

# Chapter Two

## Methodology and Related Issues

### 2.1 Aims and Objectives of the Evaluation

This study examined 10 schools that have been operated by Edison Schools Inc. for at least 4 years. It was our intention to include all 11 schools that opened during the first two years that Edison was operating schools (1995-96 and 1996-97). However, one school was left out of our analyses, Dillingham Intermediate School in Sherman, TX, because we were unable to secure any independently verifiable student achievement data.<sup>5</sup> The rationale for selecting schools that have been operated by Edison for 4 or 5 years is that we believe these schools, rather than those open for 3 years or less, provide a more convincing picture of the impact the Edison model can have on student achievement. While some of these schools may have had more difficult start-ups than others, they have been in operation beyond what many consider the start-up phase.

The overriding aim of this study was to examine the impact of Edison schools on student learning as measured by norm- or criterion-referenced tests. In order to achieve this objective, a number of specific tasks were identified:

1. Review and critically assess existing research and evaluations on the impact of Edison schools.
2. Describe the evaluation measures used by Edison Schools Inc.
3. Describe the nature and quality of the standardized test results available.
4. Compare Edison schools in terms of student achievement over time.
5. Compare Edison schools with state and national norms on standardized tests.

---

<sup>5</sup> More specifically, this was a school within a school, and the available data from the state and district were not disaggregated so that we could separate the results for the Edison half of the school. Edison did provide us some of this disaggregated data, but the district claimed it did not have this data and could not confirm its validity. These data were also limited in that they did not cover all the years when data should have been available and did not include information on the number of test takers in each group. Appendix D includes a summary of the data that Edison provided to us.

6. Compare Edison schools with local school district and state performance levels and—where possible—other similar comparison schools.
7. Develop cases for each of the 10 Edison schools that include (i) a description of the school based upon available literature and documentation, (ii) findings from analysis of norm-referenced and criterion-referenced test results, and (iii) a summary of the diverse results from the analyses of test results.
8. Based upon available literature and documentation, develop a framework for analyzing the 10 cases.
9. Analyze the case studies according to the framework, and summarize the results of this analysis.

## 2.2 Sources of Information

Student achievement data were obtained from a variety of sources. Table 2:1 lists the standardized tests in which each participating school took part. The column with district and state mandated tests include all the criterion-referenced tests (CRT), and the last three columns in the table include the three norm-referenced tests that are being used in the ten schools included in this study.

From districts and state education agencies we were able to obtain results on the criterion-referenced tests for all ten schools, as well as for the local district, the state, and control schools, where applicable. We secured data for all years these tests were administered, with the exception of Mt. Clemens Secondary Academies. Because the Mt. Clemens Junior and Senior Academies share their building with the only other middle school in the district, their test results were reported as one building/school. Data provided by Edison helped us to disaggregate the data between the two schools. The analysis conducted on the CRT took place in the spring of 2000. New data that became available after this point in time were also reported in the case studies, but in only a few cases did we rerun the chi-square and odds ratio analyses to include the new test results.

Data sets containing individual student results on the norm-referenced tests were made available to us by Edison Schools Inc. We received seven such data sets. Three of these data sets did not contain results for all possible years, which limited the length of some of the longitudinal analyses. We did not receive any norm-referenced test data for four schools. For one school, we were provided with two data sets from different norm-referenced tests. While Edison didn't inform us as to the reasons for its restricted release of data, it is possible that some districts, or the company, did not wish to share such data. The data sets contained individual student data with anonymous indicators, which enabled the tracing of individual student results over time.

Table 2:1 Standardized Test Results from Edison Schools Included in the Study

SCHOOL NAME	State- or District-Mandated Criterion-Referenced Tests	Stanford Achievement Test (SAT-9)	Metropolitan Assessment Test (MAT-7)	Iowa Test of Basic Skills (ITBS)
Roosevelt-Edison Charter School Colorado Springs, CO (1996)	DALT (district test) 1996/97, 97/98, 98/99 CSAP (state test) 1996/97, 97/98, 98/99, 1999/00			1996/97 1997/98 1998/99
Henry E.S. Reeves Elementary Miami, FL (1996)	Florida Writing Assessment 1996/97, 1997/98, 1998/99, 1999/00 Florida Comprehensive Assessment Test 1997/98, 1998/99, 1999/00	1996/97 1997/98 1998/99		
Dodge-Edison Elementary Wichita, KS (1995)	Kansas Reading, Math, and Writing Assessments 1995/96, 1996/97, 1997/98, 1998/99, 1999/00		1995/96 1996/97 1997/98 1998/99	
Jardine-Edison Junior Academy Wichita, KS (1996)	Kansas Reading, Math, and Writing Assessments 1996/97, 1997/98, 1998/99, 1999/00		1996/97 1997/98 1998/99	
Boston Renaissance Charter School Boston, MA (1995)	Massachusetts Comprehensive Assessment System (MCAS) 1997/98, 1998/99	1996/97 1997/98 1998/99	1995/96 1996/97	
Seven Hills Charter School Worcester, MA (1996)	Massachusetts Comprehensive Assessment System (MCAS) 1997/98, 1998/99	1996/97 1997/98 1998/99	1996/97 1997/98 1998/99	1996/97 1997/98 1998/99
Dr. Martin Luther King Jr. Academy Mt. Clemens, MI (1995)	Michigan Educational Assessment Program (MEAP) 1995/96, 1996/97, 1997/98, 1998/99, 1999/00		1995/96 1996/97 1997/98 1998/99	1996/97 1997/98 1998/99
Mt. Clemens Secondary Academies Mt. Clemens, MI (1996)	Michigan Educational Assessment Program (MEAP) 1996/97, 1997/98, 1998/99, 1999/00			1997/98
Mid-Michigan Public School Academy Lansing, MI (1996)	Michigan Educational Assessment Program (MEAP) 1996/97, 1997/98, 1998/99, 1999/00		1996/97 1997/98 1998/99	
Washington-Edison Elementary School Sherman, TX (1995)	Texas Assessment of Academic Skills (TAAS) 1995/96, 1996/97, 1997/98, 1998/99			1996/97 1997/98 1998/99

Note: For the three norm-referenced tests, the table indicates all years of available data when conducting our analysis. The years marked in blue indicate data made available to us by Edison Schools. Where we are uncertain if test data actually exists, we have used *italics*.

Other sources of data and information that were reviewed throughout the course of the evaluation included relevant documents, research reports, and literature.

The ten schools that were included in the evaluation are listed below (the year when the school began operation is indicated in parentheses):

1. Roosevelt-Edison Charter School, Colorado Springs, CO (1996)
2. Henry E.S. Reeves Elementary, Dade County, FL (1996)
3. Dodge-Edison Elementary, Wichita, KS (1995)
4. Jardine-Edison Junior Academy, Wichita, KS (1996)
5. Boston Renaissance Charter School, Boston, MA (1995)
6. Seven Hills Charter School, Worcester, MA (1996)
7. Dr. Martin Luther King Jr. Academy, Mt. Clemens, MI (1995)
8. Mt. Clemens Secondary Academies, Mt. Clemens, MI (1996)
9. Mid-Michigan Public School Academy, Lansing, MI (1996)
10. Washington-Edison Elementary School, Sherman, TX (1995)

## 2.3 To Compare or Not to Compare

Since children grow and develop over time, we expect gains and learning to take place, regardless of the school program and even regardless of whether or not they attend school at all. The only way to separate the impact of the Edison model on students' formal learning from their learning in nonformal and informal settings is to compare students enrolled in Edison schools with students not enrolled in Edison schools. There are many ways to make such comparisons. The first method we utilized in this study was to compare an individual's achievement performance relative to the norms on nationally normed student achievement tests. By examining the relative ranking of Edison students in terms of national percentiles or normal curve equivalents, we can see whether or not they are gaining ground or losing ground compared with other students across the nation. In this study we made such comparisons using the MAT-7, SAT-9, and the ITBS.

A second way to make comparisons is to monitor the gains made by Edison students as compared with students in a local school with similar characteristics, with the district average, or with the state average. In some states, a comparison group of schools have students with similar characteristics and provide educational services at the same level (i.e., elementary, intermediate, or secondary levels). When we used comparisons in this area, we tried not to limit ourselves to one comparison group; rather, we compared against several groups. In nearly all cases the positive, negative, or

similar results are the same no matter which group is used. In this study, we use these comparisons with the state- and district-mandated tests.

In its first two annual reports on student progress, Edison strayed away from comparing its students with control groups. A noteworthy exception to this was the reports prepared by Dr. Robert Mislevy. John Chubb, the Chief Education Officer at Edison, informed us that the most important comparison was to follow Edison students/schools over time. He pointed out that gains made in the public schools surrounding the Edison schools should, in part, be credited to Edison. This may be partially true, since competition is bound to incite the local schools to improve the quality of their services. Nevertheless, the Edison schools are equally as free or more free to compete with the local schools and advance their students.

Edison claims its program leads to higher levels of achievement, and since—in a few cases—it refers to the fact that its schools are outperforming local district schools, it seems both fair and reasonable to make comparisons with control groups across all its schools. Also, Edison’s partners are expecting the schools they contract out or charter to improve student learning. It is obvious that they also expect the gains in these schools to exceed the gains made in the other local public schools. If they thought their own schools could match or exceed the gains of Edison’s schools, they probably would not have contracted out to Edison.

### Why Edison does not wish to compare

Edison’s second annual report (Edison, 1999) makes three major arguments against comparing Edison schools with other public schools.<sup>6</sup>

1. “It is often impossible to find achievement trends for students and schools closely matched to Edison students and schools” (Edison, 1999, p. 12).
2. “Because Edison schools are launched by partnership communities to raise achievement not only in the Edison school but, through healthy competition and the diffusion of innovations, in all schools in a community, it is not a straightforward matter to estimate the relative success of an Edison school. In a successful partnership the achievement gains in other community schools might not match those in the Edison school, but they should be substantial as well. A successful Edison school, then, might not build an ever-widening advantage over other local schools; all schools might progress together with the Edison school leading the way” (Edison, 1999, p.12).
3. “A statistician would not compare the achievement of the Edison school and other local schools as if each school were performing independently. The achievement of all of the schools would be modeled as ‘endogenous’ variables, the achievement of the Edison school influencing the achievement of the other local schools, and vice versa” (Edison, 1999, p.12).

---

<sup>6</sup> These arguments sum up the reasons we have heard from John Chubb, the Chief Education Officer at Edison, for why one should not compare Edison schools with others

In answer to the first point, we have identified extensive achievement data, although of varying quality, with which to make comparisons. While the control groups we have identified are not always equivalent in all areas, the expected advantages are almost always in Edison's favor. The Edison advantages are outlined below:

- Since half of Edison's schools are charter schools, and three of the ten Edison schools in our study are charter schools, these schools have the possibility of attracting families from throughout their districts who are more involved in their children's education and willing to search out and arrange an alternative to the school to which they were assigned. In the Edison contract or partnership schools, Edison negotiates that students who move out of the local neighborhood but remain within the district shall have the right to remain enrolled in their school. Therefore, parents who are actively involved will seek out, and if they move, arrange to have their children enrolled in the Edison school. Thus, selection bias should be in favor of Edison schools, unless they are perceived to be worse schools and parents act to get their children out. We think this should serve as an advantage to Edison. However, if communities and parents view the Edison program negatively, the selection bias could work against them. In our descriptive analysis of the schools, we traced changes in background demographics in the schools. From these analyses, it would appear that parents who are more likely to exercise choice are doing so in favor of Edison. Given that the differences in background characteristics of the control groups largely change over time in Edison's favor, one cannot claim that using control groups will be to Edison's disadvantage.
- A second advantage is that among the schools where we could identify student demographic data over time, there was often a slight tendency for the Edison schools to exhibit a decrease in numbers of students qualifying for free and reduced lunches compared with the control groups.
- A third advantage is that over time the Edison schools appear to exclude more students from tests than the control groups.
- The last advantage is Edison's own school model. The control groups are not equal in terms of expenditures or in terms of time at school. Edison advertises that it invests heavily in its schools (an average of \$1.5 million per school) and generally spends more money per pupil than the control groups. The Edison schools have a longer school day and longer school year than the students in the control group.

In response to the point that comparisons with local schools fail to capture the impact that Edison has had on these schools due to "healthy competition and the diffusion of innovations," we should point out that we used state-level comparisons in addition to the local district and/or control school comparison to control for this. Where possible, we also used national percentiles or normal curve equivalents (NCEs) to compare the levels of achievement in the Edison schools with national norms for students.

The argument that is made by Edison in the third point refers to implicit nesting of schools within districts, regions, or intermediate school districts. The question we ask is whether this should inhibit comparison or can it be taken into consideration in the design of the study? Important factors resulting from the nesting that would/could affect student achievement include budget/financial

allocations and local building level autonomy (site-based management of schools). Also, at the elementary level, the nesting impact is less, due to the “local” nature of each school. The nesting effect probably increases as you move from elementary into intermediate and upper-secondary schools. Because elementary schools usually include students from the immediate neighborhood, they are more heavily influenced by the local neighborhood since the schools at the intermediate and upper secondary levels are fed by many smaller neighborhood schools.

The argument that Edison’s schools contribute to gains made in local schools is not a hindrance to comparing, but rather a testable hypothesis. While the available data and the scope of this study did not permit us to fully test this hypothesis, we took it into consideration in the comparisons we made and in the discussion of findings. Since we selected control groups from the state level statistics and also made comparisons with national and state norms where available, we think we controlled for this factor in our evaluation. For example, if the performance in an Edison school and in the district goes up while the performance level of the state remains stable or declines, one might suggest that an Edison school helped raise performance levels in the district in which it resides. However, if the trends for the state and district are similar and the trend at the Edison school does not match the state and district, one might suggest that the Edison school is not having an impact on raising performance levels in the district in which it resides. As one will see in our case studies, the latter pattern is far more prevalent than the former pattern.

In summary, we agree in part with Edison’s third argument that the schools in the communities are interlinked. In half of Edison’s schools (the contract schools) there is one governance structure: the local district school board that contracted with Edison to operate one or more of its schools and which also governs the local public schools. However, in terms of policy and day-to-day decision making, there is a clear separation between the Edison schools and the local public schools in all but a few cases where the Edison school has a district employee serving as the principal administrator (this occurs in the schools within schools where the principal is a district employee but two separate vice principals are assigned to the two school entities that share the building). While the leadership of the Edison schools is based in New York City, the local public schools are governed locally. Edison also has a separate budget from the local public schools. Because the Edison schools have largely separate governance, educational programs, and budgets from the local district schools, we think that comparison with the local district is valid. In handouts prepared by Edison and in a few instances in its annual reports on student progress (Edison, 1999, 2000), Edison does state that its students are gaining more than students in local district schools. For these reasons, we believe that comparisons can and should be made between the Edison schools and local district schools.

## 2.4 Statistical Methods Utilized

We employed several different statistical methods and analyses which, taken together, provide a composite picture of student performance on standardized tests at ten Edison schools. These methods are described and discussed in the sections that follow.

## Assumptions guiding our analyses

It is important to recognize some of the more general assumptions we made in order to make the various statistical comparisons and how these affected our conclusions regarding the impact of the Edison model on student achievement. First, we utilized a wide variety of student achievement data, e.g., nationally administered achievement tests and various state- and district-mandated tests. These tests represent only a subset of possible indices of student academic performance, and in many situations it can be argued that nationally and state normed tests do not adequately describe a student's achievement level. Consequently, our first assumption was that national and state normed tests do provide a common and valid assessment of student achievement that allows for meaningful comparisons. Appendix A contains a description of each standardized test considered in the ten cases.

Secondly, we assumed that attrition rates are low and stable over time in the longitudinal cohort<sup>7</sup> analyses (panel analyses) and at a level similar to the comparison samples in the consecutive cohort analyses. Edison (1999) noted that its rate of mobility is very low (7 percent as compared with a national average of 17 percent) and indicated that this is a form of market accountability. In most of the 10 schools included in this study, we found a higher rate of mobility; however, the rates of mobility are similar to, and seldom exceed, the mobility rates of the local districts in which the Edison schools reside.

A third assumption, particularly important in the consecutive cohort analyses, was that the later cohorts would have more exposure to the Edison effect. That is, once a student enters an Edison school, he or she is assumed to be matriculating through the consecutive grades. For example, the Martin Luther King Jr. Academy (MLK) in Mt. Clemens, Michigan, became an Edison school in 1995. First graders entering MLK in that year had nearly four years of exposure to the Edison effect when the 4<sup>th</sup> grade state assessment test (MEAP) was administered in the spring of 1999. Fourth grade students taking the state test in spring 1998 had three years of exposure to the Edison program, fourth grade test takers in 1997 had only two years of exposure, and fourth grade test takers in 1996 had just completed their first year in an Edison school. Given the expectation that students enrolled in an Edison school will improve achievement levels faster than students in a traditional public school, we should see a gradual rise in academic performance on a test like the MEAP from 1996 to 1999.

We also conducted the analysis under the assumption that students enrolled in Edison schools would improve more quickly than students not enrolled in its schools. We are aware that many of the schools that Edison operates have average performance levels that are lower than those in the local schools. In fact, many of the schools that districts contract out to Edison are the lowest performing schools in their districts. Because of this, we are more interested in the value added (i.e., gain scores) over time, rather than on absolute performance levels. Among the factors that underlie our

---

<sup>7</sup> In the text of this report we use the term “cohort” to reference multiyear longitudinal trends in individual student achievement data (what is often referred to as a “panel”). We use the term “consecutive cohort” to describe groups of students who consecutively pass through a particular grade.

assumption that Edison students should demonstrate larger gains than comparable groups of students, are the following:

- Edison has a longer school day and longer school year than traditional public schools.
- Edison reportedly invests an additional \$3,000 per student, above regular per-pupil funding, when it starts each new school (Edison, 1999). In 1995, Edison reported capital investments of \$5,114,000. This figure grew to \$70,233,000 in 1999 (Edison, 1999).
- Edison has a program and curriculum that incorporate a number of research-based practices.
- Perhaps the strongest factor supporting this third assumption is that Edison claims that its schools will gain more. At conferences and meetings, Edison personnel report that its schools are making large gains, and in its second annual report (Edison, 1999) it was reported that students in schools operated by Edison are making average annual percentile point gains of 5 percent on norm-referenced tests and 6 percent average annual percentage point gains on criterion-referenced tests. Edison's third annual report suggests that the performance during the 1999-2000 school year was even better than during earlier years (Edison, 2000).

Decisions by district or charter school boards to contract with Edison are based on this assumption as are decisions by parents who chose to enroll their children in a school operated by Edison.

### Description of statistical analyses utilized

We utilized three principal statistical analyses to gauge the effect of Edison schools on student learning. However, we first present and discuss descriptive summary data for each school, identifying important school-, teacher-, and student-related factors that may have an impact on student learning. Unfortunately, we are not in the position to relate these known moderators of student learning to achievement outcomes due to the limited available building-level data and also because the nature and type of indicators vary from case to case. We include the descriptive summary for each school so that readers will have a greater understanding of the context in which the schools operate and so that readers can judge for themselves the relevance and validity of the comparison groups and the differences between the comparison groups and the Edison schools.

***Longitudinal trend analysis.*** The first analytical strategy we utilized on six cases was a longitudinal trend analysis on individual norm-referenced student achievement data provided to us by Edison. Identifying variables were coded to retain student confidentiality. The outcome variables (results on standardized achievement tests) differ by school and within schools by grade and number of connective years due to the nature of the data provided to us by Edison. A detailed description of the tests with information on the grades and subjects they cover is presented in Appendix A as well as in each school case study. A repeated measures ANOVA (list-wise deletion) was examined for longitudinal trends over the available years. Parallel analyses are reported for all types of scores reported, e.g., grade equivalent (GE), standard (or scaled) score (SS), percentile rank (or national percentile rank) (PR), and normal curve equivalent (NCE) score. In all models, the assumption of sphericity was evaluated and if found to be violated, the Huynh-Feldt adjusted p-values are reported.

If the overall effect for time was found to be statistically significant, unadjusted (alpha) pair-wise comparisons were examined to identify where a difference in the means might be located.

We received individual student results on norm-referenced tests for six of the ten schools included in the study. Most of the data sets contained results on four separate scales: GE, SS, PR, and NCE. GE and SS scores should show increases over time for all cohorts for obvious reasons: the students are maturing and learning. Likewise, grade equivalents rise as grade level placements advance each year. The important question here is whether or not the students are gaining the equivalent of one year's knowledge between two separate test administrations, which roughly occur at the same time each year. In some cases, we found that students were progressing less than one grade equivalent in a given year, suggesting that while learning is occurring, the students are not learning at a rate suggested by national norms for the tests. In other cases, we found that students' grade level equivalents were increasing more than one grade for each year, suggesting that students were advancing more quickly than the national norm. In many, but not all the schools in our study, the students started out with grade level equivalents lower than their current grade level placement.

The last two scales that we report on regarding the longitudinal data are the national percentile rank (PR) and the normal curve equivalent (NCE). The percentile rank indicates the relative rank of a student or school in comparison with the national norm. A PR of 70 percent indicates that only 30 percent of the students in the national sample scored higher and 70 percent scored at the same level or lower. Given the nature of the data we worked with, we believe that the NCE is a better indicator of a student's or school's relative status.<sup>8</sup> The NCE is a normalized standard score with a mean of 50 and a standard deviation of 21.06. Percentile ranks and NCEs have a direct fixed relationship as shown in Appendix B. NCE scores are a preferred method for measuring and comparing gains made by a school over time. Percentile ranks in a normal distribution clearly do not represent the same score scale distance between equal differences in PR values. For example, the difference between PRs of 20 and 30, or between 45 and 55, means the same in terms of percent of the normal distribution, e.g, 10 percent. But in terms of raw score distance they are not equivalent. Indeed, the z-score distance between a PR of 20 and 30 is .32 (-.84 – -.52), whereas the z-score distance between a PR of 45 and 55 is .26 (-.13 – +.13). This is even more exaggerated when one compares the same difference in PRs at the extreme end of the distribution versus those in the middle.

The relationship between PRs and NCEs is linear between PR 5 and PR 95, but at the extremes it goes curvilinear.

$$\text{NorCurPR} = 1.21 + \text{NCE} - 10.6$$

---

<sup>8</sup> Since percentile ranks are so popular, it was decided that a better scale would be one that looked like the PR scale but where the difference in raw score distance was the same for the same difference in distance on the new scale, the NCE scale. The rationale behind that was that pre- to postcomparison on an NCE score scale would be a more meaningful comparison, since a difference of 10 NCE points for one student or school corresponds with the 10 NCE points for another student or school, even if they started at different points on the pretest.

Values in normative tables for converting PRs to NCEs are obtained by determining the normalized z-score associated with the PR of interest and making a transformation of the form.

$$\text{NCE} = 50 + 21.06(z)$$

Some of the concerns that arose regarding the use of NCE include the following:

- The NCE is very close in scale and meaning with the big T-score.
- The NCE and PR are too easily confused, especially by lay people.
- Using NCE and finding PR equivalents with a formula (like what you see in published tables), assumes that the distribution of NCE scores is normal. This should be empirically verified.

A number of other strategies for analyzing the individual student data might have been utilized but were not due to the limited amount of available individual student achievement data. For example, a hierarchical linear model would be a superior method for identifying the longitudinal growth of student achievement. However, the data necessary to accomplish this type of analysis were not available. Moreover, the longitudinal analyses presented in this evaluation do not incorporate a comparison sample against which to gauge student learning, which represents a major limitation of this analysis. Nevertheless, national norms represent a point of comparison; but because we cannot control for the characteristics of the students considered in the national norms, this is somewhat limited.

**Chi-square analysis.** The second analysis strategy focused on student learning outcomes as measured by district- and state-mandated tests. The next section of this report describes these largely criterion-referenced tests. We accessed composite outcomes by grade level for schools within our sample that fall under state testing guidelines. The state tests are scored along various ordinal scales (detailed in each case study). Since these data are open to the public, we were able to construct comparison groups (detailed in each case study) for these analyses. Chi-square analyses were examined to determine if the relative proportion of students falling in the various performance levels on the state-mandated criterion-referenced tests (CRT) differed between the Edison school and the comparison group. These analyses were examined separately by year, grade level, and subtest category of the state test.

Table 2:2 Construction of 2 x 2 Tables for Odds Ratio Analysis

	Fail	Pass
Edison School	a	b
Comparison School or Comparison Group	c	d

Note: “Fail: corresponds with not meeting state standards and “pass” corresponds with meeting or exceeding expected state standards.

Although there are several possible ways to define passing, we opted to define passing and failing as specified by each state. For example, if the CRT is scored along a 4-point scale (Level 1 [lowest]

to Level 4 [highest]) and the state criteria for passing is a score of 3 or 4 then we collapsed level 4 into level 3 and level 1 into level 2 to define passing and failing respectively. It should be noted that this reclassification could mask some important gains evidenced by the students in either the Edison school or the comparison group. Appendix E contains the complete results from the chi-square analyses.

**Odds ratio analysis.** The third analysis strategy examined student learning outcomes within a prospective cohort study by analyzing the collapsed ordinal responses (pass/fail categories) on the state tests. A cohort study is when subjects are selected before they are exposed to possible determinants of interest (i.e., being in an Edison school), and their exposure to possible determinants of interest (i.e., “the Edison effect”) are then recorded along with the outcome (i.e., passing or failing the state test or in other words, meeting or exceeding state standards vs. not meeting state standards). The critical design factor in a cohort study is the comparability (similarity) of the two groups at the beginning of the time period under study. If the two groups are similar, then an observed association between being in an Edison school and passing (or failing) a component of the state test can be reasonably defended. However, if the two groups are not similar, then any observed association between being in an Edison school and passing (or failing) a component of the state test may or may not be truly a function of attending an Edison school. We constructed the 2x2 tables for these analyses in such a way to represent the relative odds for a student to fail a component of the state test (see Table 2:2).

The odds ratio (OR) (McNeil, 1996) is defined as  $OR = ad/bc$  and represents the proportion of students who fail the test in the Edison school relative to the proportion of students failing the test in the comparison school. An odds ratio can take values from zero to positive infinity. Interpretation of an OR is straightforward. An OR value of 1.00 represents equal odds for failing (or passing) relative to the comparison group. Values from 0.00 to 1.00 are representative of a “protective” effect; that is, the odds of failing are lower in the Edison school. Values greater than 1.00 would represent increasing odds for failing the test if enrolled in the Edison school. As with any point estimate, a  $(1-\alpha)$  confidence interval (CI) needs to be constructed for accurate interpretation. Thus, if the CI around the OR includes 1.00, the conventional interpretation would be that there is no statistically significant difference in the relative failing rate between the two schools (i.e., if the CI included 1.00, there is no statistically significant difference). However, if the CI does not include 1.00, the OR is generally interpreted as statistically significant, either representing a statistically significant protective effect or a statistically significant increase in the odds for failing the test. Due to the truncated nature of the sampling distribution of the OR, the standard error of the OR is calculated based on the natural logarithm of the OR, similar to converting a correlation to a Fisher’s Z before constructing a  $(1-\alpha)$  confidence interval around a correlation.

The standard error of the natural log of the OR is

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{n_1 - a} + \frac{1}{b} + \frac{1}{n_2 - b}}$$

Table 2:3 Contains a summary of the various analyses used for each of the ten cases. Each case also discusses other research and evaluation studies that have been conducted in the past. In addition to these analyses, we have provided a summary and discussion of findings for both the norm-referenced and criterion-referenced test results for each case.

Table 2:3 Types of Analyses Possible for Each School Included in the Study

School	Descriptive Analysis	Longitudinal Study	Chi-Square	Cohort Study Odds Ratio
Roosevelt-Edison Charter School	Yes	Yes	Yes	Yes
Henry E. S. Reeves Elementary	Yes	Yes*	Yes	Yes
Dodge-Edison Elementary	Yes	Yes		
Jardine-Edison Junior Academy	Yes			
Boston Renaissance Charter School	Yes	Yes	Yes	Yes
Seven Hills Charter School	Yes	Yes	Yes	Yes
Dr. Martin Luther King Jr. Academy	Yes	Yes	Yes	Yes
Mt. Clemens Secondary Academies	Yes		Yes	Yes
Mid-Michigan Public School Academy	Yes	Yes	Yes	Yes
Washington Elementary School	Yes		Yes	Yes**

\* We included the findings from Shay (2000), a dissertation that was based on the analysis of longitudinal results by Reeves' students and a control group.

\*\* Breslow-Day statistic and confidence intervals could not be calculated because we lacked information on the number of test takers.

## 2.5 Criteria for Evaluating Trends and Summarizing Results for Each Case

When analyzing and comparing our findings, we are conscious that one cannot simply count the number of positive and negative findings to come up with a conclusion. The approach we took for this evaluation focused on developing individual school cases (see Chapters 3-12), with each case developed according to the available data and subsequent analyses. At the end of each case, a short description and discussion of the relative quality of the various analyses can be found along with the overall findings from that case.

Our discussions of the relative strength of the analyses will consider the quality of the study design, the sample size, quality of instruments used, etc. In terms of judging and rating research designs, we considered the analytical framework developed by Peterson (1998). Peterson was asked by The Edison Project to compare the results presented by AFT and Edison. In his analysis, he outlined nine categories of designs, with randomized experiments rated as the "gold medal" design.

The overall availability of student achievement data varies extensively among the schools included in our study. So, too, does the quality of the available data along with the quality of the designs behind the trends, as described above. For example, some of the data we analyzed were provided to us by Edison Schools, Inc., other data were obtained from internet Web sites, and still other data came from various state and local reports. Moreover, given the varied mechanisms we utilized to garner our data, we often had to rely on the reporting agencies for data accuracy and could not verify the accuracy of the data. There are even instances where our data are incomplete due to incomplete information received from the reporting agency.

Given the range of possibilities before us, that we perceive designs that involve tracing individual students to be superior to designs that only measure change in consecutive class cohorts. In some cases we were able to conduct analyses on both longitudinal analysis of individual student achievement data and consecutive class cohorts. In these cases it is particularly salient when the two different design strategies converge in their findings and interpretation.

In our summaries of trends for each case, described in detail in the next section, we rated the various trends as negative (-1), mixed (0), or positive (+1). We included only one trend for each subject grade level test. For example, for Dodge-Edison we were able to establish trends for the Metropolitan Achievement Test. We had individual student data and so were able to conduct a longitudinal panel analysis and compare the gains made by Edison students with the national norm. With this same data, we were able to make comparisons with the district based on national percentile ranks for each subject test by grade. Because the individual student data provided a stronger design, we included the trends based on individual student data in the summary for this case, but did not include the trends based on consecutive cohorts of students in the summary (see Chapter 5 for more details). Also, when various comparison groups are available for the criterion-referenced test data, we include only one trend in our summary, based on the comparison group that is most relevant.

We believe that judgments about the overall performance of a school need to be made on a case-by-case basis. In order to limit potential bias, and in order to establish a common method of making judgments, we thought it was important to establish criteria to distinguish, first, whether or not there had been change over time, and second, if any change was positive or negative. The criteria we chose are based upon Edison's own criteria, included in its first annual report (Edison, 1997, p. 6) to distinguish when changes in achievement levels are positive, mixed, or negative. It is not clear whether these criteria were used by Edison in its second and third annual reports.

Edison's criteria served as our starting point; however, we modified several criteria and added one new criterion. In evaluating trends we use the following criteria to distinguish meaningful change and when this change is positive or negative:

- *Effect sizes (ES) or differences in effect sizes of .20 or greater.* The effect size calculated for the NRT data is the omega squared ( $\omega^2$ ) (Kepple, 1991) for a one way repeated measures ANOVA and only provides the reader with an overall effect for time. It does not adequately convey the direction of change nor if the change occurs all in one year or is reflective of a gradual cumulative gain.

- *Differences in national percentile scores of 5 percentage points or more per year*
- *Differences in percentage proficient scores of 5 percentage points or more per year*
- *Differences in grade equivalents of 2 months or more, and annual gains in grade equivalents of 14 months or more per year*
- *Differences in DALT gain scores of 2 points or more per year*
- *Differences that are statistically significant (at the .05 level) when tests of significance are available.* The p-value criterion is only applied to the CRT data and not to the NRT data. Utilization of a p-value criterion in the longitudinal NRT analyses does not adequately convey the direction of a statistically significant change. That is, in a longitudinal analysis there can be a statistically significant change in both directions; therefore, the p-value is ambiguous relative to the direction of change, only the presence of change. However, in the chi-square and OR analysis, the p-value conveys a meaningful difference due to the configuration of the contingency tables, in that for these analyses the reference is to the comparison group.
- *Differences in normal curve equivalents of 3.5 or more per year*

Part of the technical complexity of this report is a function of the variety and large number of analyses conducted. We have constructed summary tables for the reader's benefit that help guide and focus the reader in distilling the overall impact of Edison in a given school. First we treated each analysis category (NRT, CRT) separately. Within each analysis category, we have rated a finding as negative (-1), mixed (0), or positive (+1) based upon the guidelines presented above. A negative finding would be an effect that meets one of the above criteria but in the opposite direction and should be relatively unusual. For example, in an NRT analysis on NCE, a negative finding would be a reflective of a 3.5 NCE drop per year over the number of years covered by the analysis. In a CRT OR analysis, a negative finding would be a statistically significant (p-value criteria) risk of failing the test relative to the comparison sample. A mixed finding would be reflective of grade-level improvement in an NRT, or even odds in an OR analysis. A positive effect could be illustrated by an average annual gain in NCEs of 3.5 points per year or more over the life of the analysis, or an OR that is statistically significant and protective. To this general scoring system we tried to determine if a trend was present when there was more than one score scale present, e.g., NRT data, or more than two years of data were present, e.g., OR analysis.

We also based our rating on a prioritized hierarchy of data. We consider a trend in NRT data to reflect the findings of a longitudinal panel of students as they progress through the life of the analysis by subject and grade. A trend in CRT data reflects the consecutive cohort findings for a specific grade and subject test over the life of the analysis.<sup>9</sup> Although we have calculated outcomes relative to various comparison groups (e.g., national, state, district, or other), we count only one trend in the combined table. For NRT data we prioritized the analyzes as follows: we considered the NCE trend

---

<sup>9</sup> By contrast, Edison counts trends in one-year change segments so a trend of data for a cohort of students over four years would be counted by Edison as three different trends, while we would count this as one trend and base our rating on the change over the life of the trend.

first if available, followed by the PR or NPr, then GE, and lastly SS. For CRT data we counted each grade and subject test separately based on the outcomes of the OR Breslow-Day findings relative to the district data.

Each case was then summarized by combining the NRT and CRT ratings into one table to derive an overall school rating. In its 2000 annual report, Edison defined the 5-point scale they used to rate the overall trends in its schools (Edison assigns one to five stars for each of the categories, from Strongly Negative to Strongly Positive, respectively). Its cut points are as follows:

- Strongly Negative when 0-19 percent of the trends are positive
- Negative when 20-39 percent of the trends are positive
- Mixed when 40-59 percent of the trends are positive
- Positive when 60-79 percent of the trends are positive
- Strongly Positive when 80-100 percent of the trends are positive

Since we considered all the trends and did not focus on the positive trends alone, we calculated a mean across the trends where a negative trend is equal to -1, a mixed trend is equal to 0, and a positive trend is equal to +1. We then applied the following 5-point rating scale to the mean trend:

- 1.00 to -0.60 corresponds with “Strongly Negative”
- 0.59 to -0.20 corresponds with “Negative”
- 0.19 to +0.19 corresponds with “Mixed”
- +0.20 to +0.59 corresponds with “Positive”
- +0.60 to +1.00 corresponds with “Strongly Positive”

For example, in Dodge-Edison (see Chapter 5, Table 5:7) we report a total of seven trends, four NRT and three CRT. Based on the criteria listed above, of the four NRT trends, two are positive (+2) and two are mixed (+0). Of the three CRT trends, 1 is positive (+1) and two are mixed (+0). Averaged together, we rate Dodge-Edison as a Positive school with a mean rating of 0.43.

## 2.6 Limitations of the Study

Several inherent limitations in this evaluation needed to be examined in order to provide a balanced interpretation of the findings we reported and the conclusions we have drawn. The limitations to this study can be grouped into three areas: methodology, data quality, and conceptual limitations. In this section we highlight what we think are the major limitations of the study that should temper all conclusions derived from this evaluation.

## Evaluation of schools based on student performance alone

Ideally, evaluations of schools should not be based on student performance data alone. Nor should such evaluations be based on measures of market accountability, such as head counts. We believe that evaluations of schools, such as those included in our study, should be based on measures of market accountability, performance accountability, and regulatory accountability. Clearly, this evaluation is focused only on performance accountability. As suggested in the title of this report, we are evaluating the performance of students enrolled in Edison schools and not the Edison schools themselves.

## Selection of Edison schools included in this evaluation

There is a possibility of selection bias related to the schools selected for the evaluation, in that it might be argued that the schools we studied were either performing more poorly or superior to nonselected Edison schools. We examined this possibility by conducting a chi-square analysis on the school ratings published in Edison's 2000 annual report for the 10 schools in this evaluation relative to the remaining 32 schools included in the report. Edison rates each school on a 5-point scale, from Strongly Positive to Strongly Negative. Our analysis indicates that there is no indication that the 10 schools we included in this study are rated by Edison any differently from the 32 schools that opened during or after 1997 and for which it reported trend data in its 2000 annual report. Thus while there may be some selection bias in our sample, there is no strong indication that the schools we evaluated are not an accurate representation of the schools for which Edison currently has trend data.

## Lack of a comparison group in the longitudinal analyses

One of the principal advantages of this evaluation also suffers from one of the major limitations. Edison Schools Inc. provided us with seven different test data sets that included individual student achievement data from a variety of different standardized achievement tests, i.e., SAT-9, MAT-7, and ITBS. These data sets covered six schools and varied in quality and quantity. Nevertheless, they allowed us to examine longitudinal trends based on individual student data in six of the ten schools in our study. Unfortunately, Edison was not in a position to include data on similar schools for comparisons. Although these longitudinal analyses provided the most complete picture of student achievement in an Edison school, there is no comparison group to gauge gains and losses against. Thus, we were left with comparisons against national norms and interpretations focused on grade equivalent scores and NCEs.

## Composition of comparison groups used in the chi-square and odds ratio analyses

The primary purpose of this evaluation was to develop a composite understanding of the effect of attending an Edison school on students' achievement. In order to do this, we needed to make performance comparisons on state-mandated testing programs relative to some comparison group; that is, we needed to construct suitable comparison groups. Much of the state-mandated data was

extracted from Internet web sites, and there was considerable variation in the quantity and quality of data from state to state. Moreover, we decided to try to define our comparison groups in such a way that the data sources would represent the same quality of information and thus have similar meaning from state to state. This reasoning, unfortunately, resulted in comparison groups that were at the least sophisticated end of the spectrum, e.g., district and state. Obviously, superior comparison groups are at least theoretically possible to identify. However, even if identified—for example, equated on gender, grade, SES, mobility, teacher experience, teacher mobility, etc.—identifying the necessary data to construct the contingency tables may or may not be readily available. For example, different states report different resolutions of performance data. Thus, it was extremely difficult to identify a common, minimum set of variables to begin developing more sophisticated comparison groups. Consequently, the validity of our comparison groups can be questioned. For example, if student mobility in the Edison school is dramatically higher or lower than in the district or state, then the longitudinal comparisons may not reflect the true magnitude of the “Edison effect.”

### Variability and completeness of Web-based reporting of the district- and state-mandated testing results

Much of the data we used in the chi-square and odds ratio analyses was extracted from the Internet. Much of these data, perhaps all, had undergone significant data filtering and cleaning by the various state agencies reporting the data. We cannot be sure that there were no data posting errors at these web sites. The data we extracted and analyzed cannot be checked for accuracy beyond what is posted on the Internet. If there were data reporting errors (we know there are rounding errors), we cannot identify or examine the data for possible bias resulting from this possibility. All we can do is assume that posting errors on the Internet, if any, are randomly distributed across the various Internet sites we used. In the event that there are posting errors, any impact would be to increase the background noise in our analyses, thus making it harder to detect differences among the groups.

### Marginal cooperation of Edison in supplying individual student achievement data, regardless of type

As stated above, Edison provided some individual student achievement data. These data, however, were very inconsistent and often incomplete. In many cases we were given only two or three years of data; yet it is documented (see Table 2:1) that much more data are available. Even without a comparison sample, the longitudinal analyses would be stronger had all available data been utilized. While promised access to the individual student data in July 1999, we did not receive any data files from Edison until the end of November 1999. During that time we expended considerable time and money attempting to build data sets with individual student data for the MEAP or to secure data sets containing individual student results from other sources. None of these efforts resulted in any usable data sets containing individual student data.

## Limited resources

The budget for the study was limited in size. NEA contracted with The Evaluation Center, but the Center ended up covering 15 percent of the expenditures for the study. During the latter part of the study, the authors contributed their own time since allocated funds for the study had been expended. Conducting evaluations of this nature with limited budgets leaves little room for errors in planning or exploration of alternative sources of data or methods of analysis. Additional resources for the study would have allowed us to conduct a meta-analysis and would have helped us include additional outside experts. Nevertheless, the evaluators are grateful for the interest and advice of a number of persons outside the Center who contributed generously with their time and expertise.

## Controversial and polarized nature of Edison Schools Inc.

Evaluations of this sort are made more complicated by the controversial and polarized nature of the reform (i.e., private, for-profit operators of public schools) and the strong vested interest of many of the stakeholders.