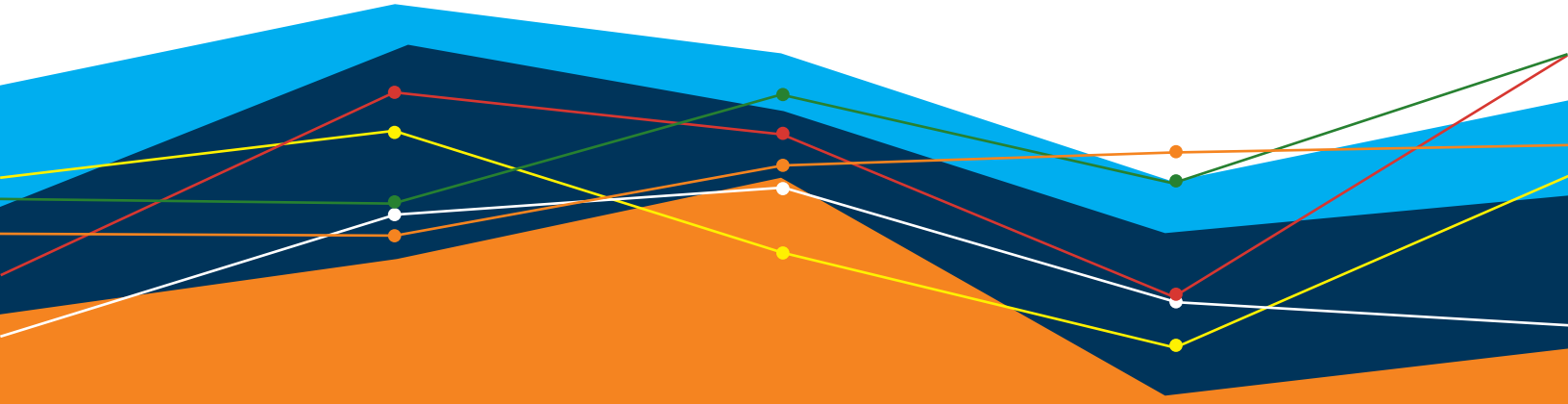
EVALUATING THE ACADEMIC PERFORMANCE OF CHOICE PROGRAMS IN CONNECTICUT:

A Pretest-Posttest Evaluation Using Matched Multiple Quasi-Control Comparison Groups



CSDE
CONNECTICUT STATE DEPARTMENT OF EDUCATION

Connecticut State Department of Education
June 2015

EVALUATING THE ACADEMIC PERFORMANCE OF CHOICE PROGRAMS IN CONNECTICUT:

A Pretest-Posttest Evaluation Using Matched Multiple Quasi-Control Comparison Groups

CSDE
CONNECTICUT STATE DEPARTMENT OF EDUCATION

Connecticut State Department of Education
June 2015

Dr. R. F. Mooney, Principal Investigator
Ajit Gopalakrishnan, Interim Chief Performance Officer
Mark Linabury, Bureau Chief, Bureau of Choice Programs
Kenneth Imperato, Education Consultant, Bureau of Choice Programs

CONTENTS

PREFACE

The Connecticut State Department of Education commissioned this evaluation to study if Choice programs were effective in raising academic achievement. Conducting an analysis of educational program outcomes is an inherently complex endeavor. This study also serves to illustrate the limitations of commonly drawn inferences made using annual assessment results. The CSDE appreciates the passion, commitment, and dedication of its former employee, Dr. Richard Mooney, for his continued partnership in designing and conducting this study.

ACKNOWLEDGMENTS

The State Department of Education wishes to acknowledge former Commissioner of Education Stefan Pryor and Chief Operating Officer Charlene Russell-Tucker for supporting the efforts required to conduct this study. We also wish to recognize the invaluable contributions made by Dr. Richard Mooney, former State Department of Education staff member, who served as principal investigator. We thank him for his continued partnership on this project. We acknowledge Ajit Gopalakrishnan, Interim Chief Performance Officer; Mark Linabury, chief of the Bureau of Choice Programs; and education consultant Dr. Kenneth Imperato, also of the Choice bureau, for their contributions to this project, as well. In addition, we thank Drs. Mohamed Dirir and Norma Sinclair from the Academic Office's Psychometric Analysis and Support Unit for sharing their expertise and for helping shape this project.

We also wish to extend special thanks to Professor H. Swaminathan of Connecticut's Neag School of Education for his review of the methodology for this analysis and helpful critique of the findings and interpretations.

EXECUTIVE SUMMARY

Public charter schools, interdistrict magnet schools, and the Open Choice program are collectively called Choice programs. One of their key missions is to improve educational outcomes of historically underperforming students from Connecticut's urban public schools. This analysis examines the academic growth and outcome performance based on the Connecticut Mastery Tests (CMT) for Choice program attendees from Connecticut's four largest cities—Bridgeport, Hartford, New Haven, and Waterbury—over a two-year period (2010 to 2012).

To conduct the most effective examination of ex post facto or pre-existing data where random assignment is either impossible or unethical (Murnane, R.J. and Willett, J.B., 2011), 30 stratified random samples of quasi-controls were generated from the co-present population of CMT test-takers. Then, the academic results for urban Choice program attendees were compared with results from the 30 samples. To counter known biasing influences such as higher baseline test performance, these quasi-control samples were matched with their respective Choice program "treatment group" on baseline test performance as well as on student background characteristics known to be related to test performance (Behuniak, P., Mooney, R. F., Cloud, R., 1990).

Results for each Choice program group and its respective quasi-control groups were tracked and compared longitudinally for the same students in two grade cohorts:

1. Grade 3 in 2010 to Grade 5 in 2012
2. Grade 6 in 2010 to Grade 8 in 2012

The use of longitudinal data allows us to ascribe academic performance gains over time to the educational interventions that have taken place; additionally, comparing gains achieved by the Choice program groups to their respective quasi-control groups enables us to control for gains that might have occurred naturally due to student maturation.

In the Grades 3 to 5 cohort, the analysis reveals statistically meaningful gains at or above the CMT Proficient level in interdistrict magnet schools operated by regional educational service centers (RESCs) and for the Open Choice program, and nearly statistically meaningful gains at or above the CMT Goal level for the RESC-operated interdistrict magnet schools.

In the Grades 6 to 8 cohort, public charter schools alone showed statistically meaningful gains at or above Proficient ***and*** Goal levels on the CMT.

This study remains an ex post facto, or "after the fact" analysis, thus not allowing causal attribution of the program outcomes. Hence, in practice it cannot be said with certainty that clones of these Choice programs, or an exportation of specific pedagogical techniques and strategies used, will necessarily ensure similar performance successes for urban students in general.

INTRODUCTION

Public charter schools, interdistrict magnet schools, and the Open Choice program are collectively called Choice programs. One of their key missions is to improve educational outcomes of historically underperforming students from Connecticut's urban public schools. The purpose of this evaluation is to answer a specific question: Are Connecticut's Choice programs succeeding in helping urban students (i.e., students from Connecticut's major cities of Bridgeport, Hartford, New Haven, and Waterbury) to close the academic achievement gap by enabling them to make greater academic gains as compared with peers who did not participate in these programs?

Four distinct Choice programs will be evaluated separately: public charter schools, interdistrict magnet schools operated by regional educational service centers (RESCs), interdistrict magnet schools operated by local school districts, and the Open Choice program. The academic outcomes used for this evaluation were based on results from the Connecticut Mastery Test (CMT).

Statement of the Problem: The central challenge in this study is that the performance of students attending these Choice programs tends not to be representative of the general urban population. As compared with neighborhood peers, urban students who choose to attend Choice program schools tend to reflect higher performance at baseline on standardized academic assessments. They also may differ in terms of background characteristics that are known to predict better test performance; for instance, Choice program attendees tend to have fewer special education students (SPED) or English language learner students (ELL). In order to conduct a fair, balanced, and meaningful evaluation, the key concern is to find an appropriate comparison group. So, who are the comparable peers?

DESIGN AND METHODOLOGY

Overview: To evaluate Choice programs in a fair, unbiased way, a carefully considered methodology is critical. The attribution of cause remains a major obstacle in the evaluation of any study where the treatment has occurred in the past (Campbell, D.T. and Stanley, J.C., 1963). A study that takes place after the fact is technically known as an ex post facto study. A traditional experimental research design using random assignment of subjects to an experimental and a control group prior to the treatment intervention is the only legitimate way to claim causal inference. Nevertheless, it is important to recognize the need for more effective examinations of ex post facto studies—as in the case of this Choice program evaluation—where random assignment is either impossible or unethical (Murnane, R.J. and Willett, J.B., 2011). Hence, we seek to conduct the best possible investigations, and draw the best possible conclusions despite the inherent limitations of ex post facto research. In other words, we should strive to do post hoc investigations in the spirit of sound scientific inquiry in spite of the fact that true causal conclusions are not possible.

Back in 1972, Harvard statisticians Frederick Mosteller and Senator Daniel Patrick Moynihan expressed frustration with the conclusions of the famous Coleman Report. They called for "better research designs" and "representative prospective longitudinal data in order to understand more comprehensively the impacts on children of investments in schooling" (Murnane, R.J. and Willett, J. B, 2011, p.6). Therefore, a good first step is to use pretest-posttest longitudinal academic assessment data. Measuring growth over time allows us to ascribe any performance gains to the educational interventions that have taken place from time 1 to time 2. A pretest-posttest only design is shown in figure 1:

Figure 1: Pretest-Posttest Only Design

O1 X O2, where

O1 = pretest group performance
X = treatment, i.e., Choice program
O2 = posttest group performance

Such a pretest-posttest design has important benefits over single-point-in-time studies because academic test-score gains can be directly attributed to educational teaching and learning that occurred during the period of study. While this design is preferable to a static or single-point-in-time analysis, one important limitation is that it does not control for student maturation (Campbell, D.T. and Stanley, J.C., 1963). Maturation refers to the relative educational test performance gains of students who attended a regular neighborhood school and did not attend a Choice program. Thus, it becomes impossible to tell what academic test performance gains might have occurred for Choice program attendees had they not received any treatment intervention. Therefore, to better understand what normal gains might have occurred without any intervention, an untreated and unbiased control group is needed.

Matching: Although a true random control group cannot be obtained after the fact, the next best option is to define a matched group of students who possess comparable academic achievement and background characteristics at baseline. We can do this by identifying student test-takers who are very similar to choice program attendees at baseline based on standardized test performance and performance related background characteristics. We can then use this information to track their performance and compare those results to Choice program attendees at outcome. Such a sample is intended to reflect a meaningful "quasi-control" comparison group.

Thus, in order to counter known biasing influences such as higher educational skills and abilities at baseline, the quasi-control group must be matched with the Choice treatment group on both baseline test performance and background characteristics known to be related to test performance.

To improve on the pretest-posttest only design (shown in figure 1), a matched quasi-control group can be added to help control for these known biasing influences and the maturational effects. Because it is not a true experimental design with random assignment to experimental and control groups prior to the introduction of the experimental intervention, it is known as a "quasi experimental" design (Campbell, D.T. and Stanley, J.C., 1963). Figure 2 (below) displays the structure of the pretest-posttest design with the addition of a matched quasi-control group:

Figure 2: Pretest-Posttest Quasi-Control Group Design

O1 X O2

C1 C2 Match, where

O1 = pretest group performance
O2 = posttest group performance
X = treatment, i.e., Choice program
C1 = mean of one matched control group at pretest
C2 = mean of one matched control group at posttest

The dotted line in figure 2 represents the fact that the treatment and control groups, although matched across critical attributes, are nonetheless not randomly assigned to treatment and control groups *prior* to the treatment intervention. This design is more powerful than the pretest-posttest only design shown in figure 1, but still cannot explain other possible underlying causal differences between students who opt to enlist in the Choice programs and those who do not—for example, student motivational factors.

While this design does not support causal interpretations concerning program outcomes, the addition of a matched quasi-control group nevertheless strengthens the pretest-posttest only design significantly because it allows for comparisons between the treatment group and untreated students with similar performance and background characteristics. It provides a meaningful lower-bound expectation of the outcome performance of very similar matched students who did not enter a choice program, *provided* the control group is a sufficiently valid representation of the treatment group at baseline.

Finally, we need a way to identify meaningful performance differences at outcome between the Choice program attendees and the quasi-control group. While many true experimental studies are faulted for basing conclusions on tiny samples that are insufficient to achieve statistical significance, an equally important practical problem in ex post facto studies is that comparisons can be identified as significant, regardless of their practical or substantive meaning, due to large samples. Technically, this somewhat obscure statistical problem is known as "inflation of the type 1 error rate." This very real technical problem occurs when researchers apply standard statistical tests intended for true experimental studies based on random samples to studies of large extant groups (Henkel, R.E., 1976).

What this means in practice is that the application of standard statistical tests can increase the potential for finding "statistical significance" even when such differences are trivial or not meaningful. Thus, ex post facto evaluations such as this Choice study, which has relatively large groups, can yield false-positive findings if standard statistical tests are applied. While these problems may seem exotic to the nontechnical reader, they are nevertheless legitimate and have been cited in critiques of recent national investigations of charter school performance (see Maul, A. and McClelland, A., 2013).

To resolve these problems, a novel strategy employed in this study is to use "matched *multiple* quasi-control groups" rather than just one matched quasi-control group. Thirty stratified random samples of students were selected from the regular test-taking population of CMT archives that match both the overall test performance ***and*** performance related background characteristics of each individual Choice program group. Figure 3 (below) shows the diagram of the design used in this study. It differs from figure 2 only because it uses *multiple* matched quasi-control samples. These are identified by the number "30" in parentheses after each representation of the matched quasi-control groups (see figure 3).

Figure 3: Pretest Posttest Design with Multiple Quasi-Control Groups Design

O1 X O2

C1(30) C2(30) Match, where

O1 = pretest group performance
O2 = posttest group performance
X = intervention, i.e., a Choice program
C1(30) = means of 30 matched control groups at pretest
C2(30) = means of 30 matched control groups at posttest

This approach has two critical benefits. First, it stabilizes the performance estimates for students in the general population who share the same performance and performance-related background characteristics for each of the Choice group comparisons; these estimates in total—the grand mean obtained from the assembled distribution of 30 means—best reflects the true population mean ***and*** provides a greater degree of stability for the performance estimates. Second, having 30 sample means rather than just one describes the empirical sampling error and, therefore, provides a practical and meaningful empirical benchmark of expected performance for students of similar background and ability against which to compare specific treatment group outcomes. The range of those sample means will be normally distributed, regardless of the shape of the parent distributions from which the samples were derived (Kerlinger, F.N., 1973). By establishing a predetermined cutoff value on the sample distribution at outcome and without resorting to standard statistical procedures, the two groups—those who attend the Choice programs and those who do not—can be compared with statistical rigor, while avoiding the type 1 error rate problem.

Stated differently, the matched quasi-control groups can thus be considered "virtual peers" and their performance at outcome will help level the playing field so that comparisons between treatment and controls can be both fair ***and*** meaningful. The virtual peers established at baseline are tracked co-presently over the same grade levels and time as the Choice program attendees and used to compare the relative test performance results at outcome or posttest. The specific detailed steps for obtaining the virtual peer samples are presented in appendix A.

Operational Definitions:

Cohorts: This study is based on a longitudinal matched pretest-posttest quasi-experimental design using two matched cohorts of students. Cohort 1 includes students in the Choice programs from pretest at Grade 3 in 2010 to posttest at Grade 5 in 2012. Cohort 2 includes students in the Choice programs from pretest at Grade 6 in 2010 to posttest at Grade 8 in 2012.

Dependent Measures: All student results reported in this study will be based strictly on matched student test records from 2010 to 2012 with valid pretest and posttest CMT results. Given that mobility tends to occur more frequently among urban students, this study moderately under-represents this group. Because the achievement gap affects both mathematics and reading performance, the dependent measure for this study is based on combining the results from both of these CMT tests. Thus, to achieve proficiency (which is achievement level 3 of 5), a student must be at or above the Proficient level in *both* mathematics *and* reading, abbreviated as MARD. Similarly, to achieve Goal level performance (which is achievement level 4 of 5), a student must be at or above the Goal level of performance in *both* mathematics *and* reading. Each program assessment will therefore be based on change in the percentage of students who have achieved these performance levels (henceforth referred to as Proficient and Goal) in both reading and math.

Performance over Time: To measure academic gains, it is first necessary to assess performance over time (Behuniak, P., Mooney, R.F. and Cloud, R, 1990). That is, to ascribe program gains or losses to academic treatments that have taken place, first baseline performance must be established and then, after a period of treatment—for example, membership in a specific Choice program—performance must be evaluated again at outcome. Only then is it possible to ascribe the educational activities from the treatment to the performance of the treatment group at outcome. For this analysis, student performance on the CMT is tracked from 2010 to 2012.

Achievement Gap: For this study, the achievement gap is defined as the academic test performance differential between urban and nonurban students. This achievement gap has been a critical issue for educators in Connecticut and throughout the country (Behuniak et al., 1990). For this study, the achievement gap is operationalized as a specific score range obtained from the test results stored in the CMT archives. The lower-bound estimate of the achievement gap is based on the longitudinal matched urban achievement results from the CMT for students who attended the same local school in the target districts at pretest in 2010 and posttest in 2012. The corresponding upper-bound estimate of the achievement gap is based on the summary test performance outcomes for all nonurban students (i.e., excluding students from all seven of Connecticut's largest cities) with valid CMT test results who attended the same school for the pretest/posttest period.

Hypotheses: A hypothesis is a formal assertion written in the form of a question that is central to the evaluation of a specifically defined research outcome. In this case, the hypothesis focuses on the determination of whether the Choice programs help students demonstrate better academic gains at outcome and achieve higher academic performance on MARD at posttest, over and above the benefits of attending a traditional urban neighborhood school.

A statistical test is fundamentally rooted in sampling theory and incorporates a concept known as the standard error. The standard error is a measure of variability that takes into account normal sample fluctuations.

Thus, if a particular experimental outcome exceeds the results of the control groups at posttest this in itself may not be meaningful, but if the difference also exceeds the extreme range of the empirical sampling error then the finding is considered statistically robust or "significant." In this evaluation, the interpretation that a particular finding is "statistically meaningful" describes cases where a particular Choice program exceeds the benchmark expectations or distribution of mean outcomes for the quasi-control groups.

There are two statistical tests of central importance in this study:

Hypothesis 1: Gains Test

Do the treatment gains at outcome exceed control gains in a statistically meaningful way? These gains[1] are diagrammed in the hypotheses depicted below. This hypothesis asks whether the treatment gains from baseline to outcome (T2 – T1) exceed quasi-control gains at outcome (C2 – C1).

Ho 1: $T2 - T1 \leq C2 - C1$ (Treatment gains are less than or equal to Control gains)

Ha 1: $T2 - T1 > C2 - C1$ (Treatment gains exceed Control gains)

Hypothesis 2: Outcome Test

Does the MARD performance of the treatment group exceed the control group at outcome? This answers the question: how did Choice program students perform at posttest compared with virtual peers who remained in their local neighborhood schools? This hypothesis and the alternative hypothesis are shown below.

Ho 2: $T2 \leq C2$ (Treatment outcome is less than or equal to Control outcome)

Ha 2: $T2 > C2$ (Treatment outcome exceeds Control outcome)

1. It should be noted that the pretest and the posttest measures based on CMT results (MARD) are derived from tests of different levels of difficulty (grade levels), and in this case the test performance levels have not been vertically equated on the measures of interest (i.e., mathematics and reading). The content of the tests reflects increasing pedagogical expectations from the lower to the higher grades. That said, this gains test is a good, practical way to measure student gains from grade to grade based on the Proficient and Goal levels. For example, if the percentage of students who are at or above the Proficient level at pretest are greater at posttest, and they exceed quasi-control peer samples by a meaningful extent, then this argues that academic growth has occurred for the treatment group over and above the expected growth for the quasi-control peers.

Statistical Analysis: Due to limitations related to ex post facto research designs, classical statistical tests are generally not appropriate for comparing group performance (Henkel, R.D., 1976). Instead, a novel strategy developed in this analysis is to use empirical tests using the empirical overall sample average and the empirical standard error of 30 quasi-control matched samples to determine statistical meaningfulness. An overriding principle of this study is to conduct all the statistical tests using the most conservative procedures. Hence, all tests will be conducted as if the comparisons at pretest and posttest are independent, and the *p* value[2] will be set at two standard deviations, which reflects the 95th percentile point of the normal null distribution of deviates. Thus, without resorting to the usual statistical procedures, the groups who attended the Choice programs and those who did not can be compared with statistical rigor. A detailed explanation of the empirical statistical testing approach is included in appendix B.

Reattribution: To counter the argument that Choice programs may be driving away lower performing students in order to bolster their test scores, the outcome scores of students who leave the Choice programs and go on to attend other Connecticut schools and take valid CMT tests in mathematics and reading are reattributed to the Choice programs at outcome. Unless the leavers are very high or very low performing at posttest (i.e., biased), this should have a neutral influence on outcome performance. Students who cannot be traced to Connecticut public schools will be regarded as cases of normal attrition. This approach is intended to be consistent with the overarching policy of conservative performance assessment decisions in this study. See appendix C for the attrition rates and the students reattributed.

2. *P* value is the probability or level of significance for rejecting a null hypothesis. In this instance, if a choice program evidences gains that are two standard deviations greater than the gains evidenced by the 30 quasi-control matched samples, then one can conclude with high confidence that the gains of the choice program were not by chance.

FINDINGS AND INTERPRETATIONS

The purpose of this report was to compare each of the Choice programs in order to determine their relative effectiveness at closing the achievement gap for students in Bridgeport, Hartford, New Haven, and Waterbury. Importantly, both the direct observation of the performance scores and gap indicators, as well as the development of statistical cut scores based on virtual peer samples converged and were in agreement in this analysis. A more detailed and technical discussion of the results is contained in appendix D.

Cohort 1 Findings: For cohort 1 (Grade 3 2010 to Grade 5 2012) the RESC magnet group performed best at closing the gap at both the Proficient level and the Goal level compared with the other Choice programs. The RESC magnet group made statistically meaningful gains over time of 25.4% at the Proficient level from pretest to posttest that exceeded the standard set by the virtual peer samples. In absolute terms, the RESC magnet group score at outcome was 83.6% at Proficient, coming to within -1.6% of closing the posttest achievement gap at the Proficient level on MARD.

Table 1. Cohort 1 CMT Growth (Grades 3 to 5) at Proficient level

	N	% Proficient on 2010 CMT	2010 gap %	% Proficient on 2012 CMT	2012 gap %	Change in % Proficient
Nonurban schools	18,318	78.9	--	85.2	--	6.3
Urban students						
Urban schools (non-Choice)	2,496	43.9	-35.0	48.3	-36.9	4.4
Public charter schools	184	63.6	-15.3	58.2	-27.0	-5.4
Magnet schools operated by local districts	353	48.4	-30.5	58.1	-27.1	9.7
Magnet schools operated by RESCs	55	58.2	-20.7	83.6†	-1.6	25.4†
Open Choice program	89	47.2	-31.7	66.3‡	-18.9	19.1†

† Exceeds empirical cut value ‡ Near empirical cut value

Furthermore, performance for the RESC magnet group at the higher Goal level of achievement was also very nearly statistically meaningful, attaining an absolute score of 56.4% at Goal and an absolute gap closure to within -14.9% of the majority for closing the gap at Goal (see Table 2). This reflects a substantial gain in goal performance of 21.9% for the RESC magnets. Goal level achievement is particularly important, because it demonstrates a higher level of learning and understanding.

The Open Choice group exceeded the statistical cut off compared with their virtual peer samples in terms of performance gains over time, which were 19.1% in absolute terms from pretest to posttest at the Proficient level (see Table 1). Open Choice also performed well for cohort 1 in terms of closing the gap at the Proficient level—although not doing as well as the RESC magnets. The Open Choice students very nearly exceeded the statistical cutoff value for the virtual peer samples in terms of overall MARD performance at posttest by achieving 66.3% of the students at the Proficient level. None of the other Choice programs met or exceeded these performance results for cohort 1. These findings are interpreted to mean that urban students in cohort 1 benefited more from the RESC magnet and Open Choice programs than from the other Choice programs based on MARD performance at outcome. The Open Choice program did not achieve a statistically meaningful level of performance at the Goal level for cohort 1.

Table 2. Cohort 1 CMT Growth (Grades 3 to 5) at Goal level

	N	% Goal on 2010 CMT	2010 gap %	% Goal on 2012 CMT	2012 gap %	Change in % Goal
Nonurban schools	18,318	59.1	--	71.3	--	12.2
Urban students						
Urban schools (non-Choice)	2,496	21.6	-37.5	29.8	-41.5	8.2
Public charter schools	184	39.7	-19.4	41.3	-30.0	1.6
Magnet schools operated by local districts	353	26.6	-32.5	38.0	-33.3	11.4
Magnet schools operated by RESCs	55	34.5	-24.6	56.4†	-14.9	21.9†
Open Choice program	89	24.7	-34.4	36.0	-35.3	11.3

† Near empirical cut value

An overall caution in interpreting the findings in cohort 1 is to bear in mind that the statistically meaningful findings for RESC magnet and Open Choice programs were based on rather small groups of urban attendees. The sample size of the programs was taken into account by the quasi-control statistical sampling procedure that was used, so this in no way reflects on the meaningfulness or robustness of the findings themselves. Nevertheless, as a practical consideration it has to be taken into account that smaller sized groups of urban students could potentially have received specialized educational support or interventions that might not be possible were these programs replicated to include larger numbers of urban students.

A second consideration in interpreting these cohort 1 results is to point out that the Open Choice program does not reflect a single unified educational program, but is actually a kind of omnibus "treatment" that allows urban students to volunteer to attend nonurban schools. Another confounding factor is that the number of urban students attending each individual school varies from school to school. Each school, therefore, may have somewhat different approaches to teaching and learning as well as variable levels of educational support and interventions. Again as a practical consideration, it is therefore difficult to say whether the Open Choice option can be precisely replicated or if the program would have the same successes if these schools had larger numbers of urban students in Grades 3 through 5.

Cohort 2 Findings: For cohort 2 (Grade 6 2010 to Grade 8 2012), the results demonstrate that public charter schools did best at closing the gap both at the Proficient level and at the Goal level on MARD by a statistically meaningful margin. Gains at Proficient for the charters in cohort 2 were 8.0%, widely outperforming the other Choice programs (see Table 3). The charter schools in cohort 2 posted an absolute performance score of 81.3% at the Proficient level at posttest and also demonstrated an absolute Proficient level gap closure of within -9.5% at posttest compared with the nonurban majority. These findings exceeded the empirical cut value for the virtual peer samples at posttest and also in terms of performance gains over time.

Table 3. Cohort 2 CMT Growth (Grades 6 to 8) at Proficient level

	N	% Proficient on 2010 CMT	2010 gap %	% Proficient on 2012 CMT	2012 gap %	Change in % Proficient
Nonurban schools	19,246	89.0	--	90.8	--	1.8
Urban students						
Urban schools (non-Choice)	2,352	61.3	-27.7	59.6	-31.2	-1.7
Public charter schools	326	73.3	-15.7	81.3†	-9.5	8.0†
Magnet schools operated by local districts	512	69.9	-19.1	69.3	-21.5	-0.6
Magnet schools operated by RESCs	96	75.0	-14.0	75.0	-15.8	0.0
Open Choice program	76	72.4	-16.6	75.0	-15.8	2.6

† Exceeds empirical cut value

Even more importantly, the performance at the higher Goal level showed that the public charter schools demonstrated a higher level of learning and understanding compared with the other Choice programs in cohort 2 (see Table 4). Gains at Goal for the charters in cohort 2 (10.7%) were even higher than that at Proficient (8.0%), again widely outperforming the other Choice programs. The absolute Goal performance on MARD was 60.1% at Goal, demonstrating a performance gap closure at the Goal level at posttest of -15.1% compared with the nonurban majority. This is nearly two or more times better than any of the other Choice programs at Goal for cohort 2. Furthermore, none of the other Choice programs approached a statistically meaningful outcome for cohort 2 at either the Proficient or Goal levels on MARD. This is interpreted to mean that the urban students in public charter schools benefitted more than urban students in the other Choice programs and therefore may be assumed to be better prepared for the academic demands of high school.

Table 4. Cohort 2 CMT Growth (Grades 6 to 8) at Goal level

	N	% Goal on 2010 CMT	2010 gap %	% Goal on 2012 CMT	2012 gap %	Change in % Goal
Nonurban schools	19,246	73.6	--	75.2	--	1.6
Urban students						
Urban schools (non-Choice)	2,352	37.8	-35.8	32.9	-42.3	-4.9
Public charter schools	326	49.4	-24.2	60.1†	-15.1	10.7†
Magnet schools operated by local districts	512	39.5	-34.1	34.2	-41.0	-5.3
Magnet schools operated by RESCs	96	42.7	-30.9	45.8	-29.4	3.1
Open Choice program	76	40.8	-32.8	31.6	-43.6	-9.2

† Exceeds empirical cut value

Interaction Effects for Cohorts 1 and 2: The inconsistent findings between cohort 1 and cohort 2 are puzzling. In research, this is known as an interaction effect. Once again, it is important to note that these findings showed a convergence of results for the descriptive test performance scores—such as in absolute performance terms at Proficient or Goal or on the gap closure index—as well as in terms of the findings for the empirical cut-value technique based on comparisons with random samples of matched quasi-control peers. Therefore, it seems that these overall findings support an interpretation of meaningful program performance differences. More investigations are required to draw conclusions about this interaction effect confidently. One possible explanation for the disparity is that strategies that work best for the younger students in cohort 1 may not be as effective for the slightly more mature students in cohort 2, and vice versa. Nevertheless, this is conjecture. Closer examination of these circumstances and a more complete understanding of the teaching methods and support environment would be helpful in interpreting these phenomena.

How can this study be used? This analysis provides an important benchmark for evaluating Connecticut's Choice programs. The methodology provided several ways to view performance comparisons, as well as including an innovative method of comparing program performance results at outcome to the results of matched multiple samples of quasi-control peers. These quasi-controls, or virtual peer samples, were derived from selecting matched random samples at pretest, tracking them over the same time as the treatment group, and comparing those findings at posttest with the Choice program attendees. The purpose of the quasi-controls is thus to adjust or control for baseline performance disparities or bias among the Choice programs and to provide a way to estimate the benefits of natural maturation for nonprogram attendees during the period of study (i.e., from 2010 to 2012).

This study demonstrates that pretest-posttest comparisons are clearly a superior method over traditional single-point-in-time or cross-sectional analyses for evaluating program effectiveness. Second, using matched multiple quasi-controls appears to be a useful enhancement to the traditional pretest-posttest quasi-experimental design.

Empirically determined performance benchmarks provide a useful way to compare the relative performance of the matched cohorts to help interpret the influence of normal student maturation effects and also to help determine robust and statistically meaningful program differences at outcome while controlling for type 1 errors. Short of employing true experimental designs with random assignment of subjects to treatment and control groups, replications of the analysis across different times would be a useful way of confirming the findings in this report.

Note that this study does not attend to other intangible gains that may be derived from attendance at these programs, including benefits to students that are not measured by the CMT. More needs to be known about the potential benefits of these programs beyond the limits of academic test performance alone.

A more detailed follow-up, school-level analysis of the best-performing programs is needed to determine the relative stability of these findings from school to school because higher overall program performance generally does not ensure that all schools are applying the model equally well. A school-level study can determine in more detail precisely what specific pedagogical or programmatic methods yield the observed performance gains in the better performing programs.

CONCLUDING OBSERVATIONS

This study examined the achievement of urban students from Bridgeport, Hartford, New Haven, and Waterbury attending Connecticut's Choice programs from 2010 to 2012 from several different perspectives. The primary finding is that urban students in cohort 1 (Grade 3 2010 to Grade 5 2012) benefitted more in terms of MARD test gains and performance at outcome from the RESC magnet schools at the Proficient level. These results in terms of absolute gap closure were consistent with comparisons with mean performance results for matched groups of virtual peer samples.

The Open Choice program also benefitted cohort 1 students on MARD at the Proficient level only in terms of performance gains. The RESC magnet program achieved near statistically meaningful performance at the Goal level as well. While demonstrating good performance gains and gap closure at the Proficient level is important, Goal level achievement is interpreted as reflecting a higher level of learning and understanding. It was noted that both of these programs are relatively small compared with some of the other Choice programs and, furthermore, that the Open Choice program actually reflects a broad diversity of treatments because each receiving school may be quite different. These issues raise questions about the broader applicability of these models.

In addition, the urban students in cohort 2 (Grade 6 2010 to Grade 8 2012) benefitted more from public charter schools both in terms of performance gains and absolute gap closure at Proficient as well as at Goal on MARD, and also exceeded the cut value established by the combined performance of matched groups of randomly selected peers. The charter performance effects at outcome demonstrated statistically meaningful performance outcomes at the Goal level of achievement, reflecting a substantially higher level of learning and understanding than achievement at the Proficient level. This is notably important for urban students because the middle school years are a gateway to high school, and this is the critical period urban students tend to decline in academic performance as compared with their nonurban peers.

An important caveat is that this study remains an ex post facto or "after the fact" analysis, thus prohibiting causal attribution for these program outcomes. This may in turn affect program transferability. Therefore, it cannot be said with certainty, for example, that clones of these choice programs, or a piecemeal exportation of the pedagogical techniques and strategies used, will necessarily ensure similar performance successes for urban students generally. Validation of these findings by identifying key elements of the programs and randomly assigning some or all of these programmatic or pedagogical methods of these programs to other urban schools would be a very helpful way to crosscheck these results further.

Also note that the Choice programs vary widely in terms of students' test performance at baseline compared with the neighborhood urban target districts examined in this study. Nevertheless, regardless of the causal reasons behind these baseline performance disparities, higher overall group performance at baseline was not always shown to result in better performance at outcome. Therefore, selecting students for admission to Choice programs based solely on better baseline achievement should not be considered a substitute for effective teaching and learning strategies.

This study is an example of a longitudinal pretest posttest study of student performance that compares the same students from time 1 to time 2. This study built on and extended the work of a prior report (Behuniak et al., 1990) by not only examining student performance over time, but also using several novel and innovative strategies for examining, reporting, and interpreting student performance. These include the development of absolute gap indicators as well as the development of a newly developed statistical test strategy based on empirical student performance of matched cohorts of students with similar baseline performance

and performance-related characteristics. This study therefore represents a new standard for innovation and thoroughness regarding the evaluation of programs designed to improve student performance.

Although this work provides a very important perspective on urban student performance in the various Choice programs, it is equally important to consider a number of issues not included in this study as these findings are reviewed and considered. For instance, while all the Choice programs showed improved performance over the students who remained in their urban local schools, these differences were not always statistically meaningful after taking into account baseline test performance and normal maturational effects using quasi-control groups, but may still have important long term benefits for urban student attendees above and beyond test performance alone.

In addition, apart from test performance outcomes alone, it is also important to take into account other practical considerations that could not be included in this study but which may also affect student achievement in the longer term. For example, the Choice programs may have many important long-term benefits for student participants, including social integration factors and elements other than academic test performance as measured on the CMT. Finally, in interpreting these results, keep in mind that prior program experience in the Choice programs occurring before the measured periods of cohort 1 and cohort 2 may help to explain cases of higher test performance at baseline.

REFERENCES

Behuniak, P., Mooney, R., Cloud, R. 1990. *Special Connecticut Mastery Test Research Report: Students at Risk Academically.* Report to the Connecticut State Department of Education, Hartford, CT.

Campbell, D.T. and Stanley, J.C. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally College Publishing Company.

Cook, T.D. and Campbell, D.T. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston: Houghton Mifflin Company.

Ferguson, George A. 1981. *Statistical Analysis in Psychology and Education (5th Ed).* New York: McGraw-Hill Bood Company.

Hambleton, R. & Swaminathan, H. 1986. *Item Response Theory: Principles and Applications.* Norwell, MA: Kluwer Nijhoff Publishing, Kluwer Academic Publishers.

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.

Kerlinger, F. N. 1973. *Foundations of Behavioral Research 2nd Ed.* New York: Holt, Rinehart and Winston, Inc.

Maul, Andrew and McClelland, Abby. 2013. *Review of "National Charter School Study 2013."* Boulder, CO: National Education Policy Center.

Mooney, R. 2013. Towards a Fairer Ex-Post Facto Program Evaluation Model Using "Virtual" Control Groups. Presentation at the Northeastern Educational Research Association Conference, October 23–25, in Rocky Hill, CT.

Mooney, R and Beaudin, B. 2009. Looking Backward: Using Data Mart Technology to Track Low-Performing Grade 4 Census Test-Takers in Connecticut. Presentation at the Northeastern Educational Research Association Conference, October 21–23, in Rocky Hill, CT.

Mooney, R and Beaudin, B. 2008. Using a Longitudinal Data Mart to Examine the Effects of Student Mobility on Test Performance Over Time. Presentation at the Northeastern Educational Research Association Conference, October 22–24, in Rocky Hill, CT.

Murane, R.J. and Willett, J.B. 2011. *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* New York: Oxford University Press.

Spector, P. E. 1981. "Research Designs." Sage University Paper Series on Quantitative Applications in the Social Sciences, series number 07-023. Beverly Hills and London: Sage Pubns.

APPENDIX A — QUASI-CONTROL SELECTION AND MATCHED TEST PERFORMANCE COMPARISONS ON PRETEST IN MATH AND READING (MARD)

Cohort 1 Grades 3 to 5

Title	sampcnt	mard_n	prespedpct	preellpct	prelunchpc	minorpct	preedfac	presocfac	premardp
1.0 CITIES OUT	--	18318	5.8	4.0	19.6	20.2	9.5	17.8	78.9
2.0 LOCAL DIST	--	2496	3.6	14.3	93.0	86.3	17.4	85.0	43.9
3.0 CHARTER DIST	--	184	1.6	1.1	68.5	98.4	2.7	97.3	63.6
3.1 RANDOM	373	185	1.6	1.1	71.5	97.9	2.7	97.9	60.4
4.0 MAG NOT RESC	--	353	2.5	7.9	81.3	82.7	10.5	80.7	48.4
4.1 RAND	132	349	2.6	8.0	68.5	88.2	10.6	88.0	51.6
5.0 MAG RESC	--	55	3.6	0.0	45.5	92.7	3.6	92.7	58.2
5.1 RAND	704	55	3.6	0.0	56.6	93.6	3.6	93.5	57.8
6.0 OPEN CHOICE	--	89	6.7	2.2	68.5	94.4	7.9	94.4	47.2
6.1 RAND	63	93	6.5	2.2	70.5	93.2	8.6	93.1	48.4

Cohort 2 Grades 6 to 8

Title	sampcnt	mard_n	prespedpct	preellpct	prelunchpc	minorpct	preedfac	presocfac	premardp
1.0 CITIES OUT	--	19246	7.1	1.8	17.3	16.4	8.8	14.4	89.0
2.0 LOCAL DIST	--	2352	4.3	11.4	90.6	85.8	15.2	84.3	61.3
3.0 CHARTER DIST	--	326	3.1	5.2	73.9	99.4	8.0	99.4	73.3
3.1 RANDOM	86	351	2.8	4.8	72.2	96.8	7.7	96.7	69.6
4.0 MAG NOT RESC	--	512	2.3	7.0	84.8	93.0	9.4	91.8	69.9
4.1 RAND	57	532	2.3	6.8	71.9	94.6	9.0	94.4	67.9
5.0 MAG RESC	--	96	2.1	5.2	62.5	88.5	5.2	88.5	75.0
5.1 RAND	59	95	2.1	5.3	68.3	93.9	7.4	93.8	73.3
6.0 OPEN CHOICE	--	76	14.5	3.9	57.9	93.4	17.1	93.4	72.4
6.1 RAND	46	88	12.5	3.4	63.0	85.8	15.9	85.6	71.2

Specific Steps for Determining the Stratified Virtual Peer Samples — The test-performance-related background characteristics for population and random sample groups for the Grades 3 to 5 cohorts and the Grades 6 to 8 cohorts are displayed above. The key to this strategy lies in the ability to obtain fair representation of the quasi-control group performance and background characteristics at baseline in order to properly assess the performance outcome of quasi-control peers who did not attend the Choice programs. The challenge is to tailor the quasi-control groups to each of the experimental or Choice program groups that received the educational treatments because each program had somewhat different combinations of baseline performance and background characteristics that could influence performance at outcome. The problem is that the total census distribution of test scores also reflects the students' background social disadvantage factors. Specifically, this means that the urban students as a subpopulation tend to have lower than average test scores as well as higher socio-economic factors. Therefore, selecting on social disadvantage factors tends to limit the range of test scores and selecting on test scores limits the range of social disadvantage factors.

This range restriction or sampling problem was resolved using a stratified sampling technique that blended both univariate and multivariate sampling techniques. Univariate sampling means simply that examinees were selected for a single trait, whereas multivariate sampling means that examinees were selected on multiple traits simultaneously. By combining these two strategies, it is possible to select quasi-control groups based on key background traits effectively. First, because the Choice program samples had relatively few students with disabilities as compared with the urban subsample generally, these two groups were randomly sampled from the general population as if they were independent, univariate samples. For example, if 1% of a given Choice program group were special education students, then an equivalent 1% of special education students who were not Choice program attendees were randomly sampled from the general population and retained. A similar univariate subsample of English language learners were also randomly obtained from the general population.

After obtaining the two small samples of students with disabilities and English language learners, and retaining those, a much larger multivariate sample of students who were either jointly or independently either eligible for free/reduced-priced school lunch and/or from minority racial/ethnic backgrounds was randomly sampled from the general population. This was the largest sample because it reflected the majority of the subset of urban Choice program attendees. The reason for doing this sample jointly is to more accurately reflect the reality that most urban students attending the Choice programs are either minority, school-lunch eligible, or both. Next, another relatively small group of nonminority, non-school lunch students was also sampled from the general population. In the last step, all four of these random subsamples were combined together to create one complete, random quasi-control group "candidate" sample. This candidate sample was then tested and compared with the true Choice program group. If the baseline test performance for the candidate quasi-control group was within 5% of the true Choice program group test performance and the background characteristics were within 15% of the true program group background characteristics, then the summarized candidate results are retained. If the criteria are not met, the results are discarded and another random sample sequence begins. The process continues until 30 summarized samples are obtained.

APPENDIX B – STATISTICAL TESTING APPROACH

Empirical Statistical Testing: The multiple quasi-control matched group samples used in this study are intended to establish a concrete and meaningful statistical performance benchmark in order to fairly assess the relative test performance of the Choice program participants at outcome or posttest. That is, each of the random sample quasi-control groups that are matched or tailored to each of the four individual Choice program groups at pretest become, in effect, a test-performance benchmark of virtual urban peers. Then by tracking these same quasi-control sample groups from pretest to posttest, the summarized results of the test performance for the quasi-control groups therefore provide a valid and meaningful point of comparison for each of the appropriate Choice program test results at outcome.

A related issue has to do with determining a proper statistical test for deciding whether a Choice group has meaningful performance increases on the CMT over and above the matched quasi-control groups at outcome. Using normal statistical tests for ex post facto analyses of census data such as this is improper because these tests are based on the assumption of random assignment at baseline (Henkel, 1976). This leads to two problems: First, the distributional assumptions required in classical statistical analysis may not be appropriate and, second, larger groups tend to be found "statistically significant" due purely to the precision of the statistic. "…[F]inding or failing to find a significant result is often more a function of sample size than the intrinsic truth or falsity of the null hypothesis" (Henkel, p. 82, 1976). These are problems with statistical tests per se, but they are concerns relevant to the proper interpretation of the findings of an ex post facto analysis when employing classical statistical tests.

To resolve these issues, a novel strategy is employed in this analysis. Instead of sampling just one group of quasi-control "peers" 30 valid or viable samples of peer groups were sampled from the CMT archives for each group comparison of interest. Each one of these was subsequently summarized and compiled for each Choice program comparison. That is, the results for each viable sample were first individually summarized at the group level and then those summarized results were compiled together and summarized again. This process has two purposes: First, this compilation of summarized results serves to stabilize the outcome performance for the benchmark comparison groups. Second, it also provides a "null" performance distribution, which can be used to establish a robust empirical estimate of sampling error, while avoiding well-known problems associated with using traditional inferential statistics for census data.

Choosing an Appropriate Cut-Score Value: In order to decide whether a Choice program performance difference exceeds the matched null distribution, a cut value needs to be determined. Although the random samples may or may not be normally distributed, sampling theory dictates that the averaged and compiled performance levels derived from the 30 random samples of quasi-control test takers should be. Accordingly, a cutoff value of two standard deviations from the mean of the Proficient or Goal scores was selected. This choice is based on the expectation that the properties of the normal distribution can be evoked when the n-count for the number of samples is 30. Thus, although the Proficient or Goal performance of the individual sample groups may not be normally distributed, the means of the performance distributions should be. Therefore, 2 standard deviations reflect a very stringent benchmark standard that corresponds to exceeding approximately the 95th percentile of the one-tailed normal z-score distribution (Ferguson F.N. 1973). Thus, without resorting to the usual statistical testing procedures using theoretical distributions, the matched treatment and control groups can be compared with statistical rigor.

Interpretation of Statistical Cut Values: What this cut-score value means in practice is that if a Choice group performance is in the lower or middle range of the "null" distribution, then the findings imply that the program performance is not meaningfully different from the peers, but if the Choice group performance is at the extreme outer limit of this "null" distribution, then the findings imply that those results are so extreme that we say that they are probably from a completely different population. Thus, we can think of each Choice program assessment as a comparison of that single outcome to a distribution of outcomes for 30 same-sized groups of comparable controls, so the notion of a sampling distribution of these 30 samples is relevant.

Exceeding this cut value compared with a normal z-score distribution reflects a conservative, robust, and meaningful way to detect performance differences at outcome for the Choice programs over and above equivalent samples of students who did not attend the Choice programs. The advantages of this model are two: First, it attempts to take into account the critical issue of differential baseline test performance for the Choice program groups due to selection bias. Second, it provides a benchmark performance level at outcome that helps to control for or explain the effects of student maturation over time.

In summary, the proper understanding and interpretation of the benefits of the random peer group results in this study is that the statistical cut value measure establishes what might be called a lower-bound performance expectation at outcome for assessing Choice program effects, although this model does not control for the hidden effects of confounding factors such as student motivation. Also, because this study is not a true experimental design with random assignment of subjects to treatment and controls at baseline, it is not possible to attribute cause to the observed program effects. Hence, if one or more of the Choice programs are observed to be effective at closing the gap in this study, this alone will not be sufficient proof that transferring these strategies to other schools and educational programs is necessarily going to lead to similar positive findings in the future.

Statistical Analyses: To assess the above hypotheses for the four Choice program options, student performance outcomes based on the combined MARD performance metric will be investigated in multiple ways. These include examining baseline percentage at Proficient and Goal, outcome percentage at Proficient and Goal, and percentage of gains over time (the difference between pretest and posttest performance) for each of the programs.

Descriptive Comparisons: The first and most fundamental descriptive index in this analysis is the degree to which these programs help students from Bridgeport, Hartford, New Haven, and Waterbury close the achievement gap in an absolute sense. This statistic will be presented directly by subtracting the performance group outcome to the upper-bound estimate of the gap. This will express in absolute terms the percentage that remains between the program performance result and the gap target index. The index reflects an upper-bound performance expectation that will always be in the negative unless a Choice program result exceeds the overall nonurban (see equation 1, below).

Equation 1: Descriptive Gap Index

$$\text{Gap index} = \text{PostMARD } i - \text{UpperBoundGap (Cities Out)}, \; i = 1,2,3,4 \; .$$

Next the gain index for each of the four Choice programs will be examined. Normally, a dependent pretest-posttest analysis will compare each individual student's score at pretest and posttest, which is the most sensitive way to detect change. However, this test is more conservative because it will compare

independent pretest-posttest gains. The pretest and the posttest measures based on CMT results (MARD) are derived from tests of different levels of difficulty (grade levels) and in this case the test performance levels have not been vertically equated on the measures of interest (i.e., mathematics and reading). The content of the tests reflect increasing pedagogical expectations from the lower to the higher grades. That said, this gains test is a good, practical way to measure student gains from grade to grade based on the Proficiency and Goal levels in a relative way. For example, if the percentage of students who are at or above the Proficient level at pretest is greater at posttest, and they exceed quasi-control peer samples by a meaningful extent, then this argues that academic growth has occurred for the treatment group over and above the expected growth for the quasi-control peers. Accordingly, the raw gains are calculated for each program as follows (see equation 2, below):

Equation 2: Descriptive Gain Index

$$\text{Gain index} = \text{PstMARD } i - \text{PreMARD } i,\ i = 1,2,3,4\ .$$

Cut-value Comparisons: Equation 3 (below) describes the cut value for the outcome test comparing actual performance of the Choice groups to the virtual peer results. These tests mirror the descriptive tests above, except that they compare the actual posttest results for the Choice programs to the matched results for the 30 samples of virtual peers. Accordingly, this comparison to the null distribution reflects the combined results for the matched and combined virtual program groups. This allows the cut value to be more representative of the population for these particular test-takers and also makes it possible to assess a meaningful standard error derived from the 30 samples. Therefore, the cut values represent a meaningful and robust statistical comparison at outcome.

Equation 3: Empirical Benchmark Outcome Test

$$\text{Cut value} = \text{Avg(PostMARDRand)} + 2.0 \times (\text{SD (PostRand } i),\ i = 1,2,3,4\ .$$

Besides the direct outcome comparison at posttest, we are also interested in the relative gains of the Choice groups over time. To do this, we compare the pretest results for each group with the average group difference from the posttest outcome scores on MARD, and compare those quasi-control group independent gain findings to the Choice group's independent gains for the same period (see equation 4, below):

Equation 4: Empirical Benchmark Gain Test

$$\text{Cut value} = \text{Avg(PostMARDRand)} - \text{Avg(PreMARDRand)} + 2.0 \times$$
$$\{\text{SD (PreMARDRand } i) + \text{SD(PostMARDRand } i)/2\},\ \ i = 1,2,3,4\ .$$

Note that this test for independent groups is less sensitive (more conservative) than the two sample gains test, where the error term is 2 standard deviations of the gains rather than the pooled total group standard deviation.

APPENDIX C — ATTRITION COUNTS BY CHOICE PROGRAM

Grades 3 to 5 Cohort

Program	Attrition *n*	MARD_*n*	% Attrition
3.0 CHARTER	18	184	9.8%
4.0 MAG NOT RESC	12	353	3.4%
5.0 MAG RESC	*	55	*
6.0 OPEN CHOICE	*	89	*

Grades 6 to 8 Cohort

Program	Attrition *n*	MARD_*n*	% Attrition
3.0 CHARTER	25	326	7.7%
4.0 MAG NOTRESC	21	512	4.1%
5.0 MAG RESC	*	96	*
6.0 OPEN CHOICE	*	76	*

*Attrition *n* is less than 6; therefore, *n* and % attrition suppressed to protect student confidentiality.

APPENDIX D — DETAILED DISCUSSION OF RESULTS

Descriptive Comparisons: Each of the following four results tables will present the descriptive analysis for each population parameter and the cut-value comparisons as appropriate. The tables include the performance of the state as a whole (excluding the seven major Connecticut cities) to provide an upper-bound estimate of the performance gap based on MARD. The lower-bound estimate will be based on MARD performance for the four targeted districts of Bridgeport, Hartford, New Haven, and Waterbury. These indices will be calculated both at pretest and posttest. Next the summary population performance for the local schools in the four urban districts is presented, which will provide a lower-bound estimate of the performance gap.

The individual programs that were analyzed were the public charter schools, the non-RESC magnet schools, the RESC magnet schools, and finally the Open Choice program results. To reduce the size of the tables, the percentage of special education and English language learner program attendees will be combined into one index called Education Factors. A second index called Social Factors will combine the percentage of students who are minorities and/or those eligible for free or reduced priced meals.

Quasi-Control Comparisons: The total sample count in each table representing the four Choice program assessment findings reflects the total number of samples needed to obtain 30 viable quasi-control group comparison samples (see table 5). Samples of the virtual peers met the following criteria: First, they must have been within 15% of the school-lunch eligibility percentage for each specific Choice program at baseline, and second their baseline Proficient level scores had to be within 5% of the baseline test performance results on the combined mathematics and reading results (i.e., MARD).

The results for the virtual peers are reported immediately following each listing of the Choice program group results. Samples that do not meet these criteria for inclusion in the comparison groups will be counted but not included in any outcome comparisons. The reason for counting the total number of samples taken is to provide a way to gauge the relative degree of difficulty in obtaining viable random samples. The larger the total number of samples, the more difficult it was to obtain 30 viable samples. This can occur, for example, when the baseline program selects a high percentage of students with higher baseline performance on MARD.

Student performance outcomes based on the combined MARD performance metric will be investigated in multiple ways: Baseline percentage at Proficient and Goal, Outcome percentage at Proficient and Goal, Percentage of Gains over time (the difference between pretest and posttest), and relative Percentage Closing the Achievement Gap. A key descriptive index in this analysis is the degree to which these programs help students from Bridgeport, Hartford, New Haven, and Waterbury close the achievement gap in an absolute sense. This statistic will be presented directly, subtracting the performance group outcome to the upper-bound estimate of the gap. This will express in absolute terms the percentage that remains between the program performance result and the gap target index. This index will always be in the negative, unless Choice program results exceed the results for the nonurban majority.

Cohort 1 Results: Table 5 (below) displays the pretest/posttest results for cohort 1 for Grade 3 2010 to Grade 5 2012 at the Proficient level or higher. The chief benefit of a pretest-posttest analysis is that it better enables us to ascribe observed change to academic program effects, without confounding the findings with the test results from outgoing or incoming students. Hence, all students reported in table 5 had valid test results on the CMT mathematics and reading subtests (MARD) in Grade 3 and also in Grade 5. In addition, two different kinds of results are displayed in each table: census results and empirical random sample results.

The whole number rows in each table display the pretest-posttest census results for all valid MARD test-takers for the appropriate performance level. Thus, row 1.0 shows the findings for all Connecticut's students excluding the seven major cities while row 2.0 shows the combined results for the four target urban districts (Bridgeport, Hartford, New Haven, and Waterbury). The decimalized rows list the findings for the combined, randomly selected quasi-control groups, which are matched to each individual Choice program group. Thus, row 3.0 describes the census charter results and row 3.1 describes the results for the combined 30 quasi-control samples that are matched to the census charter group.

Besides the performance scores listed in table 5, descriptive performance-related background characteristics are also listed for each group. For this analysis, the critical performance related background factors have been combined for compactness. (More detailed descriptions for each of the key data elements are described in appendix A.) Hence, in table 5 the column labeled SocFac, or social factors, describes the percentage of students in each row listing with both minority status and/or eligible for the School Lunch Program.

The column labeled EdFac, or educational factors, shows corresponding percentage of students who are special education and/or English language learners. The baseline test results on table 5 are labeled Prof1 and display the combined percentage of students meeting or exceeding proficiency on both the mathematics and reading CMT subtests in Grade 3 of 2010. The field labeled Prof2 corresponds to MARD test results for the same students at Proficient or higher at outcome in Grade 5 of 2012. The column labeled ProfGain describes the difference between Prof1 and Prof2 group scores, which is an indicator of relative gain.

The absolute gap closure index on table 5 is labeled Gap1 at pretest and Gap2 at posttest. This is the percentage difference between each Choice program performance on MARD at each test-reporting level compared with the overall statewide performance for nonurban students in Connecticut. These indicators are calculated for both the Proficient level or higher in table 5 and for Goal level or higher in table 6 for cohort 1. The "Valid N" is the record count for those students with valid results in both subtests who remained in the same school from the pretest to the posttest. Line 1 reflects the overall statewide results for the nonurban students and reflects the upper-bound limit or target goal for the gap indicator.

The results for each Choice program group will be followed by the results for a matched sample results found in the decimalized rows.

Recall that the matched sample is the summary of 30 subsamples, matched on relative baseline MARD performance and on the four background characteristics of interest (i.e., school lunch eligibility, minority status, special education status plus English language learner status). The relative comparability of these groups is reflected on both the education factors and the social factors indices as well as on pretest MARD performance.

The Samp_n field denotes the total number of samples required to obtain 30 viable samples where the background parameters were held to within 15% of the true population parameter of the Choice program and the baseline test performance at pretest were held to within 5% of the true population parameter of the Choice program. The higher the Samp_n, therefore, the more difficult it was to obtain 30 viable samples. What this means in practical terms is that higher Samp_n counts reflect a higher degree of difficulty in finding sample matches that are closely matched to the census choice program groups. This occurs most typically when the program groups enroll students of higher baseline performance compared with expected performance level of the four target districts.

Row 1.0 of table 5 labeled Cities Out displays the overall results for nonurban Connecticut test-takers at the Proficient level (or higher) on MARD. Row 1.0 provides the background and performance information for the group (n = 18,318) who had valid test results from pretest of Grade 3 of 2010 to posttest in Grade 5 in 2012. The education factors for this group is 9.5% and the social factors is 17.8%. Pretest proficiency is at 78.9% and posttest proficiency is 85.2%, a proficiency gain of 6.3% (see ProfGain). The results for the target districts of Bridgeport, Hartford, New Haven, and Waterbury (n = 2,496), excluding Choice program participants, shows 17.4% of these students have education factors and 85.0% have social factors.

More importantly, table 5 also reveals the widely disproportionate academic performance disparities between the urban target districts and non-urban students in Connecticut. Row 1.0 reflects the nonurban students and thus corresponds to the upper boundary of the academic performance gap on MARD, while row 2.0 reflects the target urban districts described in the lower boundary. The column labeled Prof1 shows the Proficient level or higher performance for the Grades 3 to 5 cohort at Grade 3 in 2010. Row 1.0 shows the nonurban results at Proficient level or higher performance on MARD and for Grade 3 students in 2010, which is nearly 80% (78.9%) for most Connecticut students. Meanwhile for the target districts in row 2.0, the Prof1 result shows that only a little less than half (43.9%) of the urban target district students achieve proficiency at pretest. This reflects a performance gap of -35.0% at pretest (see Gap1 on table 5).

Prof2 in table 5 corresponds to Prof1 and shows the posttest or outcome performance on MARD two years later for the same students tracked over time at Grade 5 of 2012. Here the overall statewide performance on MARD for nonurban students has gone up from 78.9% at pretest in Grade 3 of 2010 to 85.2% at Grade 5 in 2010, a performance gain of 6.3% from Grade 3 to Grade 5 (see ProfGain in table 5). For Connecticut's urban target group, the posttest performance is up from 43.9% in Grade 3 of 2010 to 48.3% proficiency or higher by Grade 5 of 2012 (see Prof2 of table 5), and the performance gap has actually increased to -36.9% (see Gap2 of table 5).

This disparity on the achievement gap has grown worse at posttest because the gains for the targeted urban students have only grown by 4.4% for the urban target group in row 2.0, compared with 6.3% for the nonurban students displayed in row 1.0. Taken as a whole, these performance disparities reflect a troubling situation, which is that not only are the target urban students behind academically as early as Grade 3, but that that gap has increased after two years.

Rows 3.0, 4.0, 5.0, and 6.0 describe the census population results for students from the targeted four urban districts attending the Choice programs. For example, row 3.0 is the results for the 184 urban students from the target districts who had valid pretest and posttest results on MARD who attended the same charter school from Grade 3 of 2010 to Grade 5 of 2012; it also includes the exited students who were reattributed (appendix C). These students reported 2.7% of the group had education factors while 97.3% had social factors. Therefore, the urban charter attendees from Bridgeport, Hartford, New Haven, and Waterbury had fewer education factors than the student peers attending local schools but about the same level of social factors (97.9% versus 97.3%). Notably, these urban charter students were much better performers on MARD at pretest, performing at 63.6% Proficient at pretest compared with urban peers who only scored at 43.9% at pretest. Hence the performance gap at pretest (i.e., see Gap1 on table 5) is only -15.3% versus -35.0% for the local urban school attendees (see Local).

The posttest gap of -36.9% (see Gap2 of row 1.0, LOCAL DIST, table 5) for the LOCAL DIST group performance is lower than any of the Choice programs, suggesting that all the programs are beneficial to some degree. However, at outcome the posttest charter performance drops to 58.2% for a proficiency gain of

-5.4% on MARD, hence showing an increase in the proficiency gap as compared with the rest of the state of -27.0% at posttest (from -15.3 at pretest). Table 5 also displays the results for the matched random charter samples in row 3.1. Note that it took 373 samples to obtain 30 valid comparison samples for the charter group (see table 5, row 3.1 RANDOM). This reflects the relative difficulty of finding students with background characteristics at pretest sufficiently similar to the urban charter students who also had similar pretest scores on MARD at Proficient. In addition, the relative performance at proficiency for the quasi-control peer was only 60.4% for the random samples compared with 63.6% Proficient for charter students.

This finding is interpreted to reflect the fact that it was difficult to obtain 30 valid random samples of students with about 2.7% EdFac and 97.3% SocFac and comparable test performance characteristics. Nevertheless, although the random samples had lower performance at baseline compared with the charter students, at outcome the random quasi-control group did better than the charter students, achieving 64.6% Proficient on MARD at posttest versus 58.2% for the charter group.

This means that while the charter student group did better than the local urban students at posttest, this outcome performance was not statistically meaningful after taking into account the high relative baseline performance and the particular background composition of the charter group at baseline compared with the results for the charter matched random samples (see table 5, row 3.1). In more detail, the charter random samples had an outcome score of 64.6% Proficient (see table 5, row 3.0 column Prof2) with a standard deviation of 2.5% (see StdPst in table 5). Therefore, by looking at the Proficient population parameter of 64.6% for the charters, the expected outcome performance comparison cutoff score derived from the results of the quasi-random yields a comparison benchmark of 69.6% (see table 5, row 3.1, PostCut), which is over 10% above the charter Prof2 Proficient level of 58.2%.

Empirical Cut-Value Assessment for Table 5: Thus the charter population group for the Grades 3 to 5 cohort would have to have achieved a population outcome score over 69.6% at Proficient or higher on MARD to have met or exceeded the statistical cutoff level. Unfortunately, the charter schools performance on MARD at posttest was below this cut value (58.2%), and hence these results are not meaningfully different from the quasi-control outcome after taking into account the random fluctuations of the sample groups. Furthermore, the gains cutoff for meaningful departure from the empirical random samples for the charters is 8.5% (see GainCut in table 5) but the actual population gains were only -5.4% (see ProfGain in table 5). We can therefore conclude that the urban charter attendees would not exceed the expected cutoff scores on performance gains on MARD.

The results for the non-RESC magnets displayed in row 4.0 of table 5 show that these students performed only slightly higher at baseline compared with the LOCAL DIST students (see row 2.0 of table 5). The baseline results on the MARD at Proficient show 48.4% Proficient for the non-RESC magnets compared with 43.9% Proficient for the LOCAL DIST results. Therefore, the non-RESC magnet group at baseline was far more consistent with local student performance than the charter group. This is also reflected in the background characteristics for this group.

Table 5: Cohort 1 Proficient Level or Higher (Grade 3 2010 to Grade 5 2012)

Title	Samp_n	MARD_n	EdFac	SocFac	Prof1	StdPre	Gap1	Prof2	Gap2	StdPst	PstCut	ProfGain	GainCut
1.0 CITIES OUT	--	18318	9.5	17.8	78.9	--	--	85.2	--	--	--	6.3	--
2.0 LOCAL DIST	--	2496	17.4	85.0	43.9	--	-35.0	48.3	-36.9	--	--	4.4	--
3.0 CHARTER DIST	--	184	2.7	97.3	63.6	--	-15.3	58.2	-27.0	--	--	-5.4	--
3.1 RANDOM	373	185	2.7	97.9	60.4	1.7	-18.5	64.6	-20.6	2.5	69.6	4.2*	8.5
4.0 MAG NOT RESC	--	353	10.5	80.7	48.4	--	-30.5	58.1	-27.1	--	--	9.7	--
4.1 RANDOM	132	349	10.6	88.0	51.6	1.4	-27.3	59.6	-25.6	1.7	63.0	8.0	11.1
5.0 MAG RESC	--	55	3.6	92.7	58.2	--	-20.7	83.6†	-1.6	--	--	25.4†	--
5.1 RANDOM	704	55	3.6	93.5	57.8	2.2	-21.1	65.0	-20.2	4.7	74.4	7.2	14.1
6.0 OPEN CHOICE	--	89	7.9	94.4	47.2	--	-31.7	66.3‡	-18.9	--	--	19.1†	--
6.1 RANDOM	63	93	8.6	93.1	48.4	2.0	-30.5	57.7	-27.5	4.5	66.7	9.3	15.8

† Exceeds empirical cut value

‡ Near empirical cut value

** Figures not exact due to rounding*

The EdFac index is 10.5% for the MAG NOT RESC program compared with 17.4% for the LOCAL DIST students. However, SocFac is somewhat

lower at 80.7% for MAG NOT RESC compared with 85.0% for the LOCAL DIST students. Nevertheless, this better balance of baseline conditions more closely reflects the overall profile of the target urban districts as a group. This is also reflected in the smaller number of random samples needed to achieve a viable group of 30 samples, which is 132 for the MAG NOT RESC random group (see row 4.1 in table 5) versus nearly triple that (Samp_n = 373) needed for the charter random group (see row 3.1 of table 5).

Despite a baseline profile for MAG NOT RESC group that is more consistent with urban students generally (see row 4.0 of table 5), the PstCut estimated cutoff score for the random groups is 63.0%, while the performance for the MAG NOT RESC group was below this cut point at 58.1%. Hence, these results are also not statistically meaningful after taking into consideration the statistical cut point. As for the gain score cut-point comparison, the pretest to posttest Proficient gains for MAG NOT RESC were 9.7% (see table 5, row 4.0, ProfGain), which did not exceed the estimated cutoff value of 11.1% for the random quasi-control samples (see row 4.1 of table 5, GainCut), and again, this was not a statistically meaningful gain on proficiency. As was the case with the charters analysis for cohort 1, the MAG NOT RESC group also did not demonstrate sufficient performance increases at posttest on proficiency to overcome the GainCut level.

However, the picture is sharply different for the MAG RESC group displayed on row 5.0 of table 5. The MAG RESC group performance shows a meaningful departure from the cut value at posttest as well as on performance gains. The posttest cutoff score (PstCut) was 74.4% for the random samples (see table 5, row 5.1) and this was exceeded by the MAG RESC group performance of 83.6%. Notably, a substantial 704 samples were required to obtain 30 valid random samples matching the MAG RESC group's baseline performance of 58.2%.

Once again, this difficulty of obtaining 30 viable samples of students with MARD performance above the expected for students with a level of social and educational factors for the program group is explained by the fact that most urban students in the target districts are achieving at only 43.9% proficiency on average (see row 2.0, LOCAL DIST Prof1). Therefore, finding 30 random samples of students from the general population who match the profiles reflected by the MAG RESC students required many samples to be discarded.

The gains for the MAG RESC are substantial at 25.4% at proficiency (compare table 5, row 5.1, GainCut versus row 5.0, MAG RESC, ProfGain), and exceed the gains of all the Choice programs compared in table 5. In addition, these gains exceed the expected random quasi-control group gains cutoff at 14.1% and therefore can be considered statistically meaningful. This impressive level of performance gains can also be observed in the Gap2 column of table 5 for this program, which shows that MAG RESC came to within -1.6% of closing the gap completely (see table 5, row 5.0, Gap2), a result that exceeds all the cohort 1 Choice programs displayed on table 5. One drawback of this impressive performance for the MAG RESC group was that it had only 55 urban students (see table 5, MARD_n). That is, with such a small subgroup of urban students in the group, it is also possible that special resources could be brought to bear that might not be available for larger numbers of students.

The Open Choice program also performed favorably at the Proficient level from Grade 3 of 2010 to Grade 5 of 2012 (see table 5, rows 6.0 and 6.1). Baseline profiles compared favorably to the LOCAL DIST profiles, particularly for the baseline proficiency results on MARD (47.2% proficiency at baseline compared with 43.9% for the LOCAL DIST group). Posttest results were nearly statistically meaningful for the Open Choice group, where the random group cutoff of 66.7% at posttest was nearly exceeded by a score of 66.3% Proficient (See table 5, row 6.0, OPEN CHOICE, Prof2 and row 6.1, RANDOM PstCut).

Furthermore, the gains results for the Open Choice program group were 19.1%, which was beyond the cutoff of 15.8% for the quasi-control groups (see table 5, row 6.0, OPEN CHOICE ProfGain versus row 6.1, GainCut). It is noteworthy that only 63 samples were required to obtain 30 viable samples for the Open Choice random group. This is interpreted to reflect that the lower baseline or pretest Proficient level of 47.2% for the Open Choice students is more consistent with students reflecting the background profiles exhibited by the LOCAL DIST's MARD proficiency of 43.9% (see table 5, row 6.0 versus row 2.0). This suggests that the pretest performance and background characteristics are a better match for the Open Choice students for the urban population compared with those of some of the other programs.

Table 6 reflects the same students' performance but at the Goal level rather than at the Proficient level. Note however that the performance matching for the groups only occurred at baseline proficiency. Table 6 shows that the LOCAL DIST Goal performance at pretest in Grade 3 is only 21.6% on the MARD composite index. This means that only 21.6% of the local urban students in Bridgeport, Hartford, New Haven, and Waterbury with matching valid test results from Grade 3 of 2010 to Grade 5 of 2012 met or exceeded the Goal level in both mathematics and reading on the CMT in 2010.

The pretest gap for LOCALDIST students as a group is -37.5%, which is the highest pretest gap although Open Choice approaches this at -34.4% and MAG NOT RESC at -32.5%. In other words, these programs appear to be reflections of the true LOCAL DIST population than the other Choice programs. Meanwhile, the posttest gap for the LOCAL DIST group is -41.5% with a goal of 29.8%, which is lower than all the Choice programs. Hence, the implication is that all of the Choice programs are beneficial to urban students.

Continuing to look at the column describing the pretest Gap in table 6, it is evident that the charter group in line 4.0 have the lowest baseline gap at -19.4%, but this gap has increased at posttest to -30.0%, indicating that performance ground is being lost. MAG RESC has the second smallest pretest gap at -24.6% but has the lowest gap performance at posttest, which is -14.9%.

Table 6: Cohort 1 Goal Level or Higher (Grade 3 2010 to Grade 5 2012)

Title	Samp	MARD	EdFac	SocFac	Goal1	StdPre	Gap1	Goal2	Gap2	StdPst	Cut2	Goalgain	gaincut
1.0 CITIES OUT	--	18318	9.5	17.8	59.1	--	--	71.3	--	--	--	12.2	--
2.0 LOCAL DIST	--	2496	17.4	85.0	21.6	--	-37.5	29.8	-41.5	--	--	8.2	--
3.0 CHARTER	--	184	2.7	97.3	39.7	--	-19.4	41.3	-30.0	--	--	1.6	--
3.1 RANDOM	373	185	2.7	97.9	32.5	2.8	-26.6	43.7	-27.6	2.5	48.7	11.3	16.6
4.0 MAG NOT RESC	--	353	10.5	80.7	26.6	--	-32.5	38.0	-33.3	--	--	11.4	--
4.1 RAND	132	349	10.6	88.0	28.9	1.7	-30.3	39.9	-31.4	2.0	43.9	11.1	14.8
5.0 MAG RESC	--	55	3.6	92.7	34.5	--	-24.6	56.4†	-14.9	--	--	21.9†	--
5.1 RAND	704	55	3.6	93.5	34.3	5.4	-24.8	45.8	-25.5	5.4	56.6	11.5	22.3
6.0 OPEN CHOICE	--	89	7.9	94.4	24.7	--	-34.4	36.0	-35.3	--	--	11.3	--
6.1 RAND	63	93	8.6	93.1	26.5	3.3	-32.6	37.6	-33.7	3.9	45.4	11.1	18.3

† Near empirical cut value

Empirical Cut-Value Assessment for Table 6: Table 6 shows no statistically meaningful posttest findings at Goal, implying that performance at Goal was not enhanced by any of the Choice programs over the matched quasi-random groups. However, the MAG RESC program performed at 56.4%, which is nearly exceeding the posttest cutoff value of 56.6% (see table 6, row 5.0, Goal2" and row 5.0, RAND, Cut2). This can also be observed by looking at the gap closure index, which is substantially the lowest of all the Choice programs at -14.9%, meaning that the Goal performance for MAG RESC is within 15% of closing the gap at goal. In addition, the performance gains from Grade 3 to Grade 5 for MAG RESC are also nearly exceeding the cut value (see table 6). The performance result is 21.9% gains for the MAG RESC group, while the cutoff point is 22.3%. Therefore, the MAG RESC program has been found to yield statistically meaningful posttest outcome and gains at Proficient (see table 5) and nearly statistically meaningful outcome and gains at the Goal level (see table 6).

Cohort 2 Results: Table 7 compares Proficient level performance for the Grade 6 2010 to Grade 8 2012 cohort. This cohort is of particular importance because it leads up to the critical high school years. Students who perform better during this period can potentially be better prepared for the more challenging curricular demands of high school. The pretest performance gap between 1.0 CITIES OUT (n = 19,246) and LOCAL DIST (n = 2,352) is -27.7% meaning that, all other things held equal, the LOCAL DIST schools are nearly 30% below the majority of Connecticut nonurban schools on the combined reading and mathematics performance (MARD).

Furthermore, comparing these same students at posttest in Grade 8, observe that the MARD gap increases to -31.2%, meaning that the performance gap actually increases slightly during the critical pre-high school Grades of 6 to 8. In addition, the education factors are much higher in the four target districts of Bridgeport, Hartford, New Haven, and Waterbury at 15.2% compared with 8.8% for the nonurban districts (see table 7, rows 1.0 and 2.0). Similarly, the social factors that reflect a combination of minority status and school lunch program eligibility are 84.3% of the four target district students compared with only 14.4% for the nonurban students. Again, this reflects a wide disparity between urban student populations and the nonurban student majority in Connecticut.

Table 7: Cohort 2 Proficient Level or Higher (Grade 6 2010 to Grade 8 2012)

Title	Sampn	MARD	EdFac	SocFac	Prof1	StdPre	Gap1	Prof2	Gap2	StdPst	Cut2	Profgain	gaincut
1.0 CITIES OUT	--	19246	8.8	14.4	89.0	--	--	90.8	--	--	--	1.8	--
2.0 LOCAL DIST	--	2352	15.2	84.3	61.3	--	-27.7	59.6	-31.2	--	--	-1.7	--
3.0 CHARTER	--	326	8.0	99.4	73.3	--	-15.7	81.3†	-9.5	--	--	8.0†	--
3.1 RANDOM	86	351	7.7	96.7	69.6	1.0	-19.4	69.3	-21.5	2.1	73.5	-0.3	2.8
4.0 MAG NOT RESC	--	512	9.4	91.8	69.9	--	-19.1	69.3	-21.5	--	--	-0.6	--
4.1 RANDOM	57	532	9.0	94.4	67.9	1.4	-21.1	68.6	-22.2	1.6	71.8	0.7	3.7
5.0 MAG RESC	--	96	5.2	88.5	75.0	--	-14.0	75.0	-15.8	--	--	0.0	--
5.1 RANDOM	59	95	7.4	93.8	73.3	1.9	-15.7	73.3	-17.5	4.4	82.1	0.0	6.3
6.0 OPEN CHOICE	--	76	17.1	93.4	72.4	--	-16.6	75.0	-15.8	--	--	2.6	--
6.1 RANDOM	46	88	15.9	85.6	71.2	2.5	-17.8	71.3	-19.6*	3.4	78.1	0.1	6.0

† Exceeds empirical cut value

** Figures not exact due to rounding*

Note that these background circumstances shown in table 7 are quite similar to the results for cohort1 displayed in tables 5 and 6. Most of the Choice programs continue to exhibit lower than expected educational factors. The exceptions are the Open Choice program at 17.1% EdFac compared with 15.2% for the LOCAL DIST group and, to a lesser extent, the MAG NOT RESC program at 9.4% (see table 7).

The gap index in table 7 is once again a good descriptive indicator of relative differential performance status at pretest as well as posttest for cohort 2, just as it was for cohort 1. However it is apparent from examining the pretest gap (Gap1) that there is considerable variability in baseline performance among the Grades 6 to 8 Choice programs, just as was the case for cohort 1, and once again, none of these programs is performing on MARD as poorly as the LOCAL DIST target neighborhood urban schools, which indicate a baseline gap of nearly 30% (-27.7% see table 7, row 2.0, Gap1).

It is evident from this discussion that the baseline performance for students attending choice programs is typically higher than their peers who attend schools in their home districts, although this could once again be related to prior program exposure. As we found in the Grades 3 to 5 performance comparisons for cohort 1, the charter and MAG RESC programs have the best pretest performance on MARD at Proficient. The charter schools had a baseline gap score of -15.7% and the MAG RESC group posts the lowest gap pretest score at -14.0% Proficient.

The posttest gap indicator in table 7 tells a different story. First, the difference in scores for the Choice programs are much better than those of the LOCAL DIST posttest gap performance at -31.2%, implying that all the Choice programs are making gains or maintaining performance for their students (see table 7, row 2.0, Gap2). Compared with that standard, the posttest gap is lowest for the charter group at -9.5%, which is a substantial departure from any of the other Choice programs (see table 7, row 3.0, Gap2). Both MAG RESC and Open Choice are tied for the next best posttest gap performance at -15.8% (see table 7). By comparison, MAG NOT RESC program is lagging behind the other Choice programs, showing a posttest gap of -21.5% below the nonurban majority group.

Empirical Cut Value Assessment for Table 7: The dramatic findings for table 7 show the charter schools to be statistically meaningful on the Proficient level, both at posttest and also in terms of performance gains for the Grade 6 2010 to Grade 8 2012 Cohort. The charter random group predicts an outcome or posttest score of 73.5% Proficient taking into account the standard error for the 30 matched random sample comparison groups. Meanwhile, the charter school results stand at 81.3% Proficient at posttest (see table 7), and thus the findings are statistically meaningful. None of the other programs demonstrated a meaningful departure from the empirical cut value at posttest for cohort 2. In addition, the charter group also posted statistically meaningful pretest-posttest gains of 8.0%, where the random comparison group predicted 2.8% gains after taking into account the standard error.

Table 8 (below) presents the findings for the Grades 6 to 8 cohort results at Goal. Note that the pretest gap at the Goal level for charters is lowest among all Choice programs at -24.2%, but that the posttest gap is a relatively dramatic -15.1%. The other posttest gap results for the remaining Choice programs are approximately twice as low on the gap indicator, with the next best score posted by MAG RESC at -29.4%. This presents an important piece of information. It says that the charter schools are not only achieving well at the Proficient level, but in addition program attendees are also doing much better than the other Choice programs at closing the gap at the Goal level.

Table 8: Cohort 2 Goal Level or Higher (Grade 6 2010 to Grade 8 2012)

Title	SAmpn	MARD	EdFac	SocFac	Goal1	StdPre	Gap1	Goal2	Gap2	StdPst	Cut2	Goalgain	gaincut
1.0 CITIES OUT	--	19246	8.8	14.4	73.6	--	--	75.2	--	--	--	1.6	--
2.0 LOCAL DIST	--	2352	15.2	84.3	37.8	--	-35.8	32.9	-42.3	--	--	-4.9	--
3.0 CHARTER	--	326	8.0	99.4	49.4	--	-24.2	60.1†	-15.1	--	--	10.7†	--
3.1 RANDOM	86	351	7.7	96.7	42.6	2.3	-31.0	40.9	-34.3	2.4	45.7	-1.7	3.0
4.0 MAG NOT RESC	--	512	9.4	91.8	39.5	--	-34.1	34.2	-41.0	--	--	-5.3	--
4.1 RAND	57	532	9.0	94.4	42.1	1.8	-31.5	41.0	-34.2	1.5	44.0	-1.1	2.2
5.0 MAG RESC	--	96	5.2	88.5	42.7	--	-30.9	45.8	-29.4	--	--	3.1	--
5.1 RAND	59	95	7.4	93.8	47.3	3.8	-26.3	45.0	-30.2	4.9	54.8	-2.3	6.4
6.0 OPEN CHOICE	--	76	17.1	93.4	40.8	--	-32.8	31.6	-43.6	--	--	-9.2	--
6.1 RAND	46	88	15.9	85.6	43.5	4.6	-30.1	43.8	-31.5*	4.7	53.2	0.3	9.6

† Exceeds empirical cut value

** Figures not exact due to rounding*

Empirical Cut Value Assessment for Table 8: Again, the posttest Goal results exceed the empirical cut value for the charter schools only (see table 8, rows 3.0 and 3.1). The expected performance based on the 30 random samples matched to the charter group at baseline and including the standard error is 45.7% at Goal, compared with 60.1% at Goal for the charter schools' performance at posttest. This is approximately double the expected outcome for the LOCAL DIST urban Cohort as a whole at 32.9% (see row 2.0 at Goal 2) and 20 to 30 percentage points above all the other Choice programs. Most notable is the fact that at Goal, the gains at posttest for the charter group are 10.7% (see table 8, row 3.0, GoalGain), which also exceeds in a statistically meaningful way the expected outcome for the random samples at 3.0% (see table 8, row 3.1, GoalGain). Hence, the charter group has obtained a meaningful departure from the empirical cut values at posttest and for both the Proficient and Goal levels.

Levy Gillespie
Equal Employment Opportunity Director/American with Disabilities Act Coordinator
Connecticut State Department of Education
25 Industrial Park Road
Middletown, CT 06457
860-807-2071
Levy.Gillespie@ct.gov