NERA Conference Proceedings 2008

Northeastern Educational Research Association (NERA) Annual Conference

10-23-2008

Using a Longitudinal Data Mart to Examine the Effects of Student Mobility on Test Performance Over Time

Richard F. Mooney
Connecticut State Department of Education, richardmooney@hotmail.com

Barbara Q. Beaudin
Connecticut State Department of Education, barbara.beaudin@ct.gov

Using a Longitudinal Data Mart to Examine the Effects of Student Mobility on Test Performance Over Time

Rick Mooney, Ed.D., Ct. State Dept. of Ed.
Barbara Beaudin, Ed.D., Ct. State Dept. of Ed.

ABSTRACT: Our analysis shows how a longitudinal data mart can provide a simple and effective way to analyze student test performance over time. Our data mart in this case is a mega-table compiled from several years of archival student-level test data, where we have modified all of the fields so that they have a common meaning over time. Using this longitudinal data base we then compared the performance statistics and effect sizes of test results in math and reading on the Connecticut Mastery Test (CMT) series for grades 4, 6 and 8 and on the Connecticut Academic Performance Test (CAPT) in grade 10, from 2000 to 2007.

We found that students tested sequentially in grades 4, 6, 8 and 10 achieve better performance in mathematics and reading at the State, ERG and school district levels, as compared to new incoming students who began the testing sequence sometime after grade 4. This suggests that mobility relates to lower student performance on our tests, a finding that others have reported (Bourque, Mary D., 2008, Rumsberger, 2002). We conclude that student mobility should be monitored and that academic and/or social interventions may be warranted. We also conclude that a longitudinal data mart may provide a practical way to look at student test performance over time particularly when vertical scaling or vertical modulation are not available. A data mart could also serve as a simple low-tech way to cross-validate results from these techniques.

Introduction

The focus of this report is to investigate the impact of student mobility on academic test performance using a longitudinal data mart. We have developed vertical scales in Connecticut using IRT and we think that this will provide a very useful way to examine student performance over time in the future. However, it will take several years to develop sufficient test results on vertically equated forms to be able to look at long-term student performance using this model.

Meantime, we have many years of archival data that may also be useful for looking at student performance over time. Therefore, we wanted to see if data mart technology might offer a good alternative for evaluation of the longitudinal performance on Connecticut's student assessments using this existing archival data. Finally, it is often useful to have more than one way to examine data. Comparing how different models converge and diverge might provide us with a better way to evaluate the real effects and also a way to appreciate the contribution of each of them. Therefore, we also thought that

A paper submitted to the Northeastern Education Research Association (NERA)
39th Annual Conference, October 22-24, 2008

this longitudinal data mart approach might also be useful for providing an alternative view of repeated measures analyses of test results in the future.

Most experts agree that vertical scales or vertically moderated test standards are the best ways to analyze test performance over time—and they may be—but questions remain about these approaches (DePascale, 2006). For example, vertical scales require IRT scaling and linking but the vertical scale may have different constructs due to the progression in test content from grade to grade and therefore may violate underlying IRT dimensionality assumptions (Lissitz, R.W. & Huynh, H., 2003). It is also unknown how robust IRT models are to these possible violations. A popular method of vertical equating involves using anchor items, but the choice of anchor items is also an area of concern. This is because these choices define and limit the growth potential of the scale (Patz, R.J. 2007). Finally, although different horizontal equating techniques all yield very similar results, different IRT vertical scaling techniques yield different results and there is as yet no consensus on the best method (Kolen, M.J. & Brennan, R. L., 2004).

Vertically moderated standards require less stringent assumptions than vertical equating Mislevy (1992) and Linn and Baker (1993), but require a mixture of 'policy equating' and 'linear statistical adjustment' such as smoothing or extrapolation (Lissitz, R.W. & Huynh H. (2003). Although Lissitz and Huynh support vertically moderated standards the technique has at least as much art as science and therefore might not achieve the desired consistency in meaning across grade levels as was intended.

While we believe that vertical equating holds promise as the best model for measuring test performance over time, even so, it is limited to a sequence of tests that have common content measured in very similar ways. For example, our lower school or CMT series of academic assessments measure student performance using the same structures, although the content becomes progressively more difficult. However, the high school assessment or CAPT, although it covers the same general content, has sufficient differences that it would not be appropriate for including in a vertical scale combined with the CMT sequence.

Vertical equating is also limited to current and future data, rather than our archival records. However, the data mart approach is flexible enough to allow us to work with archival data and to look at performance differences that span from the CMT assessments through to our grade 10 CAPT assessments. Thus, a data mart may provide a sound practical way of assessing student test performance over time, even in cases where vertical scaling or vertical modulation are not available or appropriate.

What are the Advantages of Longitudinal Analysis?

Longitudinal analysis has important advantages over designs that compare different student cohorts over time. Traditional cross-sectional comparisons based on different groups of students are subject to considerable aggregate score volatility due to the effects of differences at baseline (Kelly & Monczunski, 2007). These efforts may mask the gains that we are trying to measure. On the other hand, by following the same students over time, we can control for baseline performance differences, thus enabling us to draw valid conclusions about educational effects occurring between one grade and the next.

Connecticut's Archival Test Data

The Connecticut Mastery Test (CMT) and the Connecticut Academic Performance Test (CAPT) have been in place for since the early 1980's. This data provides us with an extensive archive of individual student records. The tests provide grade specific criterion-referenced information about student performance for grades 4, 6, 8 and 10. Since 2006, with the advent of the NCLB laws, the CMT has expanded to test all students from grades 2 through grade 8. This change also reflects a shift from fall to spring testing for the CMT and a new test generation (generation 4).

For this analysis, our intent is to look at the performance of a cohort of students that began taking the generation 3 CMT in grade 4 of 2000. We will follow this cohort through the traditional Connecticut testing progression prior to NCLB of grades 4 (2000), 6 (2002) and 8 (2004) and finally to the high school or CAPT assessment in grade 10 of the spring of 2007. After 2006, the CMT was administered in the spring for grades 3 through 8 while the CAPT continued to be administered in the spring for grade 10. Note that because the older generation CMT for grades 4, 6, and 8 tests were administered in the fall, the content for these tests reflects the previous grades (i.e.: 3, 5 and 7). This historical detail should cause us no difficulty in interpreting these findings appropriately using the data mart approach, but this time shift would represent a significant challenge to anyone attempting to employ a vertical equating model on these archival data.

Student performance on the CMT and CAPT are categorized into five common graduated status indicators ranging from "Below Basic" to "Advanced". While these criteria are useful benchmarks for marking progress within each grade level, they may not be useful for making comparisons across grades. One might ask, is an observed performance difference due to growth or due to differences in the status indicators across the grades? Vertical moderation is a method of adjusting of status indicators to make them meaningful for making valid determinations about performance growth over time. Although vertical moderation has not been done in Connecticut—since we were already committed to vertical scales. However, as it happens the vertical scales developed for the CMT indicate that the standards for grades 2 through 8 do show a meaningful sequential progression in performance difficulty over time.

What is a longitudinal data mart?

Data marts and data warehouses are a relatively new and evolving analysis strategy developed from the fields of computer science and database technology (William Inmon, 1999). The idea is first to take data that may have some or many inconsistencies over time and restructure these so that the data are placed on a common footing with respect to meaning and interpretability. Once the data are on a common footing, the data mart facilitates generating on-the-fly analyses for decision makers. This common sense approach does not address the issue of appropriate statistical comparisons of performance data over time. We resolve this issue by comparing test performance for mobile and non-mobile groups using the same grade and year test results. We further compare the differences using effect size statistics, which are scale free and not affected by differences in group sizes. Thus, we believe that this longitudinal data mart framework provides a simple and practical way to convert archival data into a tractable structure for repeated measures analyses.

The definitions of the term data mart and data warehouse are new and remain somewhat unstable in 2008. However, there is at least a clear difference in scale. The term data warehouse refers to overarching cross-departmental analyses while a data mart works with a narrower scope, such as inter-departmental data. A data mart can refer to a system that uses record level data (for the greatest analytical flexibility) or summary data. Some experts define a data warehouse as the intersection of a collection of two or more data marts. William Inmon (1999), considered the father of data warehousing, reverses this definition and defines a data mart as more legitimately drawing data from a warehouse rather than thinking of a data warehouse as a collection of data marts.

Regardless of this debate about definitions, we have developed what we think of as a freestanding longitudinal data mart that allows us to analyze student-level test data across a variety of background variables from the year 2000 to the year 2007. We think of our model as a data mart because of the narrow purpose. Our intent was to enable us to easily process and analyze assessment trends using matched student level data. The details of the development and programming details using SQL have been described elsewhere (NERA, Mooney, R and Beaudin, B., 2007).

Defining Matching and Non-Matching Records

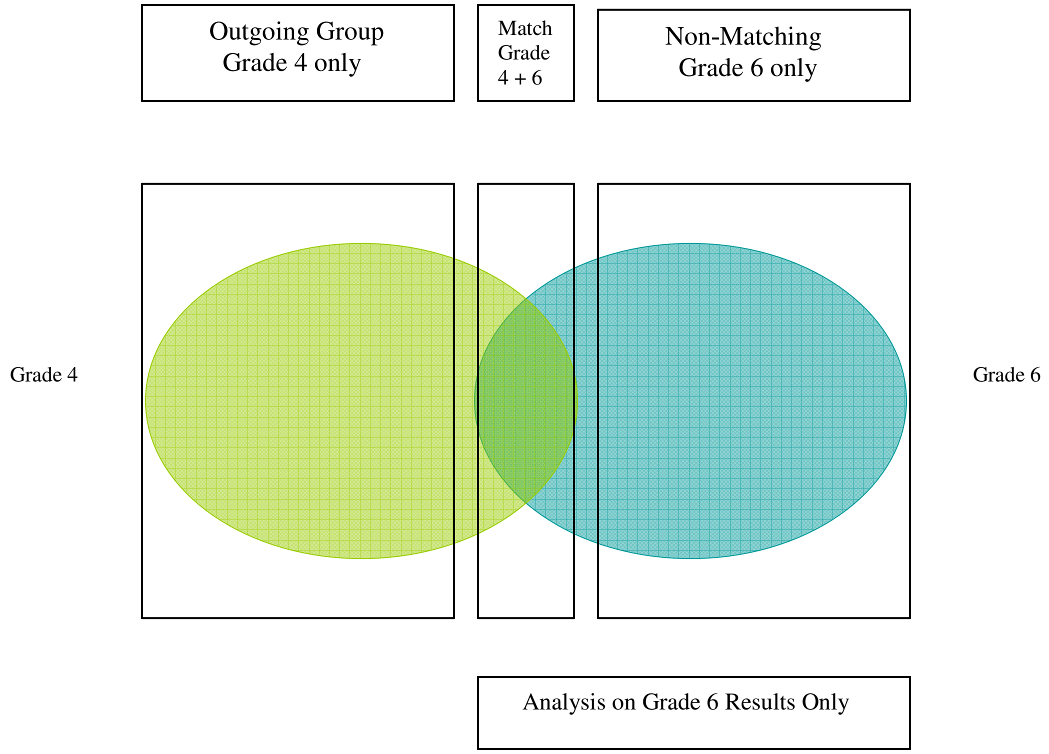
In this paper we will demonstrate our data mart system by comparing CMT and CAPT test performance for a cohort of students who progressed through our testing program from the year 2000 in grade 4 to the year 2007 in grade 10. The analyses will compare the students assessed at each grade in the 4-6-8-10 sequence to those who began testing sometime after the grade 4 test in 2000. We defined the students with a full matched sequence of test scores spanning from grade 4 to grade 10 as the "Stable" group. We define their age/grade companions who transitioned into test program at some time after grade 4 as the new or "Incoming" group. Another way to think about this is to consider students with matching records as sequential test takers and the non-matching group as those who have taken some but not all of the tests over the years.

In this analysis, matching records are thus a proxy for students who have a complete sequence of valid test records from one grade to another grade. We allow matching records to accrue in a cascading process. Thus, the records that match from grade 4 to grade 6, must also match from grade 4 to grade 6 to grade 8 and so on through to grade 10. Stated differently, as the matching sequence progresses, only the matching records from grade 4 are eligible to remain in the match group for the next grade level comparison.

As the progression moves forwards in time, outgoing students that do not participate in the next grade level examination automatically drop from the model. Thus, the grade 4 to 6 cohort can be partitioned into three groups. The first sub-group is the stable examinees that include all those students with matching test records for grade 4 as well as grade 6. The incoming sub-group reflects the records of students who took the grade 6 test but did not take the grade 4 test. The final group is the outgoing students who took the grade 4 test but did not take the grade 6 test.

Diagram 1 (See Below) illustrates the partitions for the matching and non-matching conditions. The circle to the left represents all the grade 4 records and the circle to the right represents all the grade 6 records. The intersection of the two circles describes all the matching records—those that took both the grade 4 test and the grade 6 test. The records to the left of the intersection describe those students who took the grade 4 test but did not take the grade 6 test. These are outgoing students because they left the testing program. The circle to the right of the intersection describes those examinees who took the grade 6 test but did not take the grade 4 test. We call these incoming students because they are coming into the testing sequence for the first time in grade 6.

Diagram 1: Matching and Non-Matching Records



Each of the comparisons across the grade 4-6-8-and 10 sequence will occur at the same grade and test level for each stage of the analysis. So for example, when the cohorts are compared for the first time this occurs only on the grade 6 test results. The next stage or comparison in the analysis will take place on the grade 8 results and the final comparison will occur using the grade 10 results. Another way to say this is that the records we analyze are only those that exclusively inhabit the right hand circle of diagram 1. Within that circle, the two sub-groups that we analyze are the matching group and the non-matching group. We do not analyze the outgoing group in the circle to the left.

More specifically, the matching or stable group consists of students in the intersection area of the left hand circle of diagram 1. These students have been in Connecticut public schools for both the grade 4 and the grade 6 examinations. The non-matching or incoming students, on the other hand, will consist of the data records for those students tested for the first time on the 2002 grade 6 test and who came into the testing program at some time after the grade 4 test. In the next stage of the analysis, the comparison sequence will focus on the grade 8 test. The stable group will then consist of those who took the grade 4 CMT, plus the grade 6 CMT as well as the grade 8 CMT. The records of the incoming students will reflect those who began the testing sequence at some point after the administration of the grade 4 CMT. The third stage will progress in this same manner until the grade 10 CAPT test in 2007.

Stable (or matching) records and incoming (or non-matching) records are best considered classifications that differentiate students who have been tested at each grade level in the sequence compared to those who have not been tested at each grade level in the sequence. Hence, the data mart model is less complicated than the vertical models because the performance comparisons in this case are always on the same footing, rather than based on rescaled results that span two or more testing occasions simultaneously. While the vertical scale model depends on a new scale in order to base the performance comparisons, the data mart model uses the normal on-grade scales and status indicators. This makes interpreting performance differences simpler and more direct.

Systematic and Random Error in the Longitudinal Model

Matched (stable) and unmatched (incoming) records have two sources of error: Systematic and random. Random error is a kind of noise factor that gets smaller as group size increases but systematic error is bias that is unaffected by group size. Therefore, random error is a problem when student groups are small, but systematic error can be a problem regardless of group size. This is the biggest concern in post hoc or archival analyses such as this. In order to avoid this problem, we must identify and control for all known sources of bias and be vigilant for any we may have overlooked as the analysis unfolds.

One important source of bias in this analysis has to do with the way records become matches or non-matches. A legitimate match results from the records of examinees with a progression of valid test scores over time and a non-match is supposed to reflect a new or incoming student. However, if the key fields that identify each record are ambiguous (not uniquely identified) then a variety of potential record mismatches can arise.

Connecticut has assigned a unique 10-digit numeric student ID since 2004. However in order to look back further in time, we must use an alternative strategy to link up records from when this unique ID system did not exist. In the past, we have utilized a thirteen character composite ID for inter district analyses but this model is imperfect for intra district analyses because the index is not sufficiently discriminating. To resolve this, a new composite ID was created for the data mart based on 6 letters of last name, 5 letters of first name, DOB (YY-MM-DD) and student sex. This new eighteen character ID does a better job of discriminating between matching and non-matching records statewide. Still, imperfect matches can result in two types of problems that we will consider artifacts of this type of processing. They are:

1) Mini Cartesian Joins: A Cartesian Join occurs when every record from one table matches every record from another. However, a single ambiguous record from one year that links incorrectly to one or more ambiguous records in a subsequent year will result in a duplicate record;

2) Broken Records: This occurs when a valid record from one year may not be located in a subsequent year due to an ID entry error or name change, therefore an examinee with a valid prior test score will not receive a link, and ultimately credit for that score.

To resolve mini Cartesian joins, we automatically delete all duplicate records from match groups. That is, whenever we match one set of test results from a prior year and grade to another set of test results from a subsequent grade, we automatically delete all duplicate records. These duplicates only affect a few records and are not easily resolved without detailed research, so it is simpler to delete them from the analysis. The number of duplicate records for a match set of 35,000 cases tends to be in the neighborhood of 25 records, so we feel this is sufficiently tiny to consider negligible.

Similarly, we also consider broken records to be an error effect that can be safely ignored. Broken records automatically become part of the non-matching record group and thus these are treated as new incoming students rather than as preexisting or “stable” students. These cases result from what we regard as random events such as name changes for regular students or data entry errors for rare cases of bubble-in students who move during the testing window. Although it is intuitively sensible to ignore these cases, it is difficult to know precisely how many of these exist in the files.

In cases where we have used the new random SASID ID to match cases over time, we have also found 25 to 30 records for matched sets of 35,000 cases, often due to legal name changes. For the time being, this will stand as our best guess as to the approximate number of broken records in a standard grade-to-grade match. In the future we will be able to compare results for the name based composite ID to the new unique 10 digit number ID and obtain a more precise fix on the differences between the two systems, but for now we will regard them as a kind of random error.

False Positive Records: False positives can occur when records are joined incorrectly across two testing periods. This would be the result if the matches are incorrect because one record was from one student and the matching record was from a different student. In order for this to happen, two students would have to have the identical name (given our constraints), birth date and gender. Hence, it could occur when two different students have the identical name and background information, or if a bubble-in student incorrectly enters one or more components of their name or personal information. This type of error by definition is limited to matching group. Because duplicates are already removed from the system, the only way for this to occur is when the first entry would have to reflect a student who left the system and the second student would need to be new to the system. Since we expect that this will be a very rare mistake, we have also defined it as random error.

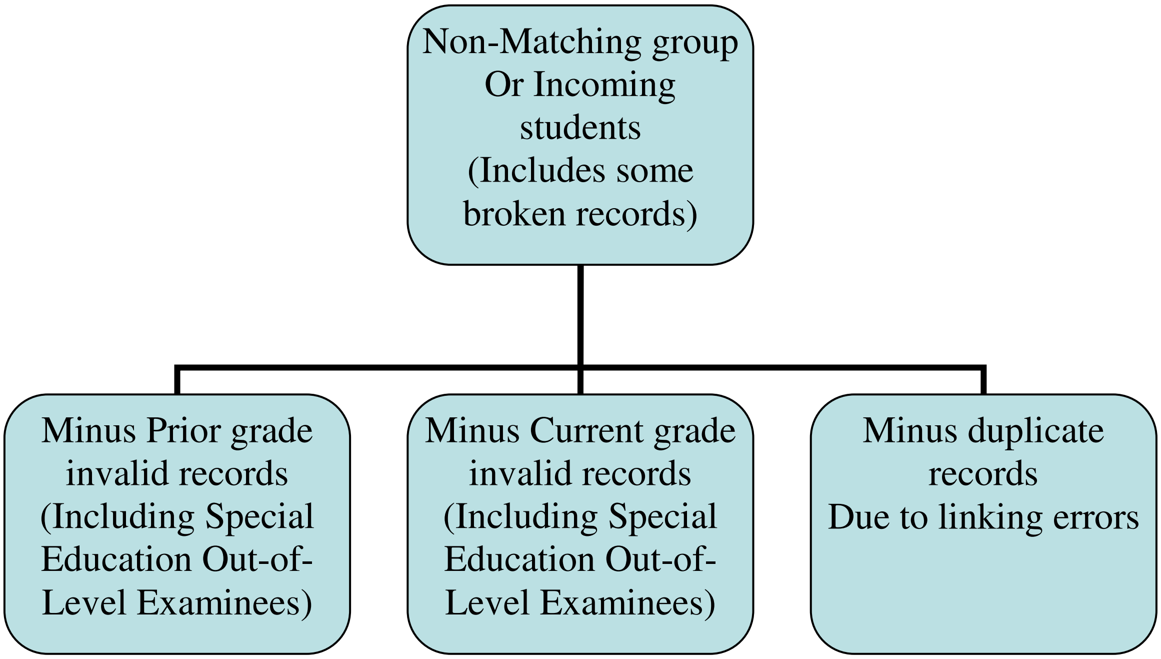
Previously Invalid Students: One known source of serious bias occurs from the scores of previously invalid students. These students submitted invalid test results in the prior grade and did not take a makeup test. This subgroup may appear to be similar to the legitimate incoming students who are not tested in a prior grade, but this group is different because they were physically present but did not take the exam. Therefore, they may or may not have characteristics similar to the real incoming students. One way in which they may differ is that they may sat for the test and had some testing experience even though they did not receive a valid test score. We will demonstrate that combining these students with the legitimate incoming students' causes lowered scores and therefore biased results.

A potential source of bias for invalid records results from the policy of having qualifying Special Education students take out-of-level examinations from the year 2000 to the year 2005 and marking the on-grade test results as invalid. This causes a bias because these examinees as a group tend to be lower performing. We will demonstrate the effects of keeping this group in the analysis or not in the results section. Our refinements applied to the non-matching group are illustrated in diagram 2 (below):

Testing Error is another source of random error. All tests measure a student's underlying ability with some degree of inaccuracy, reflected in the standard error of measurement for the test. Therefore, we assume that the influence of these random errors is negligible compared to mean performance estimates when group sizes are adequate.

Causality: Although matching records over time is a better model than using unmatched groups, claims of causality in any post hoc analysis are not warranted. These results are best seen as descriptive measures. It is the nature of this type of research that looking backwards at the behavior of human beings and attempting to draw conclusions about what happened is tricky business. Because it is a form of speculation, mistakes will inevitably occur. The real danger with post-hoc research is the temptation to make causal interpretations, especially when they seem to make compelling sense. These interpretations can be misguided when the variations that we are observing may actually caused by yet another influence that has not been included in the model. This is a type of specification error. Validating findings from this type of research would require experimental research techniques or at least direct evidence obtained from a sample of schools to determine whether these findings cannot otherwise be explained by other factors that we did not control for in this study, see Linn (2008).

Diagram 2: Defining Non-Matching Records as Purely Incoming students


Non-Matching group Or Incoming students (Includes some broken records)
Minus Prior grade invalid records (Including Special Education Out-of-Level Examinees)
Minus Current grade invalid records (Including Special Education Out-of-Level Examinees)
Minus duplicate records Due to linking errors

Research Question

Our research question is: "Does test performance differ significantly for examinees who have taken the CMT and CAPT sequentially from grade 4 through to grade 10, as compared to those examinees who have come into the testing program at some time after the grade 4 examination?"

Statewide Results

At the state level, incoming students are most likely either new to the public education system in Connecticut or else moved into Connecticut from another state (incoming students). One consideration in analyzing these data is whether to roll up incoming student data or to treat these records as incoming for some period and then after that period consider them stable? If instability has lasting effects, we reason that it would be better to included these students in the incoming group for the full experimental period (2000 to 2007). Thus, if a student comes into the testing program at grade 6, they permanently remain in the incoming group as they progress through to the grade 10 test. In our view, the biggest concern we had was that we wanted to identify whether or not there were long term effects of instability, therefore we decided to keep these students in the "incoming" pot for the duration of the analysis. If we are wrong about this, then we would expect that the worst that could happen is that the results would diminish over time.

We also want to note that there have been changes in the testing program that could also affect how these longitudinal analyses can be conducted and interpreted. For example, while grades 3 through 8 and 10 are tested now, only grades 4, 6, 8 and 10 were tested prior to 2006. Therefore if we intend to look back in time to grade 4 students in the year 2000 there will only be three legitimate comparison points (i. e.: grade 4 to 6, grade 6 to 8 and grade 8 to 10).

Another policy change that the reader should be aware of is that prior to 2006, the CMT was a fall testing program and after 2006 it became a spring testing program. This means that the tested content prior to 2006 really measures previous grade content and after 2006 the CMT measures current grade content. So to compare the results fairly, the grade content levels of the CMT will actually be adjusted to grades 3 (2000), 5 (2002) , 7 (2004) and 10 (2007). Since the grade 10 CAPT test has always been a spring test, no grade adjustment is required. For the sake of simplicity, we will continue to call the CMT tests by their actual names, which are the grade 4, 6 and 8 CMT.

We begin with a forwards-in-time analysis using the Mathematics test to examine the impact of matching against non-matching records. The time span is also important. This analysis looks at sequential test results that progress forwards in time for matching and non-matching records. These junctures occur at three different levels or stages. Stage 1 is for grade 4 students in fall 2000 to grade 6 in fall 2002. Stage 2 is from grade 6 in the fall of 2002 to grade 8 in the fall of 2004. Finally, stage 3 is from grade 8 in the fall of 2004 to grade 10 in the spring of 2007. These conditions are then partitioned at each of these three stages into new incoming students who began taking a Connecticut statewide mathematics test sometime after the administration of the grade 4 tests, the grade 6 tests or the grade 8 tests and stable examinees who took all of the tests.

The non-matching or incoming group is increasing in size from comparison to comparison and the matched or stable group is declining in size. This is because all of the incoming students from grade 4 to grade 6 are being accumulated with the incoming students from grade 6 to grade 8 and then again from grade 8 to grade 10. Meanwhile the stable group is shrinking because they can only be derived from the initial grade 4 examinees who took the CMT in fall 2000, so this cohort number can only diminish in time due to attrition.

Difference Test: To provide the reader with an index of relative differences the matching and non-matching groups appear in the “Diff Test” column of Table 1. The first column is the difference between the scale score mean of the non-matching group subtracted from the scale score mean of the matching group. The next column displays a scale free effect size statistic known as Cohan’s D (Becker, L. 2007). These tests look at scale scores differences using the Cohan’s D statistic.

We interpret classical statistical tests by comparing empirical results to the theoretical normal sampling distribution. This model assumes that a particular random sample has certain properties that ought to fall within two standard deviations of the normal sampling distribution. However, when census data are analyzed using classical statistical tests such as t-tests the when findings are based on population parameters. In addition, because the sample sizes are so great, the tests have a high degree of precision and become overly sensitive to minor fluctuations in the means. This means that any differences—no matter how trivial—are likely to be “significant.” Paradoxically, the only truly meaningful results occur when the classical statistical tests are not significant, because this means that we can be comfortable saying that there is no difference in the means!

A related problem is that people often confuse statistical significance with the magnitude of the experimental effects. Statistical significance indicates the likelihood of a finding that the expected theoretical sampling error, whereas the magnitude of the effect is often a better way to judge the finding in practical or meaningful terms (Friedman, 1972).

Effect Sizes are a useful alternative for comparing large population parameters because effect size statistics compare differences in terms of degree, regardless of group sizes or scale differences (Cohan, 1988). However, this model provides no convenient theoretical sampling distribution to identify differences that exceed expectation. So while there is no convenient way to assess these differences, Cohan, 1988 describes a difference of .2 or less as “small”, .5 as “medium” and “large enough to be visible to the naked eye” and .8 as “grossly perceptible and therefore large” (Cohan, 1988, p. 23). Therefore we find this strategy compelling and will use Cohan’s D to compare differences in test scale scores based on the data mart results. The formula for this statistic appears below, where “M” = Match group, “N” = Non-Match group, “SD” = Standard Deviation and “n” = the group size:

D = (Mean M – Mean N) / (SD M+ SD n/2)

This formula is a ratio of the left hand side, which is the differences in the means, and the denominator which is the average pooled standard deviations of the groups. Cohan does not specify which standard deviation to use but says it should be “the standard deviation of either population (since they are assumed equal)” (Cohan, 1988, p. 20). One slight practical modification that we have made is that if the ratio of the standard deviation of the two groups differs by less than .95 we opted to use the pooled standard deviation, but otherwise we will use the standard deviation of the match group as this tends to be the more conservative estimate.

Table 1: Forwards Analysis, Grade 4 to Grade 6 to Grade 8 toGrade10, 2000 to 2007, Cascading Results

Statewide Mathematics, Matching (Stable Students) vs Non-Matching Students

Grade	-------- Matching --------			---- Non-Matching ----			----- Dif Test -----	
Range*	n	SSMean	% Prof	n	SSMean	% Prof	SSDif	Effect Size
4 to 6	34166	260.4	85.9	9093	235.0	66.2	25.5	.61
6 to 8	31028	257.0	82.7	13118	228.7	59.2	28.3	.68
8 to10	26232	260.2	85.2	15734	232.9	64.0	27.3	.65

* All score results and performance comparisons are made on the higher grade

Table 1 illustrates the effect of rolling up or cascading the records. For example, the table 1 results show that the matching group count declines each year from 34166 in grade 4 until the cohort reaches grade 10 and drops from 34166 to 26232. This loss of 7934 records occurs because the Grade 8 to 10 stage of the analysis includes only the records of those students who have remained in the state public education system from the time they took the grade 4 mathematics test in 2000 through to the time they took the grade 10 mathematics test in the spring of 2007. Meantime the non-matching group has grown steadily from 9093 in grade 4 to 15734 in grade 10. This is because this group is accumulating new non-matches which roll up from grade to grade while retaining any prior non-matches.

The results presented in Table 1 show that matching or stable students who have been in the system from grade 4 through to grade 10 perform much better (approximately 83% to 86% achieving Proficiency or higher) compared with the non-matching or incoming group. The non-matching group has noticeably lower performance, with approximately 59% to 66% reaching Proficiency or higher. Thus, the non-matching or incoming group is lagging more than 20 percentage points below the matching or stable group in terms of relative Proficiency. The scale score difference column marked "SSdif" in Table 1 also shows the effect. The mean difference between the incoming and stable group is a 26 to 28 point scale score gap across the comparisons. The effect sizes, which are based on the scale scores, show differences ranging from .61 to .68 across the three comparisons. These differences exceed one-half a standard deviation and therefore are "moderate" according to Cohan's criteria.

There is a potential source of bias affecting the non-matching group in Table 1. This bias is due to including students who submitted invalid test scores for the prior test administration. Normally invalid test scores result from students who are absent, submit blank test answers or are excused from testing due to inadequate familiarity with the English language. However, as noted earlier, in Connecticut, many Special Education students took out-of-level examinations from the year 2000 until the year 2005 and these students were marked as invalid for the on-grade examinations. This means that many of the "invalid" records actually reflected Special Education students who took the out-of-level tests instead of the on-grade tests plus the more conventional reasons for invalidity such as absenteeism. This is a potentially biasing influence because Special Education examinees, especially those permitted to take an out-of-level test, are more likely to do less well on the regular on-grade test the next year.

To avoid a bias from inclusion of these special education out-of-level test takers, we simply exclude all subjects who submitted invalid results on the previous test. We did this because it more accurately reflects the spirit and intent of our analysis, which is to examine the performance impact of mobility versus stability. If we were to include previously invalid records, this may bias the performance averages for the non-matching or mobile group average by lowering their average performance. This effect is observable by looking at the Special Education percentages including and excluding previously invalid records from the prior test. If the previously invalid examinees remain in the analysis as part of the non-matching or incoming subgroup, the percentage of Special Education examinees is much higher than if they are removed (See Figure 2, below).

Figure 2: Comparison of the Percents of Special Education Participation Including and Excluding Previously Invalid Test Records

Grade Range:	4 to 6	6 to 8	8 to10
Non Match SPED w PrevInval IN:	15.2 %	19.2 %	15.1 %
Non Match SPED w/o PreInval:	10.5 %	13.8 %	11.9 %
	------	------	------
Percent Difference:	4.7 %	5.4 %	3.2 %

Table 2 (below) is a revision of Table 1. Table 2 removes all of the students who were in a Connecticut public school from the previous tested grade and submitted invalid test results either due to illness, refusal to test, inadequate English language skills or because they were in Special Education and took the out-of-level testing program. While the matching record group results in Table 2 is identical to the matching record group in Table 1, the non-matching records reflecting the new incoming examinees without including the contribution of the students who had invalid records in the previous year but valid records in the second year. We believe removing the previously invalid examinees more legitimately represents the performance effects for new incoming students to Connecticut.

Table 2: Forwards Analysis, Grade 4 to Grade 6 to Grade 8 to Grade10, 2000 to 2007, Cascading Results (previous invalids excluded)

Statewide Mathematics, Matching (Stable Students) vs Non-Matching (Incoming Students)

Grade	-------- Matching --------			---- Non-Matching ----			----- Dif Test -----	
Range*	n	SSMean	% Prof	n	SSMean	% Prof	SSDif	Effect Size
4 to 6	34166	260.4	85.9	7908	240.7	71.1	19.7	.47
6 to 8	31028	257.0	82.7	11613	234.4	64.0	22.6	.54
8 to10	26232	260.2	85.2	14643	236.1	66.5	24.1	.58

* All score results and performance comparisons are made on the higher grade

Table 2 better shows how interstate mobility influences test performance, now that the previously invalid students are out of the picture. The n-count for the non-matching group has dropped from 9093 to 7908 in Grade 4, from 13,118 to 11,613 in grade 6, and from 15,734 to 14,643 in grade 10. Scale score differences now range from about 19.7 to 24.1 points and the percent Proficient for the two classifications differ by about 20%. The effect sizes are smaller and now range from between .47 to .58. This means that the score distributions of these two groups differ by about half of a standard deviation. These are still medium sized differences according to Cohan's criteria, but they are down about ten points from the analysis that included the previously invalid students in Table 1.

Table 3 shows how the background characteristics of the non-matching group look after extracting the previously invalid students. School Lunch participation ranges from between 32.4% to 34.3%, indicating that non-matching or incoming examinees are far more likely to be in the school lunch program. Therefore, this group tends to be from low-income families, a finding supported by others (Bourque, 2008). Whites with valid test scores range from 54.0% to 57.1% for the non-matching group. This indicates that minorities are more likely to be mobile than Whites. There is a slight bias favoring Males in the non-matching group, from 53.1 to 54.4, and a substantial difference in Special Education status from between 10.5% to 13.8%, indicating that Special Education students are more likely to be mobile. ELL participation ranges from between 5.3% to 7.5% also showing that ELL students are less likely to be stable.

Table 3: Background Analyses, Mathematics Scale Score Differences between Matching Student Records (Stable) vs. Non-Matching Student Records (Previously Invalid NOT Included), 2000 to 2007, Cascading Results

Grade Range*:	4 to 6	6 to 8	8 to10
Match Lunch:	21.4	20.8	17.1
Non-Match Lunch:	34.3	35.4	32.4
Match White:	73.6	76.2	78.3
Non-Match White:	54.0	55.4	57.1
Match Black:	11.3	11.0	10.2
Non-Match Black:	17.9	18.3	19.1
Match Hispanic:	10.4	9.7	8.7
Non-Match Hispanic:	19.7	20.7	18.9
Match Male:	50.1	49.8	48.9
Non-Match Male:	54.4	54.4	53.1
Match SPED:	8.1	8.7	7.2
Non-Match SPED:	10.5	13.8	11.9
Match ELL:	0.6	0.5	0.4
Non-Match ELL:	5.3	7.1	7.5

* All score results and performance comparisons are made on the higher grade

Table 4: Forwards Analysis, Grade 4 to Grade 6 to Grade 8 to Grade10, 2000 to 2007, Cascading Results (previous invalids excluded)

Statewide Reading, Matching (Stable Students) vs Non-Matching (Incoming Students)

Grade	-------- Matching --------			---- Non-Matching ----			----- Dif Test -----	
Range*	n	SSMean	% Prof	n	SSMean	% Prof	SSDif	Effect Size
4 to 6	33888	257.3	78.9	7892	237.4	63.0	19.9	.45
6 to 8	30854	261.5	82.5	11568	236.5	63.9	25.0	.52
8 to10	26176	252.8	87.5	14530	230.1	69.8	22.7	.53

* All score results and performance comparisons are made on the higher grade

We feel that these revised profiles more fairly represent how social disadvantage factors differ for students who come into the state as compared with those students who stay in Connecticut public schools. However, retaining the "previously invalid" examinees from the earlier analysis in Table 1 would have accentuated these differences due to the number of Special Education examinees who took the out-of-level tests and were reported as invalid for the on-grade tests. When these examinees returned to the regular examination schedule the following year they did not do as well as the regular on-grade examinees on the tests as a whole, inflating both the relative background differences as well as performance disparities between the groups.

Table 4 shows the results for Reading. This is a forwards-in-time analysis of the Reading tests from grade 4 (2000) through to grade 10 (2007), extracting the previously invalid students, so this table is comparable to the Table 2 for Mathematics. The effect size differences in Reading from grades 4 through 10 range from .45 to .53. These are somewhat lower than the effect sizes in Mathematics for Incoming students, but are otherwise similar to the Mathematics results (see Table 3).

As in the Mathematics analysis, the matching or stable group declines in n-count while the non-matching group increases due to the accumulating or cascading influence as the counts roll up over time. At each point of comparison the matching students exceed the non-matching students scale scores by about twenty to twenty-five points with about a .5 standard deviation difference measured by the Effect Size statistic. Thus, we conclude that the impact of student mobility is much the same in both reading and in mathematics.

Results by ERG

Connecticut has substantial income and family educational background differences across 169 school districts that also reflect observable differences in student test performance. Therefore, in order to compare test results more fairly, we cluster like districts into sub-groups that we call Educational Reference Groups or ERGs. ERGs are highly correlated to performance on our educational achievement batteries. ERG categorizations result from a linear combination of median family income, percent parents who are college graduates, percent professional occupation, percent poverty and percent single family home. There are nine distinct ERGs ranging from "A" (the school districts with the least social disadvantage factors) to "I" (the inner-city school districts with the highest social disadvantage factors). Although Connecticut no longer reports ERG's they will serve here as solely as a broad based mechanism to stratify social disadvantage differences among the districts.

This ERG analysis compares matched and cascading non-matched records in Mathematics and Reading by ERG, again after removing the records of the previously invalid students. This analysis is therefore a comparison of students who stayed in the same ERG from grade 4 to grade 10 to students who changed ERG. This would include new incoming students from out of state or from private school or home schooling, as well as new students moving from one ERG to another. This is a different look at mobility than was provided for the statewide analysis of incoming and stable students. Our interest in examining ERG performance differences for stable and incoming students is primarily just to see whether the statewide differences hold up when the analysis is more restricted.

We are presenting these results in a more compact form than we did with the earlier results. In the interests of saving space, we are presenting only the scale score differences for each ERG level, indicated on Tables 1 and 2 and 4 as the "SSDif." This statistic presents the matching (stable) group scale score subtracted from the non-matching (incoming) group scale score. The intent is to make it easier to comprehend differences and commonalities across the ERGs.

Table 5 (below) shows the mean differences between the matched group records and the non-matched group records by ERG in Mathematics. That is, the matched records reflect students who stayed within the same ERG from grade 4 through grade 10 while the non-matched records can be interpreted as including students who came into the ERG during that same range of grades, moved into the state, or transferred from a private school or home schooling situation to a Connecticut public school. The scale scores for non-matched records are subtracted them from the scale scores for the matched records to

Table 5: Mathematics Scale Score Differences between Matching / Non-Matching by ERG, 2000 to 2007, Cascading Results

Grade	ERG								
Range*	A	B	C	D	E	F	G	H	I
4 to 6	6.9	13.0	10.8	10.1	13.3	17.3	10.0	20.5	14.2
6 to 8	9.5	13.8	12.3	17.9	16.9	20.4	15.6	20.5	16.3
8 to 10	8.5	13.5	12.8	18.8	13.6	20.2	16.5	18.7	13.5

* All score results and performance comparisons are made on the higher grade

create differences. Thus, positive numbers reflect increases in scale score performance (See Table 5, above). Note that these results exclude previously invalid records.

All the difference scores are all positive in Table 5. This means that ERG level performance for matched records (stable group) from grades 4 to 6 and 6 to 8 and 8 to 10 exceed performance of non-matched (new or incoming group) records. Positive findings reflect the impact of mobility is apparent across all grades and all ERG levels. These results range from a low scale score difference of 6.9 in the grade 4 to 6 comparison of ERG A, to a high of 20.5 in the 4 to 6 and 6 to 8 comparisons of ERG H. Another issue is that the scores tend to go up from grade level to grade level within each ERG, but this is not always the case. This suggests that the educational impact of mobility tends to persist.

Although all the effects are positive, the impact of lower performance for incoming students is less in the most affluent ERG's (A–D) than in the less affluent ERG's (F-I). ERG H has the largest performance differences. ERG G is an anomaly in the sense that the general progression towards bigger performance differences does not happen in this case. ERG G is different from the other lower performing ERG's because the districts are generally small and rural. ERG G is describes as follows: "The 16 districts in this group have a lower median family income, education level and percentage in managerial or professional occupations as in Group F" (Research Bulletin, November 1996).

Interestingly, ERG I which includes the larger inner-cities and the biggest share of social disadvantages, shows only moderate differences and therefore is also inconsistent with the trend. This disparity may be the result of general lower performance of the comparison group of matching students rather than because of a lesser impact of mobility. Another concern about the ERG's is that they are based on census data from prior years' and therefore may not adequately reflect current circumstances. Nevertheless, the ERG's are a good although potentially imperfect indicator of a social advantage-disadvantage continuum.

Table 6 (below) shows the cascading mean differences between matching and non-matching records for the Reading test by ERG. Again, the entries are all positive differences, just as they were in Table 5 for Mathematics. This also indicates that incoming students do less well than stable students in Reading across ERG's. The differences for Reading by ERG are very similar to the differences found in the Mathematics analysis by ERG. As in the case of the Mathematics analysis, we conclude that the impact of student mobility in Reading is also common across all the ERG levels, although the problem seem less prominent in the ERGs where families are wealthier and better educated (e.g.: ERG's A-D) than in the less affluent ERGs (e.g.: ERG's F-I).

Once again, ERG H in Reading has the largest differences and as in the previous Mathematics analysis. We also believe that the ERG I differences are an artifact of the lower performance by the matching or stable group. Notably, ERG E has decreased the gaps between stable and incoming examinees in the 4-6 grade range and in the 8-10 grade range. It would be interesting to know more about this. ERG E consists largely of small districts and is the smallest total number of students in all the ERG's. As in the case with

ERG G in Table 5, smaller numbers of subjects could result in greater sampling fluctuations. It also may be that smaller schools mean that fewer students are allowed to escape the vigilance of teachers and therefore some of these differences get eradicated due to more attention being paid to lower performing students generally.

Table 6: Reading Scale Score Differences between Matching / Non-Matching by ERG, 2000 to 2007, Cascading Results

Grade	ERG								
Range*	A	B	C	D	E	F	G	H	I
4 to 6	7.7	12.4	10.4	13.1	9.2	15.3	10.1	19.7	13.6
6 to 8	8.8	15.4	13.0	19.7	15.5	22.2	17.7	23.9	19.4
8 to 10	8.6	12.4	15.5	15.3	9.5	20.5	18.8	19.8	13.6

* All score results and performance comparisons are made on the higher grade

We conclude from looking at these ERG results that student mobility is an important factor in performance on academic achievement tests at the ERG level. We also recognize that the effect of mobility seem somewhat less prominent in the more wealthy ERGs (e.g.: A-D) than in the less affluent ERGs (e.g.: F-I). The smaller ERG's E and G tend to show some anomalous results, achieving better results than might be expected.

This finding may be due to the impact of being educated in a smaller school, where teachers may have the opportunity to notice struggling students and work with them to ameliorate some of the performance problems. On the other hand, multiple social disadvantage factors tend to haunt the mobility group and there may be in fact a tipping point where the shere number of mobile students can cause a disruption factor that Rumsburger refers to as "functioning in a setting of pervasive chaos" (3003). Thus, in the lowest performing districts, being stable and staying in the same ERG have have relatively less benefit than might be the case in ERGs F through H.

Results by District

The intent of this district level analysis is to look more closely at the broad effects of test score differences for matching and non-matching groups. So far, we have shown that moves into the public education system seem to result in lower test performance statewide and at the ERG levels, although the impact is less in the wealthier ERG's. However, a natural question arises as to whether this finding would also be true at the district level. For example, a move from a different state or private school or even from one ERG to another could be quite different from a move from one school district to another. In this next phase, we are going to look at the effect size differences for math and then reading scale scores by individual school districts in Connecticut. Thus, we are looking at whether moves into a district, either from another district or from out of state or from private school or even from home schooling, will demonstrate a performance difference at the district level.

In the interests of space and time, this analysis is limited to a comparison of grade 4 of 2000 to grade 6 in 2002. It also seemed unreasonable to include all the districts in Connecticut, simply because forty districts are so small that they many not be meaningfully compared with districts that have a substantially larger number of students. Therefore, we arbitrarily limited our investigation to 126 districts that met the following criteria: A match group size of 50 or more students and a non-matching group of at least 10 students. This leaves us with 126 school districts out of 166.

Using this reduced collection of 126 districts we found that in mathematics, the average effect size was .34. While this shows that there continues to be a performance difference favoring stable or matching groups over incoming or non-matching groups, this is a substantial reduction from the statewide findings in Table 2, which showed an average effect size of .47.

Table 7 (below) shows the overall effect sizes, averages and n-counts for the entire population of 166 districts. This analysis shows a downward shift in group sizes across the different grade level groupings and also a reduction in performance averages as compared to the state level analysis in mathematics (shown in Table 2). Thus, the number of students in the matching group is down from 34116 for the state to 30245 for the district sample. More students are in the non-matching group because more students move from one district to another than from one state to another. The non-matching group increases from 7908 for the state to 9651 for the sample districts (see Table 2). The means for both the matching and non-matching groups are both up substantially (by 5 points for the matching sample districts and 11 points for the non-matching sample districts).

Table 7: Forwards Analysis, Grade 4 to Grade 6 (previous invalids excluded)

Statewide Mathematics, Matching (Stable Students) vs Non-Matching (Incoming Students)

	Grade	-- Matching --		-- Non-Match --		- Dif Test -
	Range*	n	SSMean	n	SSMean	Effect Size
Total State	4 to 6	34166	260.4	7908	240.7	.47
166 Districts	4 to 6	31338	252.2	10582	252.2	.31
126 Districts	4 to 6	30245	265.1	9651	252.4	.34

* All score results and performance comparisons are made on the higher grade

Table 8 shows the percentages of districts for each .10 in effect size from 0 to 100. These Mathematics score performance differences for grades 4 (in 2000) to 6 (in 2002) show that 87% of all the districts have an effect size difference greater than .10, 75%

have differences greater than .20, 56% have differences greater than .30, and 43% have differences greater than .40.

Table 8: Math Scale Score Effect Size Differences between Matching / Non-Matching by District, Grade 4 (2000) to Grade 6 (2002)

Cell Percent and N’s	Cumulative	Decrementive
.10 or less: 12.7% (16)	12.7% (16)	100.0% (126)
.10 to .20: 11.9% (15)	24.6% (31)	87.3% (110)
.20 to .30: 19.8% (25)	44.4% (56)	75.4% (95)
.30 to .40: 12.7% (16)	57.1% (72)	55.6% (70)
.40 to .50: 20.6% (26)	77.8% (98)	42.9% (54)
.50 to .60: 12.7% (16)	90.5% (114)	22.2% (28)
.60 to .70: 4.8% (6)	95.2% (120)	9.5% (12)
.70 to .80: 3.2% (4)	98.4% (124)	4.8% (6)
.80 to .90: 1.6% (2)	100.0% (126)	1.6% (2)

These mathematics results in Table 8 support the conclusion that performance differences between matching (stable) and non-matching (incoming) groups continue to favor the stable groups at the district level. They also show that the effects are common such that nearly all districts experience this to some degree. However, there remains a question about why there is a decline of the effects from the state level to the district level. For instance, this may be due to mixing students with substantial life changes—such as moving from one state to another—with more local changes, such as a move from one district to another. These differences could therefore reflect greater or lesser degrees of personal disruption. If the change from state to state has a big impact on performance and the change from district to district has a milder effect, combining these two groups in the district analysis might appear to reduce the overall impact of transience.

Bourque (2008) proposes one interesting possible explanation for these reductions in the effects of transience. She contends that very high mobility rates at least in a school setting lowers performance both of the mobile students as well as the stable students. Bourque hypothesizes that this may be due to higher levels of systemic disruption, instructional repetition, narrowing of the curriculum and other factors. Her contention is that these effects may conspire to reduce the performance for the stable students as well as the mobile students in affected schools. Thus, schools that exceed a certain tipping point with respect to overall school mobility levels might find themselves in a state of pervasive disruption that affects learning of all students, stable as well as incoming.

Bourque identified schools in the Chelsea area of Massachusetts with three levels of mobility: Level 1, 9.9 % or less, Level 2, 10 to 19.9 % and Level 3, 20% or more. Although she used different criteria for tracking mobility, we were interested in seeing if our results support her findings. We hypothesize that our effect sizes would change across the rank ordering of the percent of non-matching students such that low mobility districts would have lower effect sizes and as mobility increases the effect sizes would also increase up to a point, but that this progression would begin to diminish as the degree of mobility reaches the highest levels.

Our data reflect an indirect approximation of mobility by looking across from grade 4 CMT administration in 2000 to grade 6 administration in 2002, hence we have two years of mobility to consider whereas Bourque was looking from year to year. Bourque was also looking at schools, whereas we are looking at districts. Nevertheless, after examining our distributions of the percent of non-matching examinees from 2000 to 2006 we found 5 levels of non-matching percentages that appear appropriate, and that these 5 levels roughly correspond to doubling Bourque's levels and expanding to add two additional levels (see Table 9).

Table 9 Math District Effect Size ordered by Percent of Non-Matching students

Level	NonMch%	N	Ave ES	SD ES
1	< 20	19	.31	2.1
2	20-29.9	50	.33	2.7
3	30-39.9	35	.35	2.3
4	40-49.9	12	.31	2.5
5	50 >	9	.39	10.8

Table 9 shows an increase in the average effect size, although modest, is apparent from level 1 to level 3 and then begins to decline in level 4, a pattern which is consistent with Bourque's contention. However, in level 5 the average effect size increases dramatically, as does the standard deviation of the effect sizes. Therefore, we have some confirmation of Bourque's findings for levels 1 through 4 but also evidence of a divergence at level 5. Level 5 contains many of Connecticut's larger districts as well as some of our lowest performing districts.

We therefore conclude that the pattern predicted by Bourque is might be said to be weakly supported if we consider the first four levels independently. However, if we include level 5 in the picture this trend breaks down. This may be due to the multiplicity of issues facing larger inner-city districts and therefore normal trends may break down for these cases. On the other hand, we could also say that despite the reduced performance of stable examinees for these higher levels of mobility, test score differentials continue to increase more or less consistently.

The reading district level analysis also reflects a shift in group sizes and performance averages consistent with the changes in the mathematics analysis (see Table 10). Thus, the number of students in the matching group is down from 33888 for the state to 30017 for the district sample. Again, more students are in the non-matching group for the 126 districts as compared to the state. This is because more students move from one district to another than from one state to another. The non-matching group increases from 7992 for the state to 9643 for the sample districts. Also, the means for both the matching and non-matching groups are both up substantially from 257.3 for the state to 262.6 for the matching districts and from 237.4 to 250.5 for the non-matching groups. For reference, we have also included the overall effect sizes, averages and n-counts for the entire population of 166 districts.

Table 10: Forwards Analysis, Grade 4 to Grade 6 (previous invalids excluded)

Statewide Reading, Matching (Stable Students) vs Non-Matching (Incoming Students)

	Grade	-- Matching --		-- Non-Match --		- Dif Test -
	Range*	n	SSMean	n	SSMean	Effect Size
Total State	4 to <u>6</u>	33888	257.3	7992	237.4	.45
166 Districts	4 to <u>6</u>	31087	262.1	10571	251.2	.28
126 Districts	4 to <u>6</u>	30017	262.6	9643	250.5	.30

* All score results and performance comparisons are made on the higher grade

As was the case for mathematics, a similar pattern is apparent for the district sample and the state in reading (see Table 10). The matching group has a mean difference of 5 points and the sample has decreased from 33888 to 30017 a difference of almost four thousand, while the non-matching group has gained 1,651 cases (9643-7992). The performance has also increase from 237.4 for the state to 250.5 for the district sample (a gain of 13 points). The overall effect size has declined by .15. It is also important to note that these reading effect sizes are smaller than the math effect sizes, a pattern that is consistent throughout all of our findings.

Table 11, below, examines the reading effect sizes by district. The Reading distributions of the effect sizes in Table 11 shows that 87% of the districts have an effect size greater than .10, 69% have differences greater than .20, 51% have differences greater than .30 and 29% have differences greater than .40. These reading are in the same direction and therefore favor stable examinees over incoming examinees, but interestingly results are less dramatic for reading than for mathematics. The total average Effect Sizes for reading are down compared to mathematics so the effects appear to be less deleterious for incoming students in reading than in math.

This finding may be due to differences in the teach-ability of the two content areas. Thus, perhaps Mathematics skills are more discrete and perhaps therefore more likely to be affected by skills gaps, whereas reading skills may be more innate or at least less sensitive to particular skill gaps or otherwise less affected by changes in schools and teachers. Nevertheless, Reading results also support the conclusion that performance differences between matching (stable) and non-matching (incoming) groups favor stability at the district level.

Table 11: Reading Scale Score Effect Size Differences Between Matching / Non-Matching by District, Grade 4 (2000) to Grade 6 (2002)

Cell Percent and N's		Cumulative	Decrementive
.10 or less:	13.5% (17)	13.5% (17)	100.0% (126)
.10 to .20:	17.5% (22)	31.0% (39)	86.5% (109)
.20 to .30:	18.3% (23)	49.2% (62)	69.0% (87)
.30 to .40:	22.2% (28)	71.4% (90)	50.8% (64)
.40 to .50:	14.3% (18)	85.7% (108)	28.6% (36)
.50 to .60:	7.9% (10)	93.7% (118)	14.3% (18)
.60 to .70:	1.6% (2)	95.2% (120)	6.3% (8)
.70 to .80:	4.0% (5)	99.2% (125)	4.8% (6)
.80 to .90:	.8% (1)	100.0% (126)	0.8% (1)
.90 or more	0.0 (0)	100.0% (126)	0.0% (0)

Concluding Observations

Our research question was "Does test performance differ significantly for examinees who have taken the CMT and CAPT sequentially from grade 4 through to grade 10, as compared to those examinees who have come into the testing program at some time after the grade 4 examinations?"

We have shown a series of analyses using a longitudinal data mart which describe differences in the performance of students with matching (stable) and non-matching status (incoming new students). We followed these results sequentially in three stages or cycles beginning from grade 4 of 2000 through to grade 10 in 2007. We interpreted matching records as the stable group and non-matching records as new incoming students. We analyzed this data using a longitudinal data mart. We have shown that students who have matching records (indicating greater stability) perform better in mathematics and reading over time as compared to new incoming students at the state, ERG and district levels. This also means that new incoming students have lower performance at the state, ERG and district levels. This finding may reflect the adage that children do not respond well to change, but more analyses would be required to demonstrate this conclusively.

Statewide, the percentage of incoming students meeting Proficiency showed about 15 to 20 percentage points below the stable group. Matching students that have been stable in the same system over time in Connecticut meet or exceed Proficiency 82.7% to 85.9% of the time in mathematics and 79.6% to 85.9% of the time in reading. Their incoming student counterparts meet proficiency between 64.0% and 71.1% in mathematics and 63.0% to 69.8% in reading. Clearly, these are substantial differences.

Our ERG analyses show that this mobility effect is relatively consistent apart from the very highest ERG category (ERG A—the districts with the wealthiest and best-educated parents). The scores of students in ERG A are generally much higher than the scores in the other ERG's so the potential for growth may be somewhat limited. Nevertheless, these differences show that, regardless of other ERG related factors, student mobility continues to have an important impact on test performance. While additional efforts to improve the scores of matching students in ERG A may quickly hit a point of diminishing returns, it is unlikely to be the case with most of the other ERGs.

At the district level, these performance differences between stable and incoming students are common to nearly all the districts, at least for grades 4 to 6 in Mathematics and Reading. The impact of mobility consistently appears to relate to lower student performance at the district level, but we found that the impact is greater in Mathematics than it is in Reading. This may be due to differences in the nature of the two content areas. Perhaps mathematics skills are more discrete, whereas reading skills may be less compartmentalized, and to that extent less affected by changes in schools and teachers.

We must exercise caution in interpreting these results. We need to be aware of the potential that this mobility factor may indeed be a stand-in for other underlying factors that are actually causing the performance discrepancies—such as the relative stability of the family. It is fair to say that for whatever reason students who stay in the same system from the earliest grades tend to do better on our statewide achievement tests as compared with students who are in transition. What we cannot say for sure is whether the general decline in performance for students in transition is due to the transition itself or whether transition is in itself a proxy variable that reflects other unobserved factors that actually cause the lower performance. It could be that stress caused to the student by living in a family that relocates which in turn may reduce test performance. We have also shown that other social disadvantage factors are more prominent in these groups.

Regardless of the true cause of this effect, we concur with Bourque (1980) that incoming status may be an important factor or identifier for students at risk. The good news about this is that we can use this information to inform us about these students in more detail. We can monitor the prior and current academic performance of incoming students more carefully and determine if there is a downturn in the new setting.

We therefore conclude that it would be good policy to at least look at incoming student performance on prior test scores and to assess current levels of progress to determine whether a slump is occurring. Should a downturn be evidenced one simple intervention strategy is to offer additional academic programs and special monitoring for new incoming students who appear to be in trouble. Another approach would be to implement a blanket policy of content review for all new incoming students, regardless of current or past test performance or academic functioning. The advantage of using the blanket approach is that it would help de-stigmatize the intervention. Our concern here is that if this negative performance effect is due in part to social adjustment factors then identifying only certain students for an academic intervention might increase the negative impact.

Following proposals recommended by Rumsberger (2002), we also concur that it may also be useful to employ methods to assist with social adjustment in order to help new students acclimatize to new school circumstances. There is a potential for disruptive family situations or events such as divorce, job displacement for parents, loss of standard of living and other disruptive factors that may influence new incoming students disproportionately.

Schools might consider teaming up new students with volunteer students who might be willing to help introduce them to the new school. Some schools invite new students to spend a day at the school with their parents, getting to know where things are and perhaps meeting other students. Another strategy might be to assign a teacher to monitor or advisor for incoming students in order to assess how the transition is going for the child from an academic and or a social prospective.

Following Bourque (2008), we also concur that it also may be valuable to inform parents about the possible concerns of changing schools and propose ways to mitigate these effects. One approach might be to put off family moves until summer. Perhaps the hardest hit school districts with high social disadvantage factors might consider transportation and other supports for students who are moving, so that they can potentially remain in the same school even after a family move, a model that Bourque introduced in Chelsea.

We note in conclusion that the change in Connecticut's testing program from fall to spring testing may reduce the differential performance between stable and incoming new students in the future. This could occur because parents with young school aged children may tend to move between June and September so that this might reduce the school adjustment factor for them. However, when the CMT was a fall testing program, these students would not only experience a school and social adjustment, but would also be required to take the CMT at the beginning of the school year. However, the new spring testing program could ameliorate this effect, because new students would have had more to acclimatize to the new circumstances and may do better on the tests.

References:

Becker, L. (2007). Effect Size (ES). Http://web.uccs.edu/lbecker/Psy590.

Bourque, Mary D. (2008): "The impact of student mobility on urban school districts." Unpublished dissertation, Boston University School of Education.

Cohan, Jacob (1988). Statistical power analysis for the behavioral sciences, Revised Edition. Orlando Fla: Academic Press, Inc.

DePascale, Charles A. (2006). Measuring Growth with the MCAS tests: A consideration of vertical scales and standards. A paper presented at the National Center for the Improvement of Educational Assessment.

Drewek, K, (2005). Data warehousing: Similarities and differences of Inmon and Kimball. Article URL: http://www.b-eye-network.com/view/743.

Friedman, Herbert (1972): Introduction to statistics." New York: Random House.

Kelly, S and Monczunski, L, (2007). Overcoming the volatility of in school-level gain scores: A new approach to identifying value added with cross-sectional data. Educational Researcher, vol. 36 No. 5 pp. 279-287.

Glass, Gene (2004). Introduction to quant methods. http://glass.ed.asu.edu/gene/.

Glass, Gene (2000). Meta-Analysis at 25. http://glass.ed.asu.edu/gene/.

Henkel, R.E. (1976). Tests of significance. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-004. Beverly Hills and London: Sage Pubns.

Inmon, W. H. (1997). Data warehouse and data mining. Article URL: www.taborcomunications.com/dsstar/97107/1000001.html.

Inmon, W. H. (1999). Data mart does not equal data warehouse. Article URL: DMReview.com.

Kolen, M.J. & Brennan, R.L. (2004, 2nd Ed.). Test equating, scaling, and linking: Methods and practices. NY: Springer.

Linn, Robert L. (2008). Validation and uses and interpretations of state assessments. National Center for Research on Evaluation, Standards, and Student Testing. Unpublished paper prepared for Council of Chief State School Officers.

Linn, R.L. & Baker, E. L. (1993; Winter). Comparing results from disparate assessments. The CRESS Line, pp 1.2. Los Angeles: National Center for Research on Evaluation, Standards, & Student Testing.

Lissitz, R. W. & Huynh Huynh (2003). Vertical Equating for State Assessments: Issues and solutions in determination of adequate yearly progress and school accountability. Practical Assessment, Research and Evaluation, PARE online.net.

Mislevy, R.J. (1992). Linking educational assessments: Concepts, issues methods, and prospects. Princeton, NJ: Educational Testing Service.

Mooney, R. and Beaudin, B (2007). How to build and Use Assessment Data Marts for Policy Decision-Makers. NERA, October.

Patz, Richard J. (2007) Vertical scaling in standards-based educational assessment and accountability systems. Pre-publication copy: Council of Chief State School Officers, January 2007.

Popham, W. J. (1978), Criterion-Referenced Measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Prowda, P. and Thompson, J. (1996), Research Bulletin. Hartford CT: Connecticut State Department of Education.

Rumsberger, R.W. (2002). Student mobility and academic achievement. ERIC Digest, University of Illinois, June, 2002, EDO-PS-021.

Streifer, Philip (2002): Data-driven decision-making: What is knowable for school improvement? NCES Summer Data Conference in Washington DC, 24-26 July, 2002.

APPENDIX A: Statewide Mathematics Results Grades 4 to 6, 6 to 8 and 8 to 10, 2000-2007

```
TITLE     MYPCT  FIXED REF_GRP MYSUBSET
G4 to G6: 99.999 NO    STATE   TOTAL


CENSUS GROUP   TOTAL N:  44311      TOTAL INVALID:  2834


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     04     2000   CENSUS   0.4   1.3   3.3   6.9        12.4  18.5  19.4  12.9        6.1   1.9   0.0   0.7


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     04     2000   CENSUS   41477 250.1  44.6     82.0     60.2       8.0      10.0      21.8      42.4      17.8


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     04     2000   CENSUS   41477  50.8  49.2  70.1  12.4  11.1   2.2    0.4   3.8   7.9  26.1   1.4


CENSUS GROUP   TOTAL N:  45153      TOTAL INVALID:  1879


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     06     2002   CENSUS   0.3   1.0   2.9   6.2        12.1  17.5  21.8  18.9       10.2   3.7   0.4   0.4


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     06     2002   CENSUS   43259 255.1  44.8     81.8     61.0       8.1      10.1      20.7      40.7      20.4


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     06     2002   CENSUS   43259  51.1  48.9  68.9  12.7  13.0   2.6    0.3   2.4   9.6  24.8   1.9


Match Group


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     06     2002   MATCH    0.1   0.5   1.9   5.0        11.0  17.4  22.9  20.5       11.4   4.2   0.5   0.5


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     06     2002   MATCH    34166 260.4  41.9     85.9     65.7       5.3       8.8      20.2      43.1      22.6


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     06     2002   MATCH    34166  50.1  49.9  73.6  11.3  10.4   2.4    0.3   1.9   8.1  21.4   0.6


Non-Match Group PREV INVALD OUT:  1185


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     06     2002   NOMTCH   0.7   1.9   5.2   9.5        16.2  18.5  19.3  14.4        6.4   2.3   0.3   0.2


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     06     2002   NOMTCH    7908 240.7  47.0     71.1     48.0      14.4      14.5      23.1      34.6      13.4


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     06     2002   NOMTCH    7908  54.4  45.6  54.0  17.9  19.7   3.3    0.5   4.6  10.5  34.3   5.3


Statistical Tests


MCH_TEST     DUPSOUT NONMCH_AVE MCH_AVE DIF_AVE NONMCH_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***      24      240.7   260.4    19.7      47.0   41.9   -5.1     89.2       0.0       .4707 .4460 .4954


GRP_TEST_1            PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              250.1   260.4    10.3     44.6   41.9   -2.7     93.9      43.3       .2383 .2239 .2527


GRP_TEST_2              CURR_AVE MCH_AVE DIF_AVE CURR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***               255.1   260.4     5.4    44.8   41.9   -2.9     93.6      43.3       .1235 .1093 .1377


GAIN_TEST             PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              253.2   260.4     7.2     42.9   41.9   -1.0     97.7      42.1       .1711 .1561 .1861


TITLE     MYPCT  FIXED REF_GRP MYSUBSET
G6 to G8: 99.999 NO    STATE   TOTAL


CENSUS GROUP   TOTAL N:  45315      TOTAL INVALID:  1168
```

```
TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     08     2004   CENSUS   0.1   1.6   5.1   9.3        13.4  16.8  19.6  17.2        8.5   2.1   0.5   0.1


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     08     2004   CENSUS   44146 248.6  46.1     75.7     55.7      12.2      12.1      20.0      36.2      19.5


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     08     2004   CENSUS   44146  51.3  48.7  69.4  13.2  13.5   2.8    0.3   0.7  11.8  26.0   2.7


Match Group


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     08     2004   MATCH    0.0   0.3   2.6   7.2        12.3  17.1  21.6  19.6       10.1   2.5   0.6   0.1


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     08     2004   MATCH    31028 257.0  41.7     82.7     62.7       6.8      10.5      20.0      40.0      22.7


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     08     2004   MATCH    31028  49.8  50.2  76.2  11.0   9.7   2.4    0.3   0.5   8.7  20.8   0.5


Non-Match Group PREV INVALD OUT:  1505


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     08     2004   NOMTCH   0.2   2.8   8.6  13.8        16.4  17.0  16.4  12.8        5.3   1.5   0.3   0.0


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     08     2004   NOMTCH   11613 234.4  47.6     64.0     42.8      19.7      16.3      21.2      29.7      13.1


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     08     2004   NOMTCH   11613  54.4  45.6  55.4  18.3  20.7   4.0    0.4   1.2  13.8  35.4   7.1


Statistical Tests


MCH_TEST     DUPSOUT NONMCH_AVE MCH_AVE DIF_AVE NONMCH_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***       2      234.4   257.0    22.6      47.6   41.7   -5.8     87.7       0.0       .5417 .5201 .5633


GRP_TEST_1            PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              255.1   257.0     1.9     44.8   41.7   -3.0     93.2      43.2       .0446 .0300 .0592


GRP_TEST_2              CURR_AVE MCH_AVE DIF_AVE CURR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***               248.6   257.0     8.4    46.1   41.7   -4.4     90.5      43.9       .1918 .1772 .2064


GAIN_TEST             PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              261.9   257.0    -4.9     41.3   41.7    0.4    101.0      41.6       -.117 -.133 -.101


TITLE      MYPCT  FIXED REF_GRP MYSUBSET
G8 to G10: 99.999 NO    STATE   TOTAL


CENSUS GROUP   TOTAL N:  44311      TOTAL INVALID:  2344


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     10     2007   CENSUS   2.2   2.0   2.0   6.1        12.7  18.3  20.8  15.5        7.5   3.7   0.5   0.2


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     10     2007   CENSUS   41966 250.0  47.5     77.2     45.2      10.3      12.4      32.0      25.0      20.3


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     10     2007   CENSUS   41966  50.6  49.4  70.0  13.5  13.0   3.3    0.3   0.0  10.1  23.4   3.1


Match Group


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     10     2007   MATCH    0.7   0.8   1.1   3.9        10.5  18.0  23.2  18.2        9.3   4.5   0.7   0.2


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     10     2007   MATCH    26232 260.2  41.8     85.2     53.4       5.1       9.7      31.8      28.8      24.6


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     10     2007   MATCH    26232  48.9  51.1  78.3  10.2   8.7   2.5    0.2   0.0   7.2  17.1   0.4
```

```
Non-Match Group PREV INVALD OUT:  1091


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
MA     10     2007   NOMTCH   3.8   3.4   3.2   9.1        16.4  19.2  17.6  11.4        4.9   2.5   0.3   0.1


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTMAPRO PCTMAGOL PCTMASTD1 PCTMASTD2 PCTMASTD3 PCTMASTD4 PCTMASTD5
MA     10     2007   NOMTCH   14643 236.1  50.0     66.5     33.4      16.8      16.7      33.2      19.5      13.9


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
MA     10     2007   NOMTCH   14643  53.1  46.9  57.1  19.1  18.9   4.6    0.3   0.0  11.9  32.4   7.5

Statistical Tests


MCH_TEST     DUPSOUT NONMCH_AVE MCH_AVE DIF_AVE NONMCH_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***       0      236.1   260.2    24.1      50.0   41.8   -8.2     83.5       0.0       .5758 .5552 .5964


GRP_TEST_1            PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              248.6   260.2    11.6     46.1   41.8   -4.3     90.7      43.9       .2640 .2487 .2793


GRP_TEST_2              CURR_AVE MCH_AVE DIF_AVE CURR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***               250.0   260.2    10.2    47.5   41.8   -5.7     88.0      44.7       .2290 .2135 .2445


GAIN_TEST             PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              259.7   260.2     0.5     40.0   41.8    1.8    104.5      41.4       .0114 -.006 .0285
```

APPENDIX B: Statewide Reading Results Grades 4 to 6, 6 to 8 and 8 to 10, 2000-2007

```
TITLE     MYPCT  FIXED REF_GRP MYSUBSET
G4 to G6: 99.999 NO    STATE   TOTAL


CENSUS GROUP   TOTAL N:  44311      TOTAL INVALID:  3235


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     04     2000   CENSUS   0.1   0.8   4.2   9.5        13.9  18.5  20.0  17.3        9.3   3.7   0.5   0.2


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     04     2000   CENSUS   41076 249.7  45.0     70.7     56.9      19.9       9.4      13.8      35.4      21.5


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     04     2000   CENSUS   41076  50.6  49.4  70.3  12.3  11.0   2.2    0.5   3.7   7.2  26.0   1.3


CENSUS GROUP   TOTAL N:  45153      TOTAL INVALID:  1974


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     06     2002   CENSUS   0.4   2.0   4.3   8.0        12.6  16.6  20.2  18.2        9.5   2.8   0.7   0.2


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     06     2002   CENSUS   43163 251.5  47.3     74.2     64.1      17.8       8.0      10.0      45.3      18.9


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     06     2002   CENSUS   43163  51.0  49.0  68.8  12.8  13.0   2.6    0.4   2.5   9.4  24.8   1.9


Match Group


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     06     2002   MATCH    0.1   1.0   3.0   6.7        11.8  16.8  21.4  19.8       10.4   3.2   0.8   0.2


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     06     2002   MATCH    33888 257.3  44.2     78.9     69.0      13.6       7.5       9.9      48.1      20.9


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     06     2002   MATCH    33888  49.9  50.1  73.7  11.4  10.3   2.4    0.3   1.9   7.5  21.2   0.6


Non-Match Group PREV INVALD OUT:  1383


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     06     2002   NOMTCH   0.9   3.9   7.2  11.3        15.2  17.2  17.3  14.0        7.0   1.6   0.4   0.1


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     06     2002   NOMTCH    7892 237.4  50.0     63.0     52.0      27.4       9.6      11.0      38.7      13.3


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     06     2002   NOMTCH    7892  54.5  45.5  54.1  18.0  19.6   3.3    0.5   4.5  10.3  34.2   5.3


Statistical Tests


MCH_TEST     DUPSOUT NONMCH_AVE MCH_AVE DIF_AVE NONMCH_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***      24      237.4   257.3    19.9      50.0   44.2   -5.8     88.5       0.0       .4511 .4264 .4758


GRP_TEST_1            PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              249.7   257.3     7.6     45.0   44.2   -0.8     98.2      44.6       .1706 .1562 .1850


GRP_TEST_2              CURR_AVE MCH_AVE DIF_AVE CURR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***               251.5   257.3     5.8    47.3   44.2   -3.1     93.5      45.7       .1268 .1126 .1410


GAIN_TEST             PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              252.6   257.3     4.8     43.8   44.2    0.4    100.9      44.1       .1079 .0929 .1230


TITLE     MYPCT  FIXED REF_GRP MYSUBSET
G6 to G8: 99.999 NO    STATE   TOTAL


CENSUS GROUP   TOTAL N:  45315      TOTAL INVALID:  1065


TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 LOW   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
```

```
RD     08     2004   CENSUS   0.8   2.4   4.3   7.9        12.4  16.9  18.1  16.1       10.4   4.1   1.4   1.2

TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     08     2004   CENSUS   44249 251.7  52.8     75.2     64.9      17.3       7.5      10.3      41.2      23.7

TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     08     2004   CENSUS   44249  51.3  48.7  69.4  13.3  13.6   2.8    0.3   0.6  11.8  26.0   2.6

Match Group

TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     08     2004   MATCH    0.2   0.8   2.5   6.1        11.3  17.3  19.8  18.4       12.0   5.0   1.6   1.4

TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     08     2004   MATCH    30854 261.5  47.9     82.5     72.5      10.9       6.6      10.0      44.9      27.6

TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     08     2004   MATCH    30854  49.5  50.5  76.2  11.0   9.7   2.4    0.3   0.5   8.1  20.8   0.5

Non-Match Group PREV INVALD OUT:  1827

TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     08     2004   NOMTCH   1.5   4.3   6.9  11.1        15.3  17.2  15.5  12.1        7.5   2.5   1.0   0.8

TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     08     2004   NOMTCH   11568 236.5  54.6     63.9     52.4      26.5       9.7      11.5      35.7      16.7

TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     08     2004   NOMTCH   11568  54.4  45.6  55.6  18.3  20.6   3.9    0.4   1.1  13.6  35.4   7.1

Statistical Tests

MCH_TEST     DUPSOUT NONMCH_AVE MCH_AVE DIF_AVE NONMCH_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***       2      236.5   261.5    25.0      54.6   47.9   -6.7     87.7       0.0       .5219 .5002 .5436

GRP_TEST_1            PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              251.5   261.5    10.0     47.3   47.9    0.6    101.2      47.6       .2092 .1945 .2239

GRP_TEST_2              CURR_AVE MCH_AVE DIF_AVE CURR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***               251.7   261.5     9.8    52.8   47.9   -5.0     90.6      50.3       .1936 .1790 .2082

GAIN_TEST             PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              258.7   261.5     2.8     43.5   47.9    4.4    110.1      46.8       .0595 .0437 .0752

TITLE      MYPCT  FIXED REF_GRP MYSUBSET
G8 to G10: 99.999 NO    STATE   TOTAL

CENSUS GROUP   TOTAL N:  44311      TOTAL INVALID:  2254

TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     10     2007   CENSUS   0.7   1.9   5.1   9.2        15.8  19.3  20.7  14.3        6.3   3.1   0.7   0.2

TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     10     2007   CENSUS   42056 242.9  47.4     79.7     45.6       7.5      12.7      34.2      25.6      19.9

TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     10     2007   CENSUS   42056  50.6  49.4  69.9  13.7  13.0   3.2    0.3   0.0  10.2  23.4   3.0

Match Group

TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     10     2007   MATCH    0.2   0.7   2.7   6.4        14.5  20.2  23.7  16.9        7.7   3.7   0.8   0.3

TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     10     2007   MATCH    26176 252.8  42.8     87.5     53.4       3.4       9.2      34.1      29.5      23.9

TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     10     2007   MATCH    26176  48.8  51.2  78.3  10.3   8.7   2.5    0.2   0.0   6.7  17.1   0.4

Non-Match Group PREV INVALD OUT:  1350
```

```
TEST   GRADE  TST_YR GROUP  G_100 G_125 G_150 G_175 L0W   G_200 G_225 G_250 G_275 HI   G_300 G_325 G_350 G_375
RD     10     2007   NOMTCH   1.3   3.2   8.1  13.3        18.1  18.4  16.6  10.6        4.5   2.1   0.5   0.2


TEST   GRADE  TST_YR GROUP   NCOUNT   AVE    SD PCTRDPRO PCTRDGOL PCTRDSTD1 PCTRDSTD2 PCTRDSTD3 PCTRDSTD4 PCTRDSTD5
RD     10     2007   NOMTCH   14530 230.1  49.2     69.8     34.7      12.4      17.8      35.1      20.4      14.3


TEST   GRADE  TST_YR GROUP   NCOUNT  MALE   FEM WHITE BLACK  HISP ASIAN INDIAN OTHER  SPED LUNCH   ELL
RD     10     2007   NOMTCH   14530  53.1  46.9  57.0  19.3  18.7   4.6    0.3   0.0  11.8  32.5   7.1


Statistical Tests


MCH_TEST     DUPSOUT NONMCH_AVE MCH_AVE DIF_AVE NONMCH_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***       0      230.1   252.8    22.7      49.2   42.8   -6.3     87.1       0.0       .5304 .5098 .5510


GRP_TEST_1            PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              251.7   252.8     1.1     52.8   42.8  -10.0     81.1      47.8       .0219 .0066 .0372


GRP_TEST_2              CURR_AVE MCH_AVE DIF_AVE CURR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***               242.9   252.8     9.8    47.4   42.8   -4.6     90.3      45.1       .2178 .2023 .2333


GAIN_TEST             PRIOR_AVE MCH_AVE DIF_AVE PRIOR_SD MCH_SD DIF_SD RATIO_SD POOLED_SD EFFECT_SIZE CI_LO CI_HI
Cohan's D***              264.2   252.8   -11.4     46.3   42.8   -3.4     92.6      43.7       -.262 -.279 -.245
```