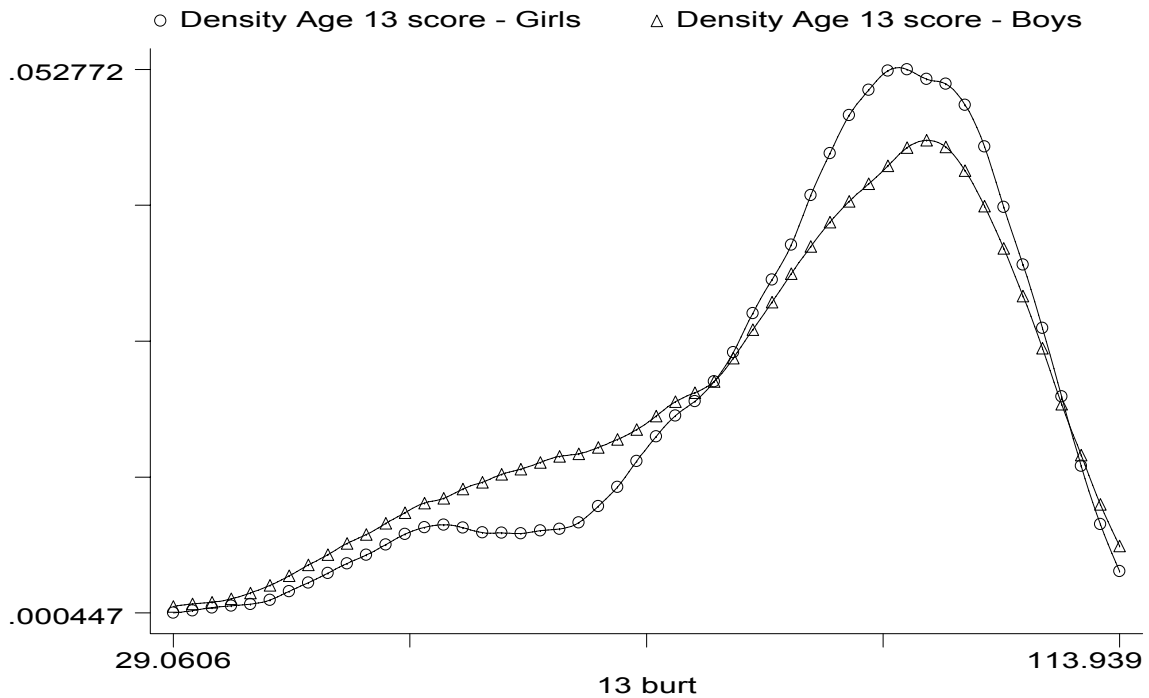

Figure 11: Densities By Gender of Age 13 Burt Scores

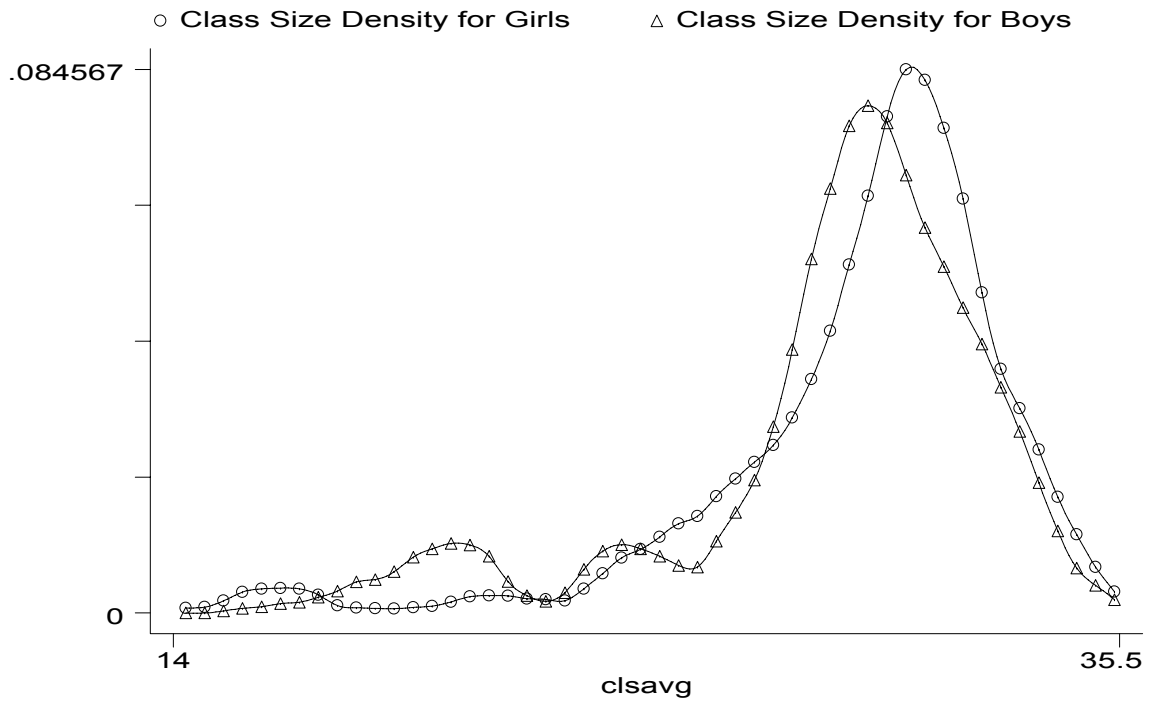

Figure 12: Densities By Gender of Class Sizes

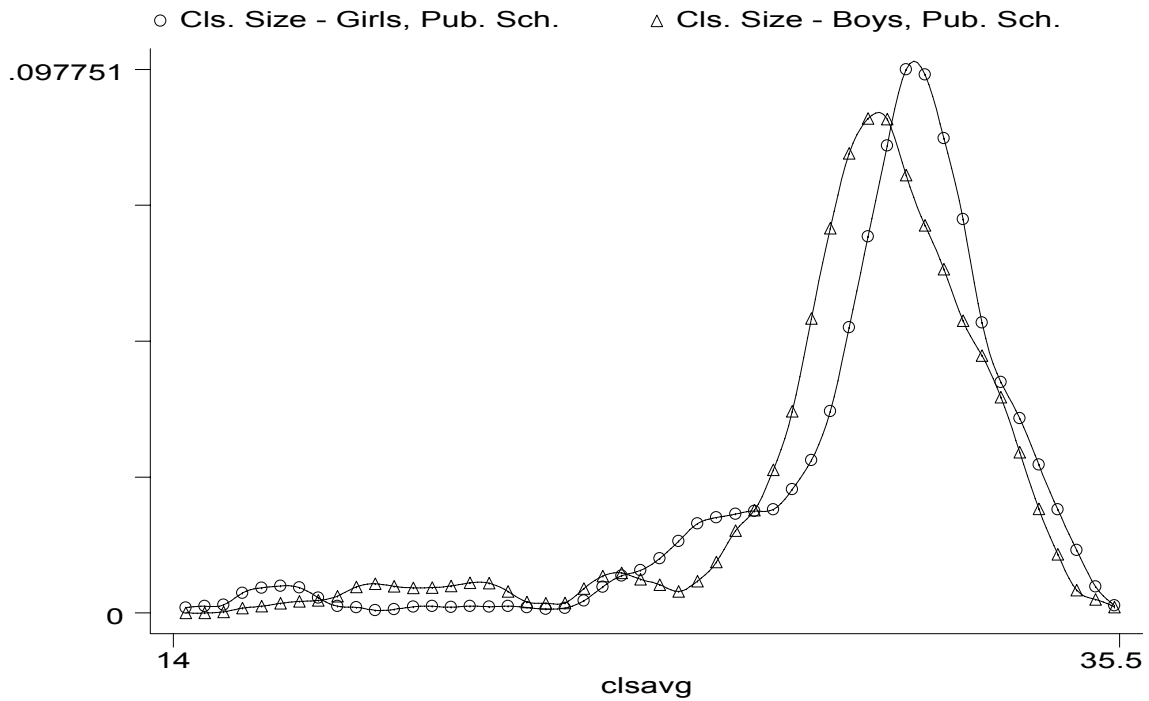

Figure 13: Dens. of Class Sizes by Gender - Pub. Schls

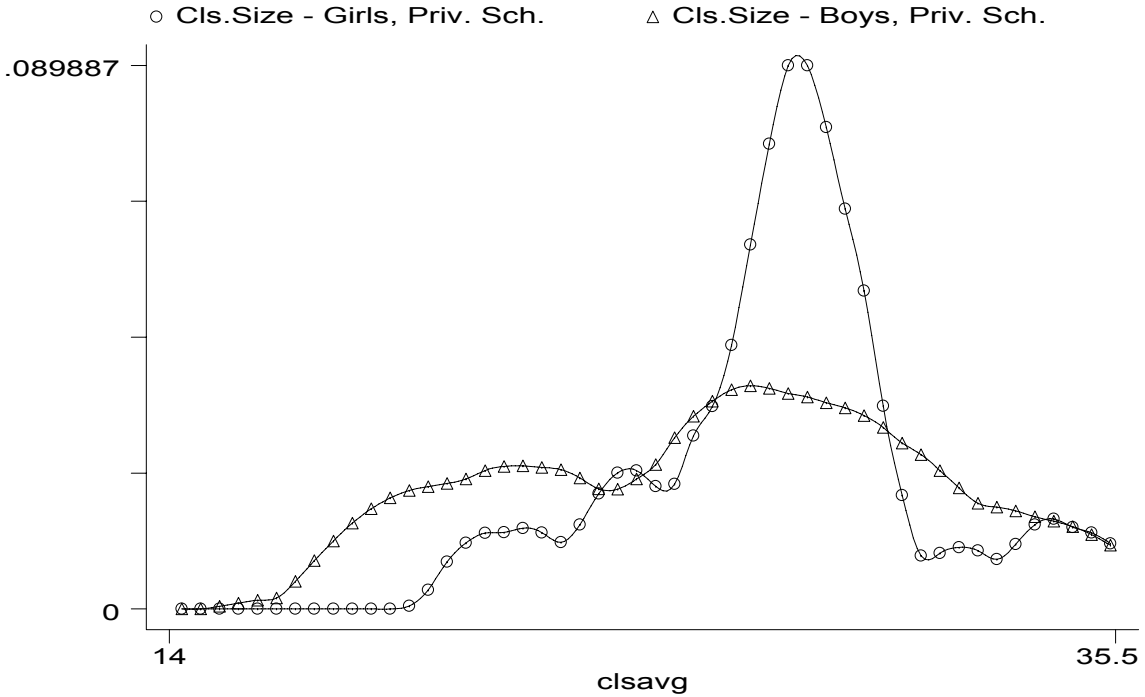

Figure 14: Dens. of Class Sizes by Gender - Priv. Schls