Examining Productivity Series

Academic Productivity of Chicago Public Elementary Schools

A Technical Report Sponsored by The Consortium on Chicago School Research

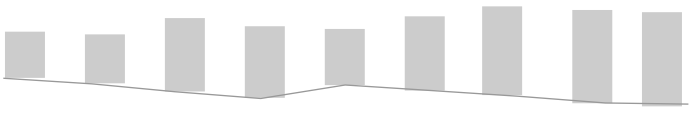
March 1998

Anthony S. Bryk
University of Chicago

Yeow Meng Thum
University of California at Los Angeles

John Q. Easton
Consortium on Chicago School Research

Stuart Luppescu
Consortium on Chicago School Research



Academic Productivity of Chicago Public Elementary Schools

A Technical Report

Table of Contents

Acknowledgments

Many people have been involved in this project throughout its long history. G. Alfred Hess, Jr., Research Professor at Northwestern University, proposed this project in 1990 as a joint effort of the Chicago Panel on School Policy and the Center for School Improvement at the University of Chicago. With the financial support of the Spencer Foundation and the assistance of the Department of Research, Evaluation and Planning of the Chicago Public Schools, work proceeded for several years. With the conclusion of the Spencer grant to the Chicago Panel, the Consortium on Chicago School Research assumed responsibility for completing the project. Although Fred Hess continued to advise on the study, the authors alone are responsible for any errors of fact or interpretation.

At the Consortium, we would like to thank Sandra Jennings for her expertise in drawing the figures and desktop publishing the report. We thank Kay Kersch Kirkpatrick for her attention to detail in managing the production of the productivity report, and we appreciate Rebecca Williams' editing and rewriting skills.

Staff members of the Chicago Panel on School Policy, especially Jesse Qualls, also contributed greatly to this study.

At the University of Chicago, we wish to thank Ben Wright on the design of the equating study and for the numerous hours he spent guiding us through the detailed analyses. David Kerbow conducted much of the research on mobility that helped us understand some of the complexities of this study. Many people who contributed along the way have since moved on. We especially wish to thank Ong Kim Lee and Paul Dean for their help over the years.

We wish to specifically thank Julia Smith (Western Michigan University) for the development of the mathematics measurement ruler used in this study. Additional work by her on pacing and content in mathematics instruction will be forthcoming.

At the Chicago Public Schools, we greatly appreciate the early support and ongoing assistance of William Rice and Carole Perlman. We would also like to thank John Delmonte, Daisy Garcia, Cynthia Gonzalez, Andrea Ross, the late James Stewart, Sandra Storey, and Peter Wallin for their help.

Introduction

The past decade saw two major changes in the governance and operations of the Chicago Public Schools (CPS). The Chicago School Reform Act of 1988 devolved substantial resources and authority to local schools and made them responsible for their own improvement. This law established locally elected school councils with authority to evaluate and select the school principal, and devise an annual School Improvement Plan and budget. Increased discretionary monies, provided as part of this legislation, have fueled local improvement efforts including hiring additional staff; purchasing instructional materials, equipment, and textbooks; and increased professional development activities.

Beginning in 1991, the Consortium on Chicago School Research initiated a number of critical probes of Chicago's decentralization reform. Our early work focused primarily on how teachers and principals in elementary schools reacted to this reform, how they used the opportunities it provided for local improvement initiatives, and the constraints they encountered in advancing school change. Over the last three years, we brought more intense scrutiny to reform of the city's high schools. In both cases, we adopted a strong formative orientation seeking to assist both school community leaders and systemwide policy makers. We have sought to chart the progress of this reform and to advance the public conversation about additional changes needed if this reform is to culminate in major improvements in educational opportunities for children.

More recent state legislation in 1995 added a new dimension to reform—it restructured the central office. The legislation created a corporate style management team, including a chief executive officer, who replaced the position of superintendent, and a Reform Board of Trustees, who are now directly appointed by the mayor. This law brought greater central accountability by clarifying the powers of the chief executive officer to deal with non-improving schools. As the system has moved aggressively to use these new powers to place over 100 schools on probation and to reconstitute some of the most problematic among them, the need to accurately identify

failing schools has become more critical. To date, the system has relied primarily on a simple statistical indicator—less than 15 percent of the students above national norms on the Iowa Tests of Basic Skills (ITBS)—for this purpose. While the CPS's efforts to intervene in failing schools have been generally lauded, criticisms have been raised about the specific criterion used.

Purpose of this Paper

Looking back to 1988, it is very clear that the Chicago Public Schools needed deep and profound changes. While there were a few pockets of excellence, taken in total it was a school system organized for failure. The 1988 Chicago School Reform Act banked on expanded local participation to challenge this dysfunctional status quo and to promote structural change at both the individual school and the system level. While reformers recognized that major changes in student learning might not come quickly, the ultimate bottom line for reform was improvements in academic achievement.

Thus, the aspirations for the 1988 Reform Act as well as the more recent accountability efforts of the central office indicate that the CPS needs a credible system for charting academic improvement. As we demonstrate below, the annual systemwide reports of student test scores, while of great public interest, are crude and sometimes seriously biased indicators for making judgments about the productivity of individual schools. For this reason, several Consortium staff and affiliates, under the initiative of the Chicago Panel on School Policy, have been working for a number of years on better ways to analyze and report standardized test score data for examining the academic productivity of the Chicago Public Schools.

This report uses ITBS scores for all students in grades two through eight from 1987 to 1996. In half of the schools, where local school councils had the opportunity to choose their own principal in 1990, these data represent six-year trends in student learning under reform. For the other half of schools, who had the opportunity to select a principal in 1991, these data represent five-year trends. In both cases, sufficient time has been afforded for significant organizational changes to occur. A body of evidence has finally been assembled that makes it now possible to investigate seriously time trends in school academic productivity.

This report differs from others distributed by the Consortium in that it is more expository in tone and somewhat more technical. We detail a set of weaknesses in the current CPS testing and reporting system, and develop an alternative approach, called a *school academic productivity profile,* for sum-

marizing the changes that have occurred in a school. The core of this approach entails estimating the value that a school adds to the learning of students taught at that school.

This report also initiates our "Examining Productivity" series. It is the first in a series of studies that will systematically examine the academic productivity of Chicago's public elementary schools. This report develops a productivity profile for each school and uses these to summarize the systemwide trends over the decade from 1987 to 1996 in reading and mathematics achievement. Subsequent reports will use these same data to investigate the characteristics of schools that have been especially effective in their academic improvement efforts.

The term *academic productivity* has a very specific meaning in the context of this report series. It refers to the contribution a school (or group of schools) makes to the learning of students receiving instruction in that school. *Improving academic productivity* means that the contribution to students' learning is increasing over time. We detail later in the report that this is the most appropriate standpoint for school accountability. To be clear, improving academic productivity does not necessarily mean high test scores. If a school enrolls a large proportion of weak students, the school may contribute a great deal to their learning, but overall test scores may still be rather low because of the limited preparation that these students bring to the school.

Before forging ahead an important caveat is in order. The analyses presented here, and in subsequent reports, are the best we can offer given the limitations of the available data. We emphasize at the outset that these data limitations are considerable. This report concludes that the CPS needs a better testing and reporting system in order to have a more accurate basis in the future for charting academic productivity. The Consortium's Steering Committee offers a number of recommendations to frame these future developments.

A Weak Indicator: Problems with Percentage of Students at National Norms

Different statistical indicators are needed for different purposes. An indicator that is useful to describe student achievement across the whole system may not necessarily be well suited for examining individual school productivity and improvements (or declines) in that productivity. The Chicago Public Schools have used a variety of statistics over the years for reporting student achievement. These include median grade equivalent scores, median percentile ranks, and "the percent of students scoring at or above national norms." Recently, this latter statistic has been

used to make important decisions about individual schools, including whether they are put on academic remediation or probation.

The percent of students scoring at or above national norms was first calculated in response to the 1988 Reform Act, which mandated a goal for each school of academic achievement "that equals or surpasses national norms." While this statistic does indicate a very real systemwide gap from the national norm, it can be problematic when used to judge changes in the

Figure 1a. Initial Distribution of Student Scores

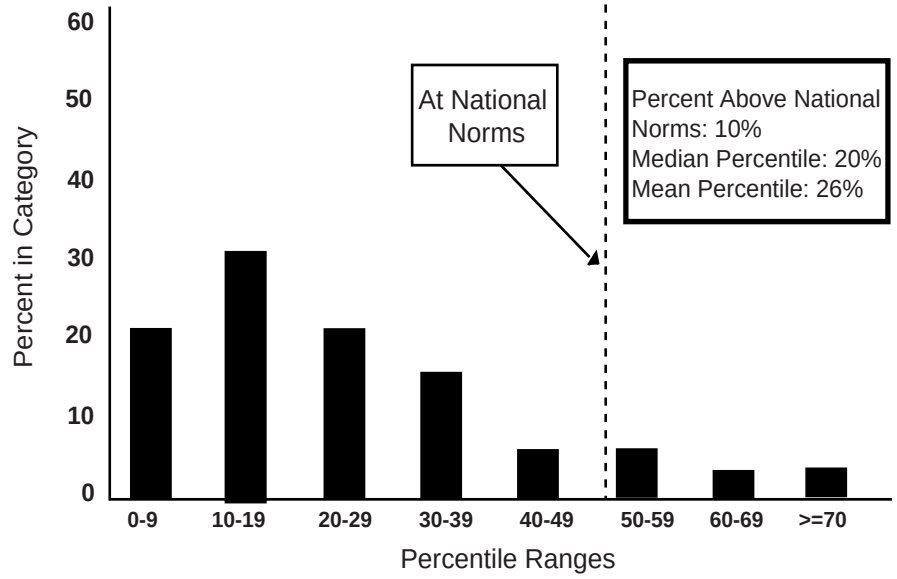


Figure 1b. Distribution of Student Scores after a Broad-Based Intervention

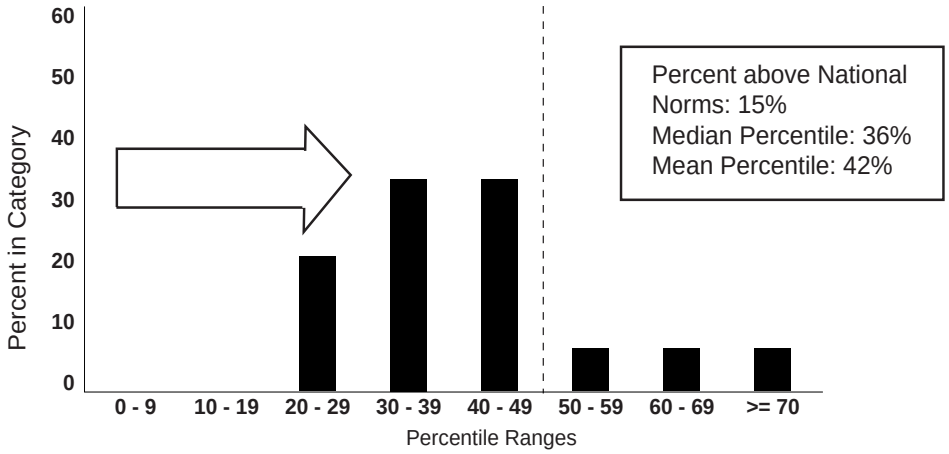
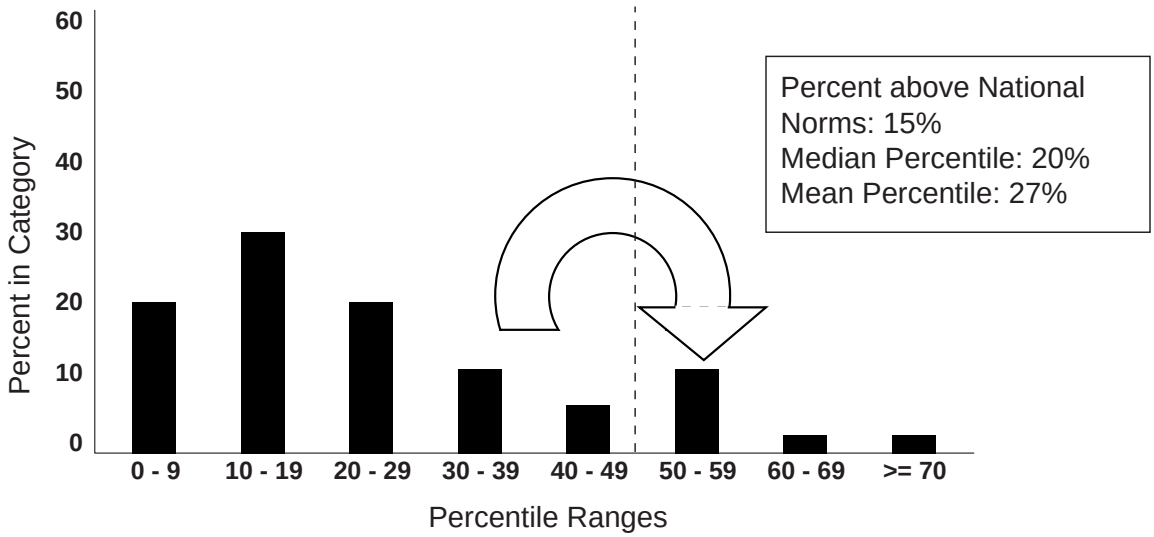


Figure 1c. Distribution of Student Scores after a Narrow-Based Intervention



performance of an individual school. The major concern is that this statistic is responsive to changes in the performance of only a subgroup of students—those who cluster close to national norms. Significant improvements in the learning of very low achieving students, for example, in the 10-30 percentile range, can go undetected. This is problematic since many Chicago schools enroll large numbers of such students. We note that the same issue arises for improvements among higher achieving students, for example, in the 70-90 percentile range. These changes also would go unrecognized.

We demonstrate the problem with a simple illustration. Figure 1a presents a profile of test scores for a low achieving elementary school with students arrayed within 10-point percentile ranges called deciles. Let's consider two possible school improvement scenarios. In the first case, a broad-based intervention is put in place that affects the achievement of all students, with more attention, however, focusing on the lowest achieving students. As a consequence, students who were originally in the lowest three deciles moved up about 20 percentile points. All other students improved by 10 percentile points. (See Figure 1b.)

In the second case, a much narrower intervention was attempted focusing only on students in the fourth and fifth deciles (i.e., the 30-49 percentile range). While this intervention was successful in moving many of these students toward or above the threshold of national norms, the vast majority of students in the school remain unaffected. (See Figure 1c.)

In terms of the indicator of percentage of students at or above national norms, these two cases are indistinguishable—both improved from 10 to 15 percent! Although these two interventions are very different in terms of their consequences for students, the principal criterion currently used by the CPS for accountability purposes would not distinguish this.

Other statistical indicators can do a better job in this regard. The median percentile for the school is a somewhat better statistic because it clearly points out the large improvement in the first case (Figure 1b) from the 20th to the 36th percentile. This statistic, however, does not detect the small improvement that did occur in the second case. An even better statistic for this purpose is the school mean achievement (i.e., the simple average of all students' test scores).[1] It correctly detects both the large improvement in the first case and the small improvement in the second case. This occurs because the school mean achievement indicator is sensitive to the performance of all students. Any changes, even small ones, will be reflected here. We build on this idea in a subsequent section when we introduce a *value-added indicator of school productivity*. This indicator, which assesses the contributions that a school makes to students' learning, is based on the mean learning gains for all children receiving instruction in a school in a given year. Here, too, the performance of each individual student affects the final results.[2]

Need for a Stable Measurement Ruler over Time: Problems Associated with Nationally-Normed Standardized Tests

The ITBS is the main achievement data gathered annually by the Chicago Public Schools and is the sole information source currently used by the system for school accountability purposes. These tests are inexpensive and relatively easy to administer and score. They are quite useful for the purposes for which they were originally intended—to tell us about how well our students perform against a national sample of students who took the same test. They were not, however, specifically designed for the purposes we now use them for—to assess improvements in schools' productivity over time.

By way of background, the ITBS is not a single test, but rather a testing system. It consists of multiple forms that were developed at different points in time. These forms are literally different tests with no overlapping items. Each form consists of multiple levels, each designed to be administered to students at a particular grade. For example, level 9 is designed for grade 3, level 10 for grade 4, and so on. Although it is

now an infrequent practice in the CPS, students sometimes have been tested "off level," such as giving level 8 to a very disadvantaged third grader or level 10 to a gifted student at the same grade.

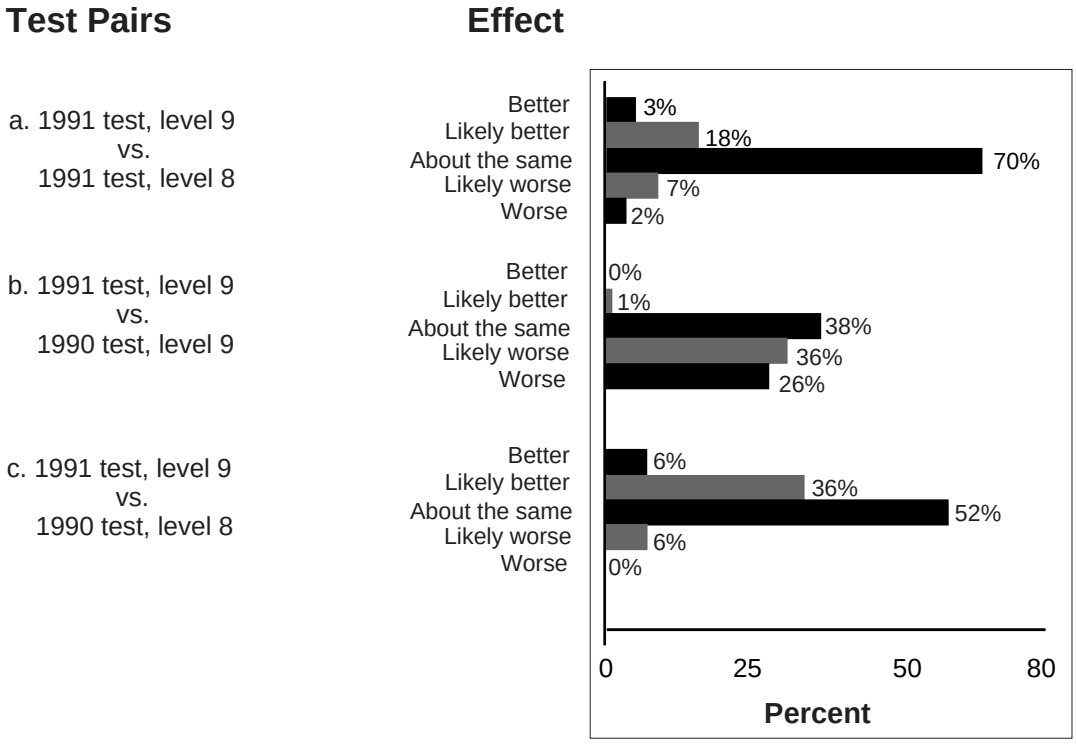
The Non-Equivalence of Grade Equivalence

The ITBS, like most nationally norm-referenced standardized tests, produces a score report called a grade equivalent (GE). GEs have a great deal of appeal to teachers and parents because they appear to describe a child's performance in developmental terms of grade level and months within grades. Since the CPS administers the ITBS in the eighth month of the school year, a fourth grader's score of 4.8 is "on grade," "at grade level," or "at the national norm." Similarly, a fifth grader who tested at grade level is assigned a GE of 5.8, a sixth grader who is on grade scores a 6.8, and so on.

Since all of the test forms and levels produce GEs, the lay user might easily think that these results are equivalent and directly comparable. In fact, this is not true. To demonstrate the problems here, we gave a sample of CPS students two different reading and math tests from the ITBS series. In one case, we administered adjacent levels (8 and 9) from CPS91 (Form G), which was administered systemwide in 1991. In a second case, we administered the same level of the test (level 9) but from two different forms, CPS90 (Form J) and CPS91 (Form G), which were used in 1990 and 1991. Finally, in the third case we changed both the level and form. A sample of students took both test level 8 of CPS90 and level 9 of CPS91.[3] The latter case is interesting because it directly represents what CPS students actually experience. That is, as students progress across the grades, they normally change test levels each year. In addition, since 1990, the CPS has been changing the form of the test administered each year. Thus, as we consider the year-to-year progress of students over time, we are actually comparing data from two different forms and levels.

A basic criterion for comparing data from any testing system is that students' score reports should not depend upon the particular form or level of the test taken so long as it is appropriate for their general ability. Thus, if we give a child two different tests, we expect similar estimates of that child's competence. While some children might do a bit better on the first test, and others might do somewhat better on the second, on average the two tests should tell us the same thing.[4] Figures 2a, 2b and 2c demonstrate, however, that this is not always the case with grade equivalent scores from the ITBS reading assessments. For example, students who were given CPS91 (see Figure 2a) were more than twice as likely to have better GE scores on the higher level test (level 9) than on the lower level test (level 8). Similarly,

Figure 2. GE Test Score Bias Due to Form and Level Differences



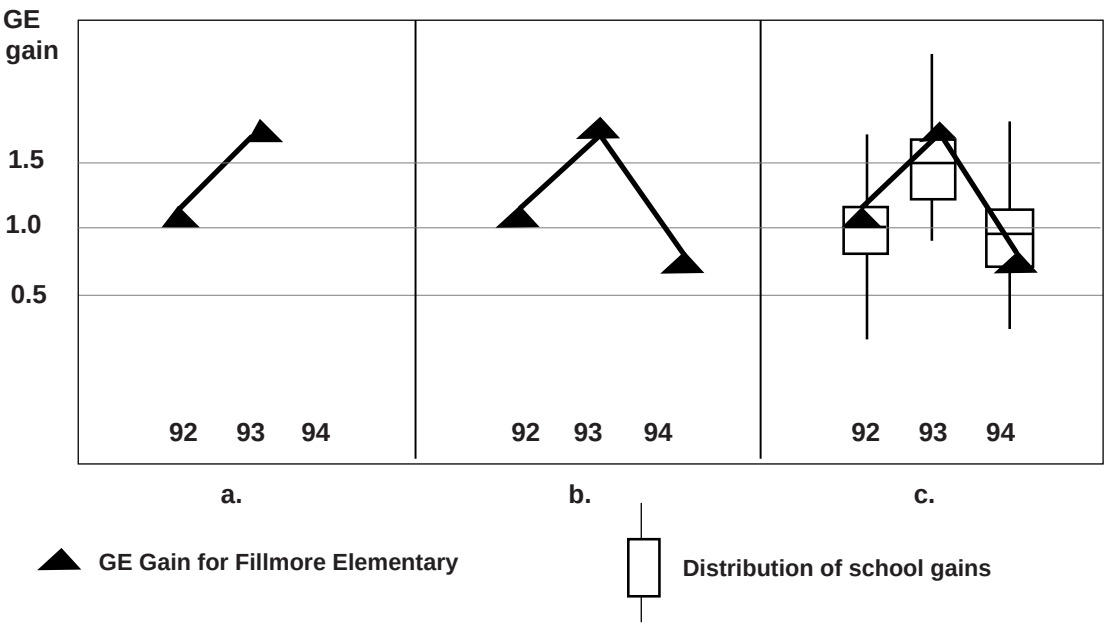
Note 1: *About the same* category is +/- 1 standard deviation from zero.
Note 2: See endnote 8.

consider the students who took the same level of the test from two different years (Figure 2b). These students were much more likely to do better on CPS90 than on CPS91. These differential score effects are equally dramatic when we consider the comparison across forms and levels (Figure 2c). Students were seven times more likely to score higher on CPS91, level 9 than on CPS90, level 8.

These empirical examples illustrate a general problem that grade equivalents are both form and level specific and can not be strictly compared. Clearly, this limits our ability to make accurate statements about how much actual learning an individual student is making over time. It also introduces a great deal of uncertainty into any assessment of whether scores may be going up or down over time for an individual school or across the whole system. While real changes in student performance are embedded here, so are the differences in the test scoring.

Figure 3 presents a clear example of the problems that this can produce when we try to interpret grade equivalent scores to assess progress over time. We illustrate the GE gains made by seventh grade students in "Millard

Figure 3. Trends in Reading Gains: A School Effect or Measurement Artifact?

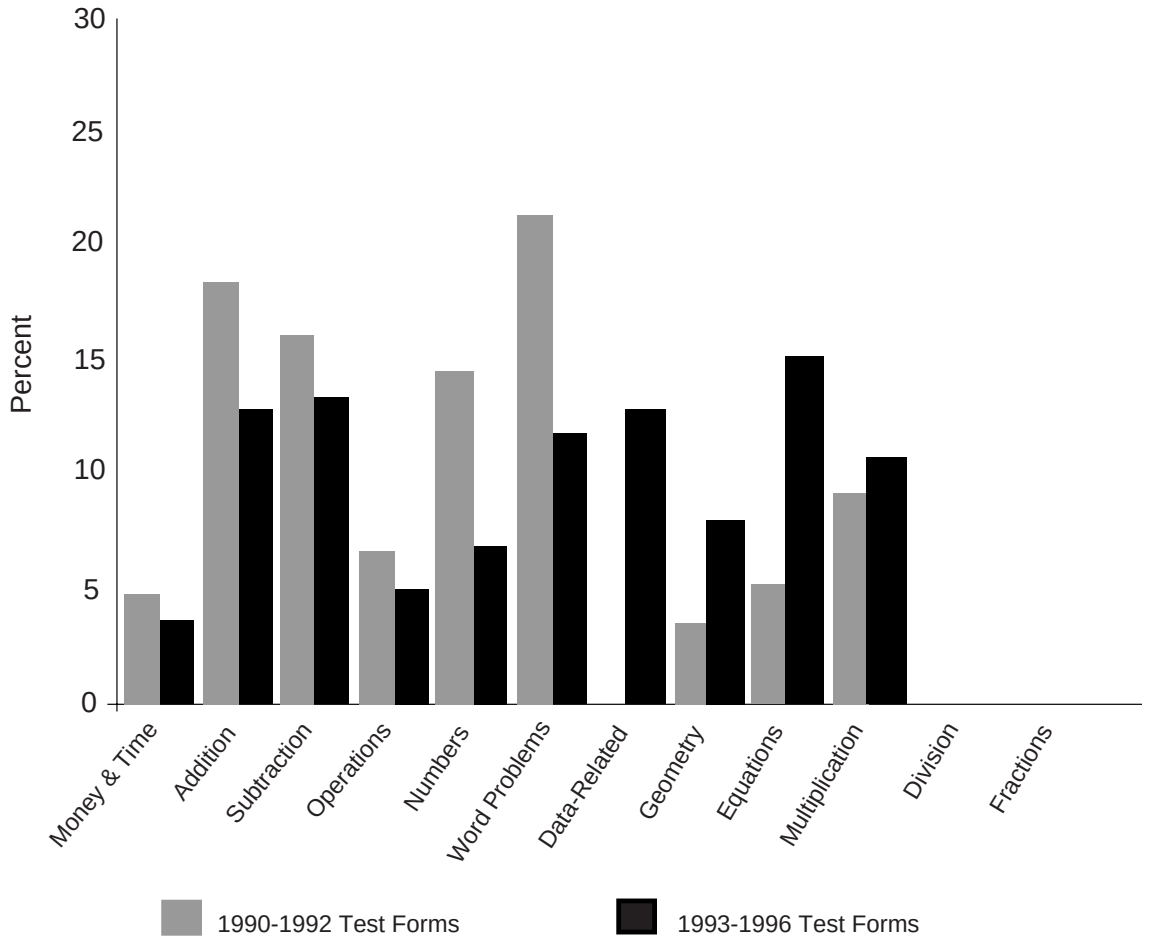


Note: Figure 3c uses a box plot to display the distribution of school gains. The area inside the box represents gains for half of the schools; the top whisker represents 25 percent of the schools with the greatest gains, and the bottom whisker 25 percent of the schools with the lowest gains.

Fillmore Elementary School" in 1992, 1993, and 1994.[5] The seventh graders in 1992 gained approximately 1.0 grade equivalents over their end of grade six performances (see Figure 3a). The following year, seventh graders gained 1.7 GEs—an improvement of 70 percent. In 1994, however, student gains fell back to 0.7 GEs—worse than where they started two years earlier! Why did this school suddenly lose the productivity improvement from the year before? What went wrong?

In fact, it is quite likely that nothing went wrong in 1994, and probably nothing went right in 1993 either. This pattern of results is not distinctive to Fillmore; it occurred generally across the entire school system. Figure 3c presents a set of box plots that displays the seventh grade gains in these same three years for all Chicago elementary schools. Notice that in most schools, seventh grade gains went up in 1993 and then fell back down in 1994. The median CPS elementary school went from 1.0 GE gain in 1992 to a 1.5 GE gain in 1993 and then back down to 0.9 GE gain in 1994. While Fillmore students gained a bit more in 1993 and lost a bit more in 1994, their results closely follow the overall system trend.

Figure 4a. ITBS Mathematics Content Changes: What the ITBS Tests in Grade 3

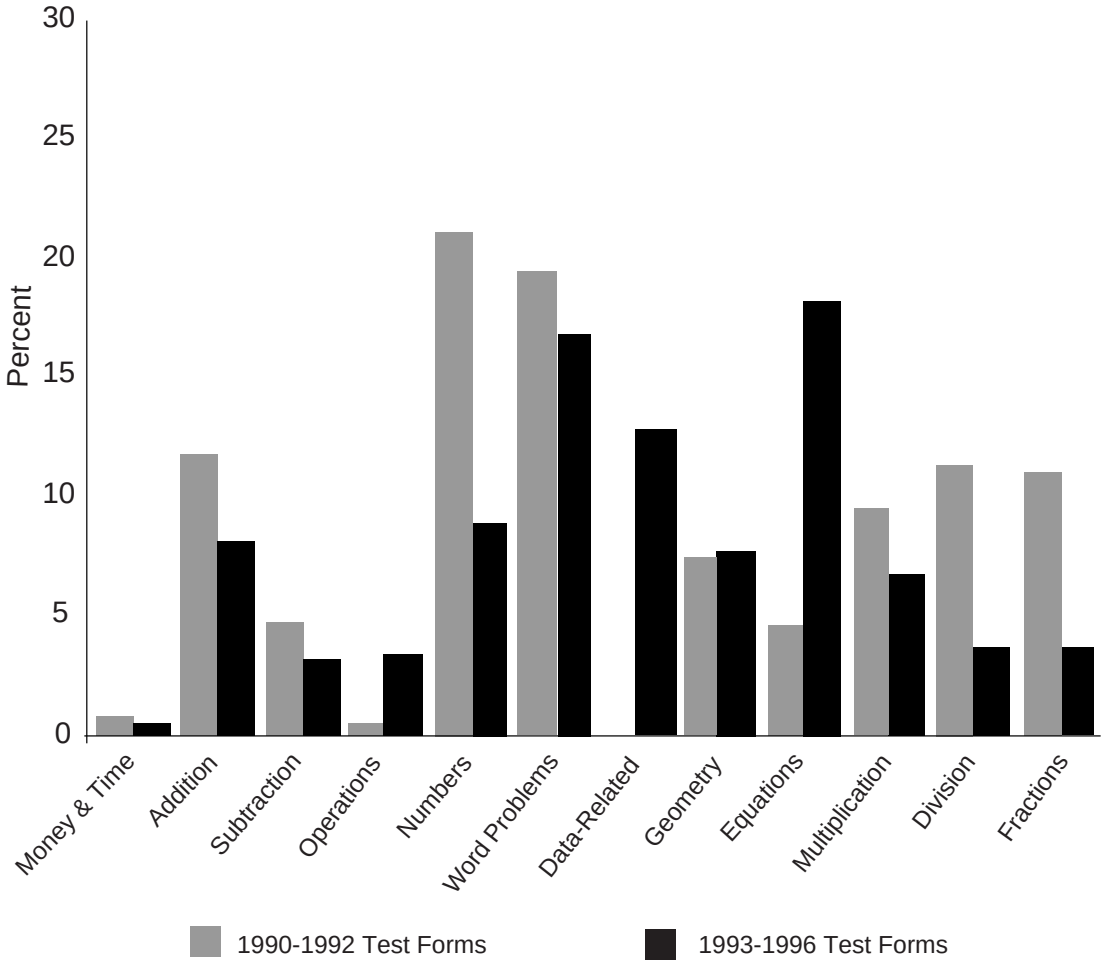


Unfortunately, many educators and most of the public are unaware of these inherent limitations in the grade equivalent metric. These score reports are simply not designed for purposes of making inferences about change over time. Clearly, a better reporting metric is needed if we wish to assess accurately whether school productivity is improving.

A Non-Standard Standard

A second problem with the use of the ITBS for productivity analysis emerges when we consider the actual content of the tests. The skills assessed by the ITBS have changed over this 10-year time period. Thus, when we look at 10-year trends in score reports, we are, in essence, judging students, schools, and the system against a moving target. Unfortunately, this changing target is largely hidden in a secure test and unknown to most educators. As a

Figure 4b. ITBS Mathematics Content Changes: What the ITBS Tests in Grade 6

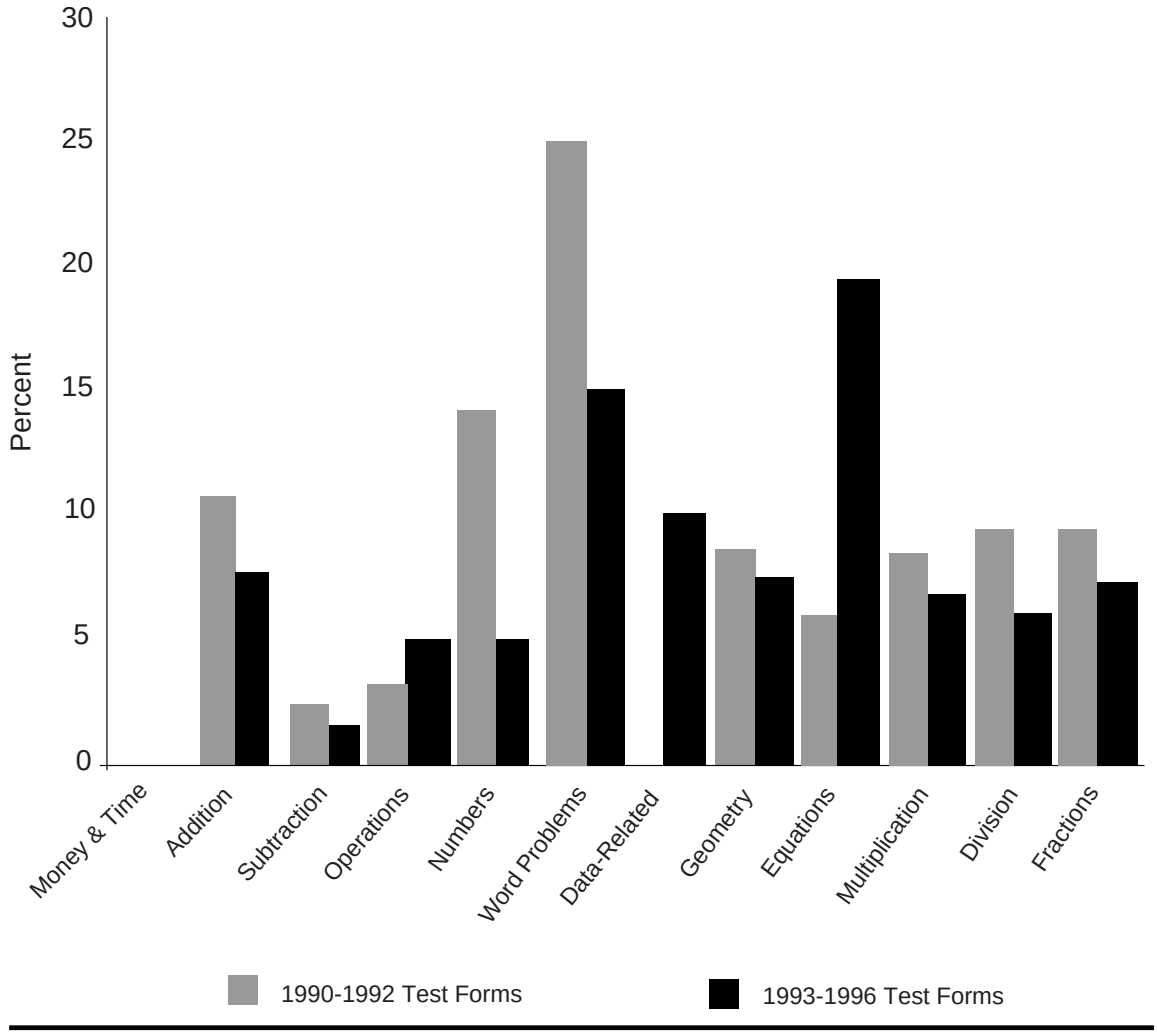


result, a teacher may see, for example, that students in her classroom clearly know more mathematics than previous classes of students, but their standardized test scores may still come back lower.

To document this problem of changing standards, we did a content analysis of the ITBS forms used by the CPS from 1990 through 1996. We grouped the ITBS math test items into 12 major categories ranging at the easy end from "money and time" and "addition" problems to the more complex tasks of "equations," "multiplication," "division," and "fractions."[6] Figures 4a, 4b, and 4c compare the relative frequency of these 12 different item types in the tests used from 1990 and 1992 with those used from 1993 through 1996 for grades 3, 6, and 8 respectively.

Clearly, a major content shift occurred beginning in 1993. A new topic on "data related concepts" appeared. There was also a major increase in

Figure 4c. ITBS Mathematics Content Changes: What the ITBS Tests in Grade 8



"equation" problems across all grades. This, in turn, was compensated by a decline in the proportion of basic computation items involving "addition," "subtraction," "multiplication," "division," and "number problems."

These patterns reflect gradually changing professional judgments about the appropriate content for elementary school mathematics curriculum. Beginning with the National Council of Teachers of Mathematics (NCTM) standards in 1989, there has been an emphasis on introducing more challenging mathematics into elementary schools. Test publishers such as Riverside, producer of the ITBS, pay close attention to these developments. In general, the content of national norm-referenced tests is deliberately designed to sample broadly from the different kinds of

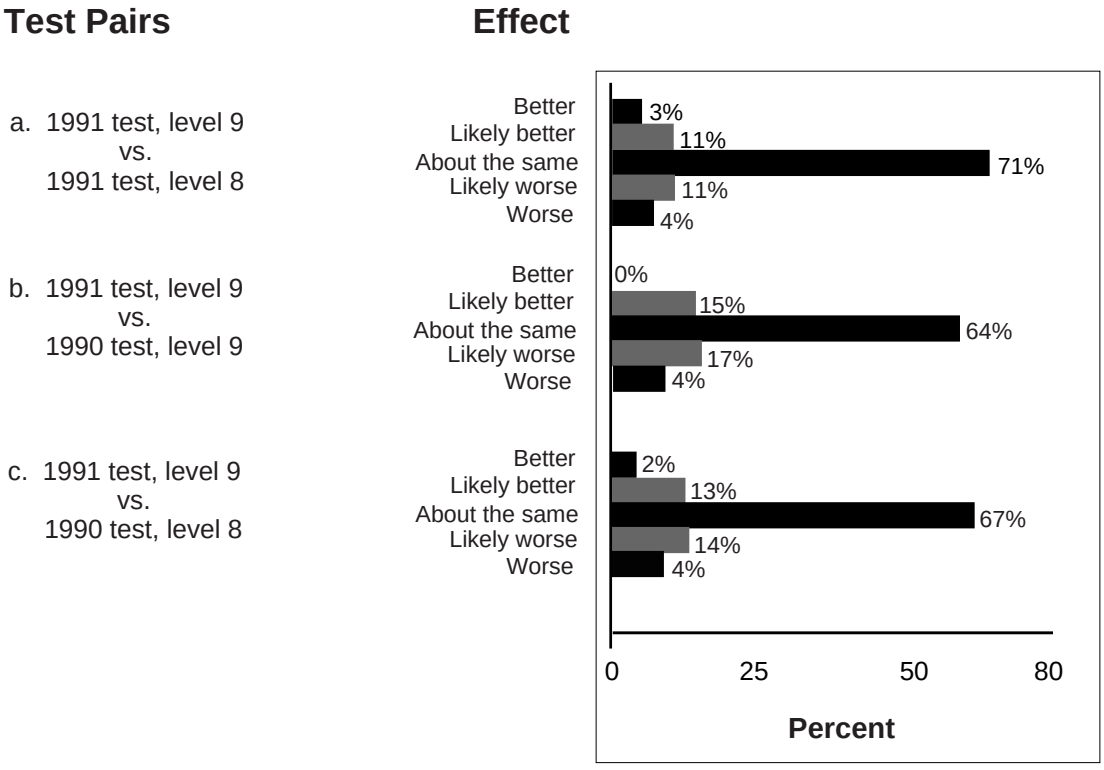
curricula that schools may be implementing in order to provide a basis for global comparisons of how students in a particular school or district compare with a national sample of children who took the same test. The tests are purposefully not aligned with any one curricular strategy so as to be useable across a wide range of schools. As a result, they are a very blunt instrument for assessing increasing productivity in a particular curriculum because only a modest portion of the test may be assessing what schools are actually trying to teach students in any given grade. For example, while the tests used in 1993 through 1996 reflect some movement toward the NCTM standards, few math educators would consider these authentic tests of the more challenging mathematics envisioned in the NCTM.

We again return to our general point. The ITBS system was simply not designed for the purposes to which it is now directed. The testing system was intended to compare the competence level of an individual or group of students relative to a national sample who took the same particular test of basic knowledge and skills. For this comparison to be relevant, the tests try to represent at least some of what children might be asked to learn in a wide range of districts. Moreover, it is quite natural to change the content of norm-referenced tests over time as ideas about instruction shift. This helps to keep comparisons across districts as relevant as possible. This latter principle, however, proves problematic when we switch purposes toward assessing changes in school productivity in a single district. An absolute prerequisite for valid studies of change is a constant measurement ruler.

The Alternative: A Content-Referenced Measurement System

The problems laid out above offer a formidable challenge to any simple assessment of changing productivity in the CPS. We found it necessary to create a new test score metric that allows us to take into account the different content used in various ITBS forms in order better to compare results across time. For this purpose, we undertook a major equating study of all forms and levels of the ITBS used in Chicago from 1987 through 1996 at grades 1 through 8. (See details of the equating study in the sidebar on page 17.) This test equating produced a content-referenced scale that offers a common metric against which persons and schools can be assessed. The scale is constructed around the relative difficulty of the test items for CPS students. Each student is then measured against this content-referenced scale. Any student's scale scores can be directly interpreted in terms of the kinds of items that the student is likely to answer correctly and those that he or

Figure 5. Rasch Test Score Bias Due to Form and Level Differences



Note 1: About the s*ame* category is +/- 1 standard deviation from zero.
Note: See endnote 8.

she is unlikely to know. In this process, we are adjusting for the variations in test content across forms and levels. A particular test is now simply a set of items, each of which has its own unique difficulty. By knowing the difficulty of the items a child got right and wrong, we can calculate a content-referenced scale score.

The major advantage of the content-referenced metric is that the scale scores of students of similar competence or ability should no longer depend on the specific form and level of the ITBS they receive.[7] In Figure 5 we present the same data as previously analyzed in Figure 2. Figures 5a, 5b, and 5c show that, regardless of the specific level or form administered, a student is no more likely to do better or to do worse. The key difference as compared to the GE metric is that, while some students still do better on the first test and some do worse, on average, there is no bias. That is, a student has an equal chance of doing better or worse on the second administration. This is reflected in Figure 5 by the fact that the percentage of students

doing better and the percentage doing worse are approximately the same in all three panels.

On balance, the results presented here illustrate the kinds of improvements that can occur when test scales are content-difficulty referenced. Our equating design involved 24 different situations or *links*, where students took two different forms and/or levels of the ITBS. The GE metric showed bias in half of the cases! The equating removed the bias in eight cases, effected improvement in three situations, and exacerbated it in one case. While this is an improvement, it is less than ideal.[8] To establish better test comparability, the mechanism for test equating needs to be built directly into the design of the testing program rather than treated as a special study as we have done in this research.

The ITBS as Content-Referenced Scales

The reading and mathematics measurement rulers—Figures 6 and 7 (included separately)—present the content-referenced scales for the reading and math series that we developed from the equating study. In both cases, the scale has been established so that test scores run from 0 to 100. These content-referenced scales form a developmental metric. Higher scores indicate more advanced student competency. The scales have been anchored such that a score of 20 is comparable to being at national norms at the end of first grade, and a score of 80 is consistent with being at national norms at eighth grade, based on the average of 1987 to 1996 ITBS scores.[9] The Chicago grade-level averages for 1996 are represented in the blue bars at the top and bottom of each scale. For example, the fourth grade average reading scale score was 48 in 1996; for sixth grade it rose to 60.5. The comparable results for math at grades four and six were 48 and 65 respectively.

The scale score for any student (or the average score for an individual school) is directly related to the specific content that constitutes the test series. For example, students with scale scores of 50 on the reading test have a 75 percent probability of answering correctly the items clustered around that scale value (e.g., items C4, D1, and E2,). They are even more likely to get the simpler items (e.g., C1 and D3) correct. They are less likely, however, to answer correctly the harder items, for example those associated with passage F.

In short, the scale score provides specific information about what students know and can do. This is what we mean by a content-referenced, as contrasted to a solely norm-referenced, testing system.

The reading scale. The reading scale is defined by the difficulty of the reading passages and the individual items associated with each passage. We

present here a sample of tasks from Form 7, which was used by the CPS up to 1989, to illustrate the content difficulty that forms the overall scale. In general, the reading tasks become more difficult as we move from left to right across the scale.[10] Each sample passage has been selected so as to illustrate what a student who is approximately on that grade level should be able to read well. For example, passage E about fireflies represents the kind of text that an on-level student in grade three should be able to comprehend and answer questions about.[11] The difficulties of selected individual items for each passage are referenced against the scale at the bottom of the page. Notice that the items vary considerably in their difficulty even within a single passage. For example, item E1 associated with the fireflies passage is relatively simple to answer and has a scale difficulty of 39; in contrast, item E3 is almost 20 scale points harder.

In general, the easiest passages (i.e., with lower scale score difficulties) involve short simple narratives. The items associated with them tend to ask simple factual questions and make little or no evaluative demands on the student. The questions associated with fireflies offer good examples of this. In contrast, passages on the right draw on more specialized subject matter and offer a more detailed exposition of facts. These passages also tend to use more complex sentences with less common vocabulary. For example, passage H is about the Floating Market in Thailand—a topic with which most Chicago students would not have had any firsthand experience. These upper level passages sometimes tap other literary genres, such as passage I, which is a poem. Items associated with such passages typically elicit the reader's overall impression (or inference) of what a passage is about in its mood, tone, and meaning.

The mathematics scale. The easiest items in mathematics probe students' ability to count, perform simple addition, and tell time. These typically have item difficulties of 20 or less. Next come subtraction and multiplication tasks which become more common around scale values in the 20s and 30s. As we move farther up the scale, the computation tasks become more complex and involve other operations such as division and fractions. Word problems and tasks involving equations become more frequent as well. Some topics, such as geometry problems, span almost the entire scale, but the questions become more complex. For example, a simple geometry problem of identifying shapes has a scale difficulty of 16; in contrast, a geometry problem involving lines and angles has a difficulty of 82.

The interpretation of students' scale scores follows the same basic logic as the reading scale. For example, the average Chicago first grader in spring 1996 had a scale score of 22. Such students are likely to able to

Equating the ITBS

We conducted a series of four separate studies to equate the six different forms of the ITBS used from 1987 through 1996.[12] These studies involved both *vertical equating* (that is, linking different levels within the same test form, such as grades three and four tests given in 1990) as well as *horizontal equating* (that is, linking similar levels in different tests, such as third-grade tests given in 1991 and 1992). In order to accomplish horizontal equating, four studies were undertaken where students completed two different tests. This created the necessary links to make scores comparable across forms.

Within each form, test levels 9 – 14 are linked by common items that appear on more than one test level. This provided the basis for the vertical equating among test levels. For levels 7, 8, and 9, which share no common items within a form, the vertical links were established by groups of students who took two of these different test levels at the same time. These groups ranged in size from 150 to 450 students. Additional data from students who took single test levels (about 1000 people per test level) were included to improve the precision of the item difficulty estimates. For those students who took two tests, the order of test administration was varied. This counterbalancing design was employed to prevent systematic effects of fatigue, boredom, and differential effort and motivation.

The actual statistical equating relied on a method of test item calibration called Rasch analysis. The Rasch model is a member of a class of scaling models based on *item response theory* (IRT) currently used by most modern testing programs such as the NAEP, the SAT, and the TOEFL. Item difficulties for all forms and levels are placed on the same scale. This is intended to assure that all measures are directly comparable.

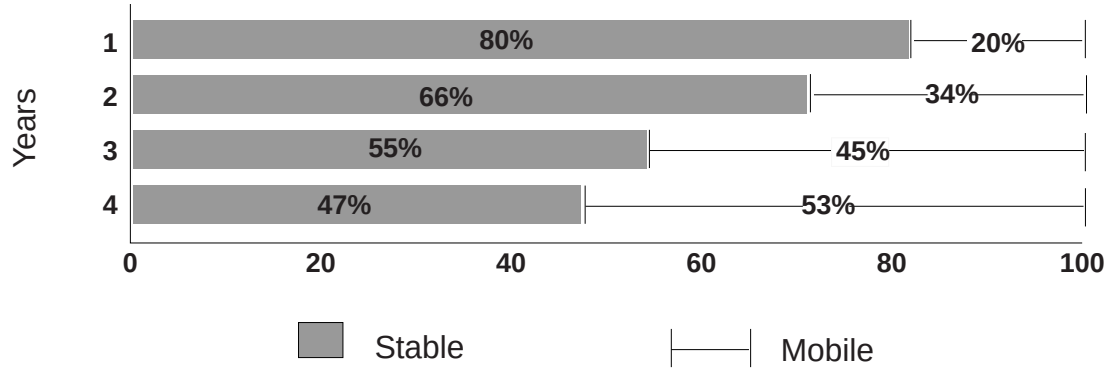
To be sure, issues of comparability can still arise as small changes in the design of a testing program can have a significant impact on observed student performance. The intent in the technical design of the assessments, however, is to assure greater comparability than is now the case.

do simple two-digit addition with no regrouping and even more likely to answer correctly simple addition and time problems. Questions that ask simple multiplication facts (e.g., 3x3=? which has a scale difficulty of 31) would likely be too difficult. Similarly, the typical eighth grader in the CPS in 1996 (scale score of 76) would likely show mastery over most computation tasks (except for the most complex division and fraction problems). But he or she would encounter difficulty with more complex word problems (e.g., the distance, rate, and time problem illustrated with a scale difficulty of 81), or with problems requiring a solution to a linear equation system in two unknowns (scale score of 88), or finding the roots of a quadratic equation (scale difficulty of 91).

A Good Indicator of School Productivity: A Value-Added Approach

We showed earlier that a school mean provides a better statistical summary of the overall attainment of students in a school or district because the

Figure 8. Average Percentage of Students Remaining in the Same School after One through Four Years



performance of every student affects the indicator value. This statistic is most useful for informing us about the overall level of students' capabilities. Moreover, if we track this indicator over time, it will tell us about possible changes in overall student attainment.

The average achievement level, however, is not an especially good indicator of school productivity and whether this is changing over time. One major problem that this indicator fails to take into account is student mobility. For example, if a group of students enrolls in a school sometime during the academic year, even on the day just before testing, their scores will be counted as part of the overall achievement level for the school. Clearly, the attainment for these students depends primarily on their previous schooling experiences and home background and tells us virtually nothing about the effectiveness of the particular school.

This concern is especially problematic in urban school districts such as Chicago because student mobility tends to be high. In the typical Chicago elementary school only 80 percent of the students tested in a given year were also tested in the same school the previous year. This means that 20 percent of the students are new each year.[13] (See Figure 8.) Over a third of the students are new to schools over a two-year period.

Additional problems arise as we examine trends over time. Consider, for example, a school in a "port of immigration" neighborhood. Many of the students enrolled in the neighborhood school will not be native English speakers and, as a result, their measured initial standardized test scores will typically be low. (Further complicating the problem, the CPS currently has no tests designed to measure how well non-native speakers are learning English.) As these students progress through a few years of schooling, their

academic attainment is likely to improve, but they may also leave the school as their family develops opportunities to move into better housing. New immigrants in the community replace these students, and the cycle begins anew. Clearly, the average attainment level for such a school is not likely to get very high because teachers are constantly working with new students. While school staff may do a terrific job contributing to the learning of students who are enrolled, few students stay long enough to significantly affect the bottom line of average student attainment.

More generally, if the student composition of a school is changing over time, the average achievement levels might well rise or fall, but this would tell us little about any possible changes in school effectiveness. Clearly, we need to take such factors into account in developing an appropriate indicator for purposes of assessing school productivity and whether this is changing over time.[14] In order to do this, we begin with a basic accountability principle: A school should be held responsible for the learning that occurs among students actually taught in that school. This suggests that rather than focusing exclusively on the average achievement levels at each grade level, we also consider the *gains in achievement* made by students at each grade in the school for each year.[15]

In addition, as we examine trends in achievement gains over time, we need to take into account other factors that might also be changing during this period that could affect the observed learning trends. For example, over the 10-year period of this study, the CPS changed its procedures concerning eligibility requirements for the testing of bilingual students. Similarly, grade retention policy changed. Both of these policy changes could very well affect the gains recorded at some grade levels and schools. As a general rule, we want to adjust for the effects of such extraneous factors so that any changes over time in a school's value-added to learning will signal real improvements (or declines) in school productivity.

The Grade Productivity Profile

With these ideas as background, we now proceed to define a productivity profile for each school. The school profile is composed of a set of grade profiles, one for each grade in the school for which entry and exit data are available. Figure 9 develops the idea of a grade productivity profile using test data from grade six at Fillmore School.

The productivity profile is built up out of two basic pieces of information for each school grade: the *input status* for the grade and the *learning gain* recorded for that grade. The *input status* captures the background knowledge and skills that students bring to their next grade of instruction. To estimate this input status, we began by identifying the group of students

Figure 9. Constructing the Grade Productivity Profile

9a. Identify 6th grade scores. Match 5th grade scores from previous year. Include only students in same school.

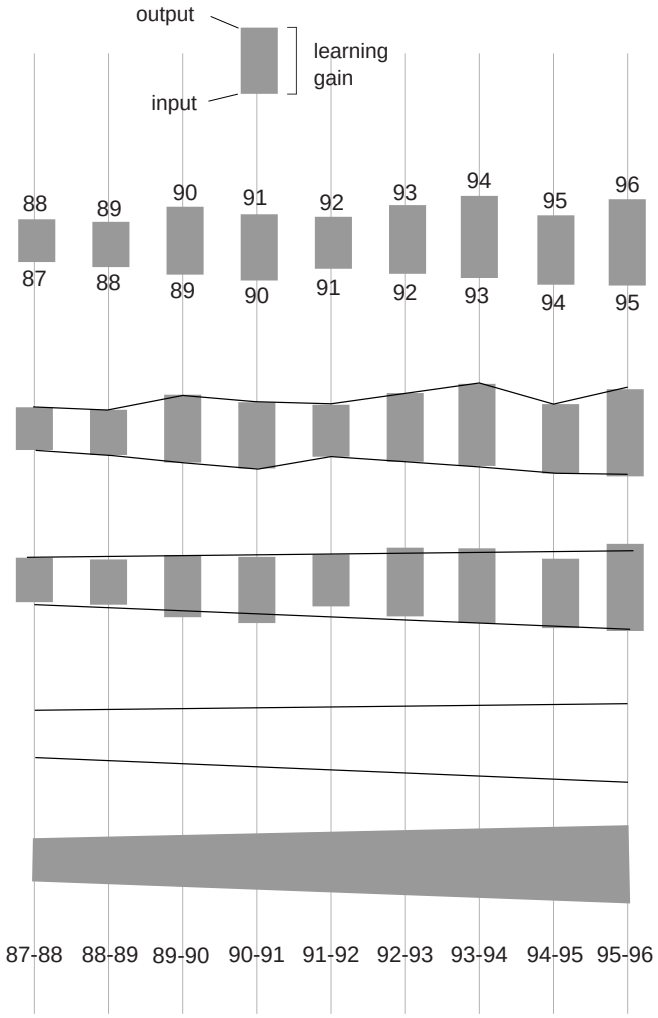
9b. Repeat for all years beginning in 1987.

9c. Add input and output trends to the profile.

9d. Compute smoothed trends using "best fitting" summary line.

9e. To make trends clearer, remove basic data.

9f. Final Productivity Profile



who received a full academic year of instruction in each grade in each school, and then retrieved their ITBS test scores from the previous spring. As noted above, students who move into and out of a school during the academic year do not count in the productivity profile for that year.[16] For our illustrative case of grade six at Fillmore School, we retrieved the end of grade five test scores for students who spent grade six at the school. The average of these students' previous year's test scores is the input status for that school grade. This input status is what teachers had to build on to advance the learning of the stable sixth grade students at Fillmore School that year.

As for the *learning gain* for each school grade, this is simply how much the end of year ITBS results have improved over the input status for this same group of students. In terms of our case example of grade

six instruction at Fillmore School, the learning gains for the stable grade six students is how much their test scores have improved over the grade five scores from the previous year. Finally, by adding the *learning gain* to the *input status* we recover the third piece of information – the *output status*. This tells us about the knowledge and skill levels of these students at the end of a year of instruction. This would be at the end of grade six in our Fillmore School example.

The grade productivity profile is organized around data from some *base year*. In our analyses of productivity for CPS schools we have selected 1991 as the base year.[17] Panel 9a displays the base year input status, learning gain, and output status for grade six at Fillmore School. We then add to this in panel 9b the grade six data for years prior to and post 1991. We have represented now all of the basic data for examining academic productivity in grade six at Fillmore School.

Our interest in changing school productivity directs attention to the variation over time reflected in these data. A visual scan of panel 9b suggests that the inputs to grade six at Fillmore School may be declining over time. Countering this, the learning gains appear to be increasing and with this, the outputs also appear to be increasing. To make this clearer, Panel 9c adds an *input trend*, and *output trend* to the profile. Notice that each of these trend lines varies considerably from year to year. This variability in the data tends to obscure any overall pattern. To highlight this better we compute *smoothed trends* that involve estimating the best summary line that fits these data. These are presented in Panel 9d. To make the trends even clearer, Panel 9e presents the trend lines with the basic data removed.

Indeed, the inputs to grade six have declined, but the learning gains increased. The latter is reflected by the fact that the input and output trend lines spread apart over time. Moreover, since the learning gains increased faster than the input decline, a positive output trend is the net effect. Key to making such judgments is the estimation of smoothed trend lines through the use of a statistical model. (See the Appendix for a description of the model and discussion of estimation issues.) The analysis generates our most concise visual summary of a grade productivity profile. Panel 9f illustrates the final representation of this.

The fitting of a statistical model to smooth the trend lines also serves another important function. It allows us to adjust the trend estimates for other factors that might be changing over time besides school effectiveness. In seeking to develop the best possible estimates of school productivity for the CPS, we considered a range of factors including changes in a school's ethnic composition, percentage of low income students, retention rates,

percentage of students enrolled who are old for their grade, and the proportion of bilingual students. Generally, the effects associated with these factors were not large. In addition, most CPS schools did not vary much on most of these factors over the 10-year period from 1987 to 1996. As a result, the adjusted trends were quite similar to the unadjusted estimates.[18]

Finally, we use our estimate of a school's learning gain trend to quantify school improvement in the form of a *learning gain index* (LGI). This quantity assesses the relative change in student learning over the last five years as compared to the amount of learning that occurred across the system in the base year, 1991.[19]

Classifying Productivity Profiles

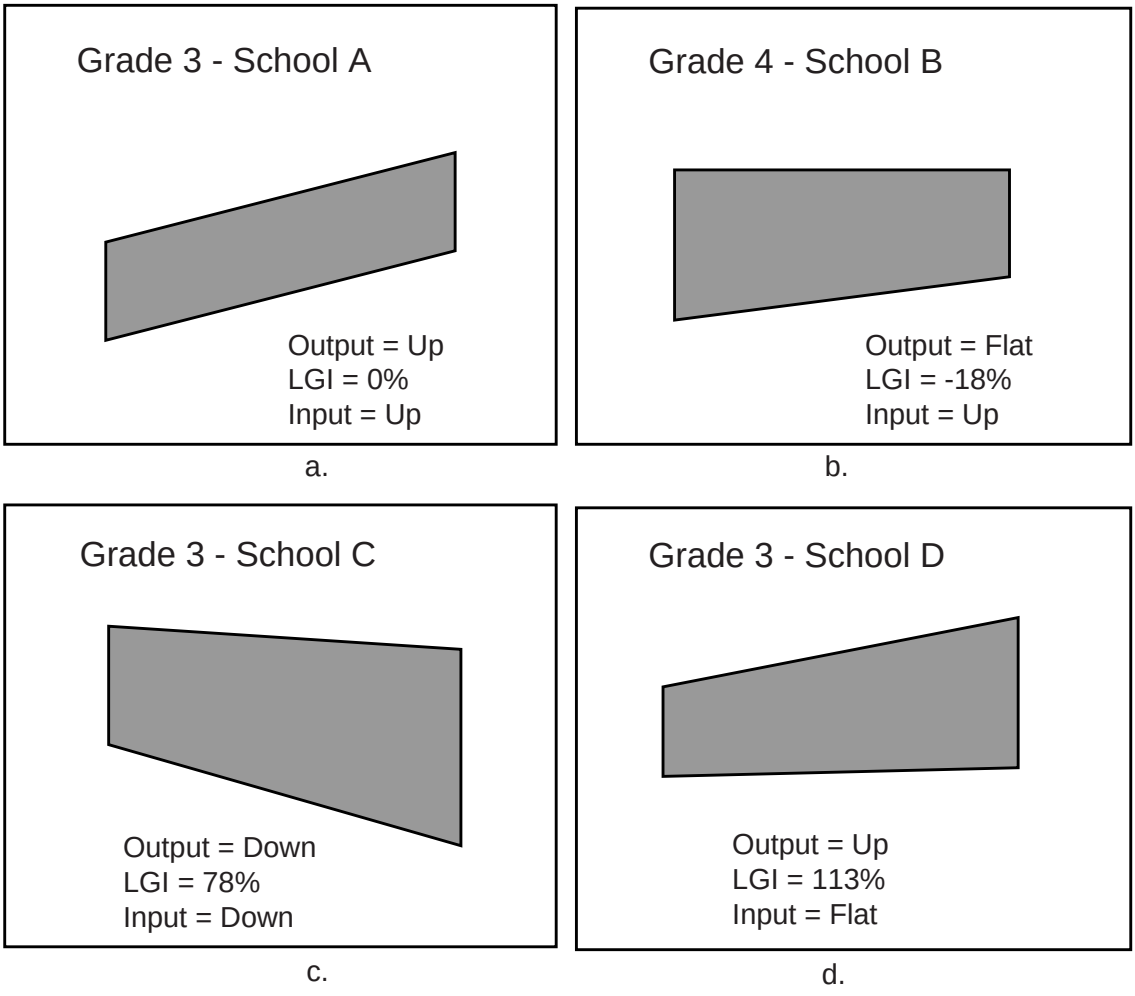
Each grade profile involves three different trends: input, learning gain, and output trends. If we know any two of these, the third can be inferred directly. Observing only one of the three, such as when we monitor an output trend or a gain trend separately, can be misleading.

Much of the recent literature on school accountability emphasizes use of the learning gain trends for purposes of judging productivity.[20] As we began this study, we intended to focus exclusively on the learning gain trend or value-added to student learning for judging school productivity.[21] Gradually, however, we came to conclude that while the statistical arguments for using the learning gain or value-added trend were sound, these arguments were too narrow on both educational and policy grounds. We elaborate our concerns through two examples.

First, consider the grade profile in Figure 10a. Notice that the output trend is up substantially. However, the input trend for the grade is also increasing at the same rate, and the estimated learning gain trend is flat. (Formally, the estimated LGI is 0 percent.) Visually the input and output trends are parallel lines, implying no change over time in the value added to student learning. While most educators would consider the output trend to be indicative of a reform success, focusing only on the learning gain trend would lead us to conclude that no significant change in instruction had occurred in this school grade.

Let's think about what might actually be occurring educationally in "School A." The students entering each year are more advanced than the previous year's students (i.e., the input trend is positive). The teachers must recognize this and each year modify their plans of instruction. Since at least some of the instruction will be new each year, teachers must also engage in continuous formative evaluation—trying to figure out what is working and what is not and adjusting accordingly. In the absence of such teacher activity, we might expect a profile more like Figure 10b. Here, the improving

Figure 10. Grade Productivity Profiles



Note: LGI = Learning Gain Index, computed for 1992-1996.

inputs go unrecognized, teachers continue to teach as they have in the past, and succeeding student cohorts make less progress because, increasingly, instruction is simply a repeat of past lessons. (The learning gain trend is actually negative here. The LGI is -18 percent.) In essence, one could argue that Figure 10b, and not Figure 10a, is the "no change" case in that Figure 10b represents the trends that we might expect to occur if teachers are not proactive change agents.

Now let's consider another case represented in Figure 10c. Both the input and output trends are declining, but the input trend is declining at a faster rate. This pattern results in a positive learning trend (LGI=78 percent) that is reflected in the distance between the two trend lines increasing over time. While from a strict value-added perspective, this is a case of

reform success (improving learning gains over time), it would still be problematic to hold up this case as an exemplar of improved performance. At a minimum, we would want to distinguish it from a school grade with a productivity profile more like Figure 10d. Here, both the output trend and learning gain trends are improving over time. This clearly is a success story!

Examples such as these have led us to conclude that we should employ a dual indicator comparison scheme. Specifically we need to look simultaneously at both the learning gain trends and output trends to classify improvement efforts. Taken together, these two trends provide a detailed summary of changing school productivity over time.

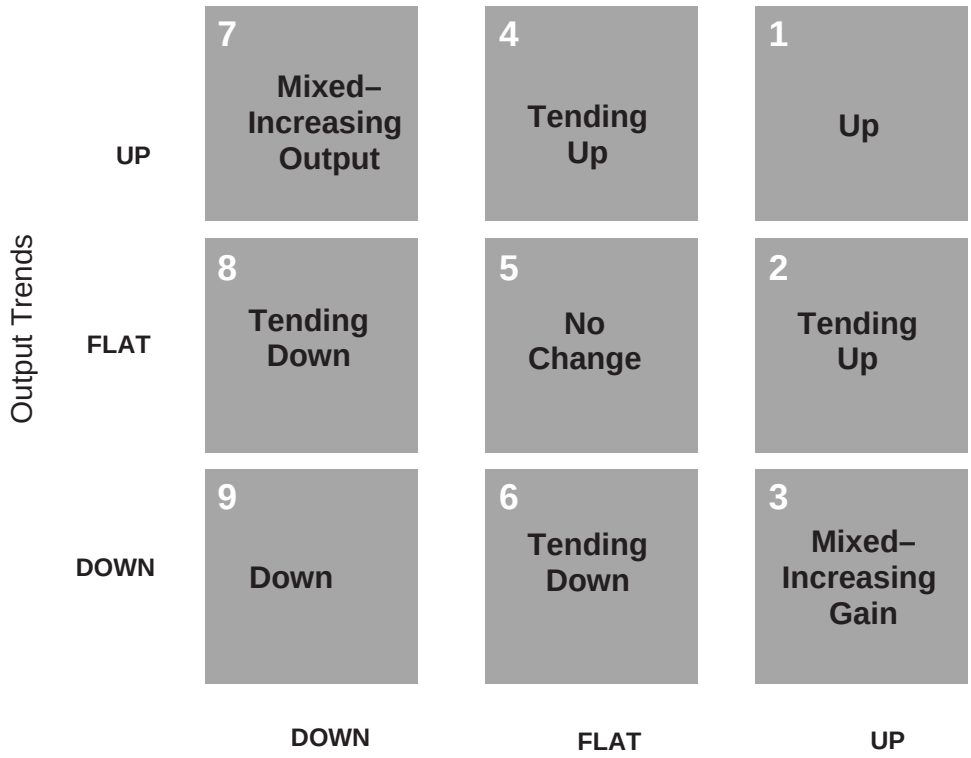
From visually inspecting a large number of grade productivity profiles, we were able to identify nine distinct patterns among output and learning gain trends. These are presented in Figure 11. Each cell in this table is based on whether the output and learning gain trends are up, flat, or down respectively. Some patterns, such as 1, 5, and 9, are straightforward to interpret. These represent "Up," "No Change," and "Down" in academic productivity. We describe patterns 2 and 4 as "Tending Up" since there is some evidence of improvement in either output or learning gain trends. Similarly, we describe patterns 6 and 8 as "Tending Down" because there is some evidence of real decline. Patterns 3 and 7 are the hardest to interpret since the learning gains and output trends are going in opposite directions—one is improving while the other is declining. Without knowing more about the particulars of a school case like this, we call these "Mixed" profiles. The result is a 7-category scheme for describing grade productivity trends.[22]

Summarizing School Productivity

While we compute productivity profiles for each grade, we do not recommend that an accountability system use only single grade information. Our statistical analyses have identified negative relationships among profiles in adjacent grades. That is, improving productivity at one grade tends to be followed by some declines at the next, and the reverse is also true.[23] As a result, judging a school by looking at only selected grades can be misleading. We would be better off, from a statistical perspective, to average across adjacent grades to develop a more stable estimate of school productivity.

Educational concerns also push us in this same direction. The design of a good accountability system should promote cooperative improvement efforts among a faculty in articulating curriculum across grade levels, evaluating improvement efforts, and tracking the progress of students through schooling. This too suggests aggregating adjacent grade level profiles together to focus accountability analyses on the performance of meaningful

Figure 11. A Typology of Productivity Profiles



sub-units within a school. In this way, the accountability system creates incentives for more cooperative teacher work, which has long been a major organizational concern for schools.[24]

In sum, for both educational and statistical reasons, we have grouped grade level profiles to summarize a school's overall productivity. Assuming the basic Chicago elementary school structure of kindergarten to grade eight, we report the following:

- a summary profile for grades three and four;
- a summary profile for grades five and six; and
- a summary profile for grades seven and eight.[25]

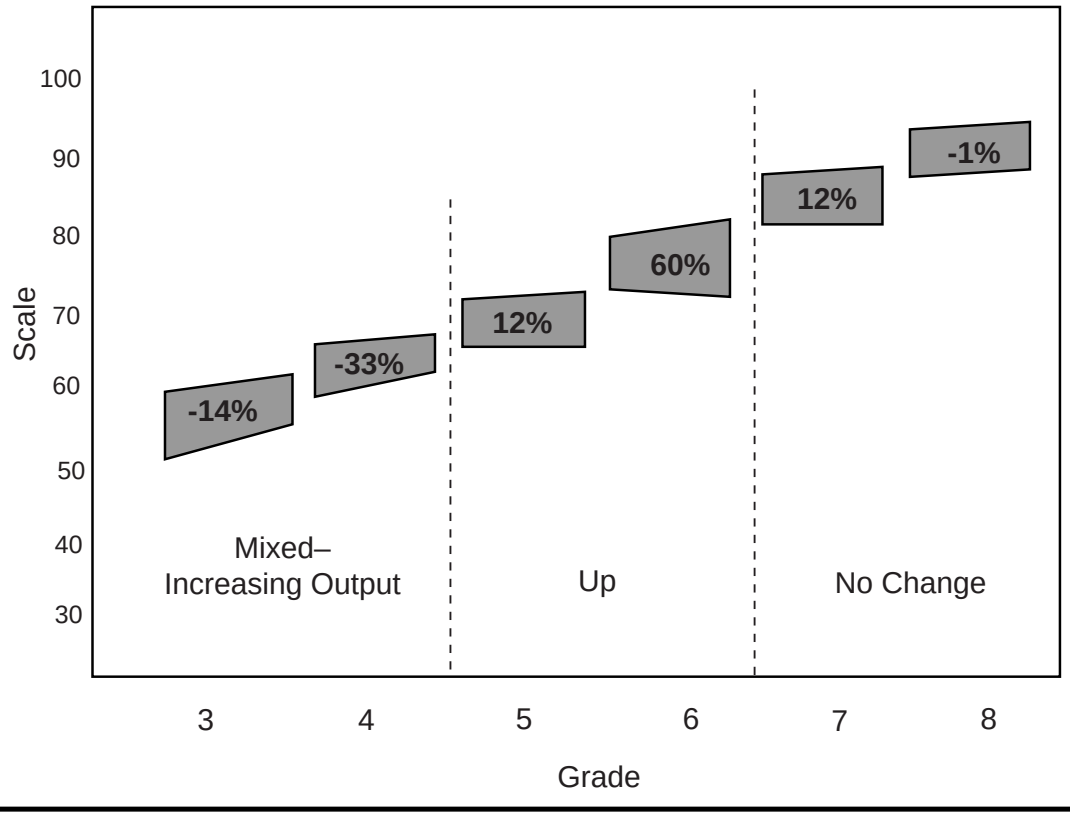
Since we lack comprehensive systemwide data prior to grade two, we have no basis for judging improvements in productivity for pre-kindergarten, kindergarten, and grades one and two. Nonetheless, it is important to chart whether the output of this grade grouping is changing over time since this represents the basic inputs to the next sub-unit. For this reason, we add

the output trend from grade two to our school summary profile. In general, any changes observed in the grade two outputs can be attributed to either change in the kinds of students enrolled or changes in a school's effectiveness. Without more detailed student background data than is available currently through CPS central records, we were unable to sort between these two competing explanations. Individual schools, however, have access to considerable additional information which may allow them to interpret better their second grade output trends.

For our illustrative case of Fillmore School, the productivity summary for reading achievement was "Mixed, Increasing Output" for grades three and four; "Up" for grades five and six; and "No Change" for grades seven and eight. (See Figure 12.) The output trend for grade two (not shown) was also "Up." The vertical axis on this and subsequent figures represents the 0-100 content-referenced scale.

Two examples of school productivity trends. We have argued that the school productivity profile is a better way to examine school effectiveness

Figure 12. Reading Productivity for Fillmore Elementary School



Note: Percentages associated with each grade productivity profile are the percentage improvement in learning gains (LGI) over the base year period (1991).

Figure 13a. Reading Test Score Results at Garfield School

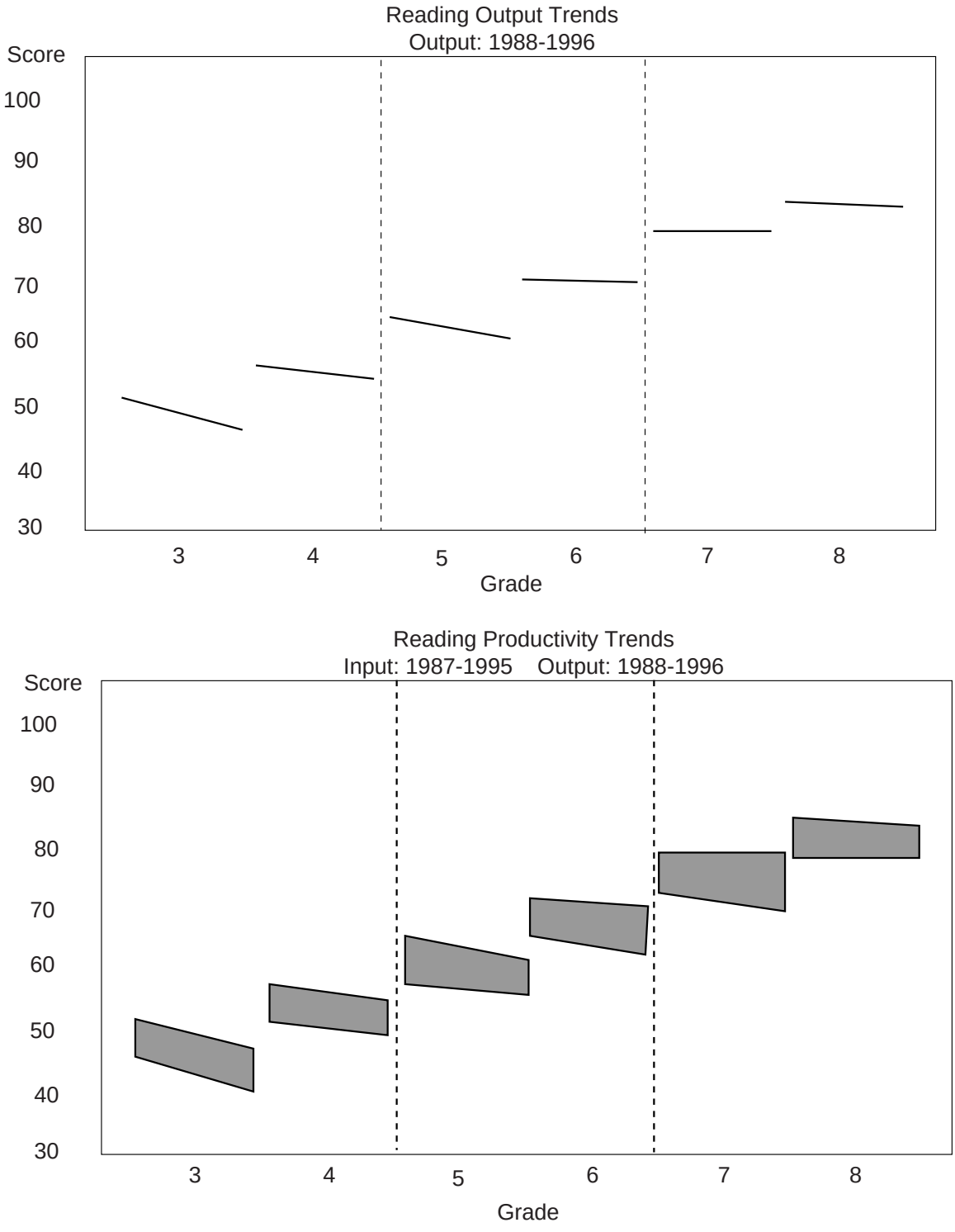
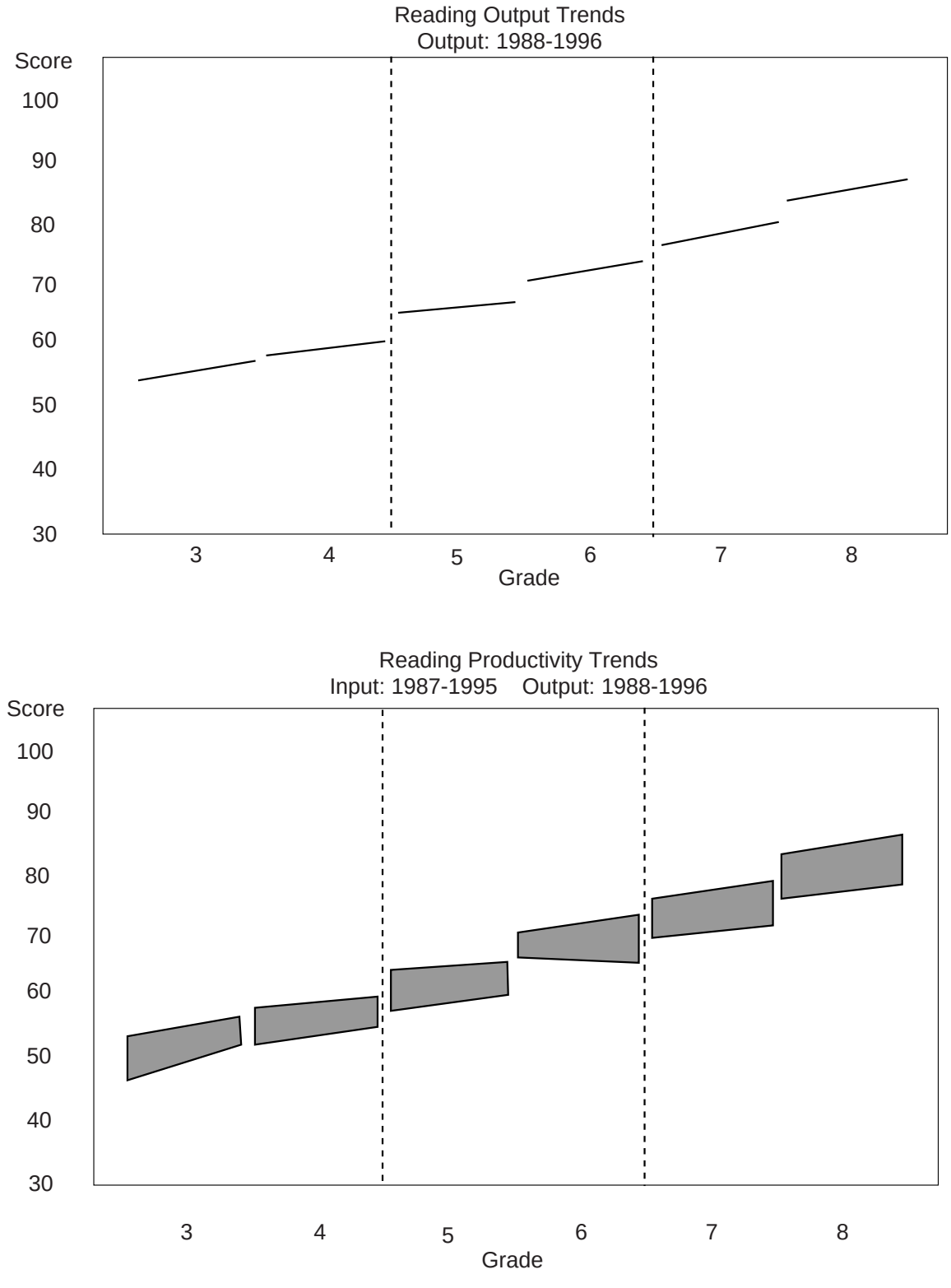


Figure 13b. Reading Test Score Results at Polk School



than simple trends in percentage of students at national norms. We illustrate this with an analysis of reading data from two schools shown in Figure 13. In Figure 13a (top), we see the reading output trends for grades three through eight for the years 1988 to 1996 in Garfield School. The output trends are clearly down in grades three through five and, at best, they are flat in sixth, seventh, and eighth grades. This surely looks like a deteriorating school! A look at this school using the productivity profile, however, tells us more about what is happening here. We see in Figure 13a (bottom) a sharply declining input trend in grade three. The gains in grade three are actually improving over time. A similar pattern appears in grade six, where, again, the gain trends are improving over time. We don't know why these declining input trends are occurring. The school populations might be changing because of community demographic changes; the opening of a new school nearby may be siphoning off some of the stronger students; there may be serious instructional problems in the primary and preprimary programs. Whatever the reasons, it does appear that the middle and upper elementary grade teachers are making a serious effort to respond to the increasing educational needs of students appearing at their classroom door. The productivity profile helps in this case to understand better the output trends and suggests at least some possible clues about what might be happening here.

Polk School provides a different story. The output trends in Figure 13b (top) show increases at every grade level—a big success across the board! But again, the productivity profile (Figure 13b, bottom) provides a more nuanced view. The input status to grade three has gone up markedly in these years. Though the output is up, it has not moved nearly as quickly. (The grade three LGI is actually negative.) This same pattern appears in fourth and fifth grades where the gains are decreasing over time. The input status grows stronger year after year, but the school is not capitalizing fully on this. Again, we don't know fully what is occurring here. This school may have developed an extraordinarily effective primary reading program. If this is the case, this would be something that we should investigate more closely. Alternatively, it may just be recruiting better students. Genuine improvements appear to be occurring at certain grades, such as six and seven, and this we should definitely look at more closely as possible guidance to other schools. Even though the output trends are up for all grades, it is less than clear that all grades are actually improving. In sum, while Polk School still appears successful, we now have a somewhat more circumscribed view about the extent and location of these successes.

Evidence for Systemwide Improvement

Preliminary Results

As we have argued throughout this report, the differences in test forms and levels of the ITBS used by the Chicago Public Schools over the last decade cause a significant problem in drawing inferences about changing school productivity. For this reason, we undertook a cross form and level equating study. While the equating study results are an improvement over the grade equivalent metric for assessing change over time, a great deal of noise remains in these data.

Fortuitously, there is one set of comparisons embedded in the post-reform testing series that is not afflicted by these problems. The test form administered in 1993, Form K, was repeated in 1995, and the form used in 1994, Form L, was repeated in 1996. As a result, the 1994 and 1996 learning gains are directly comparable because they are based on the same pairs of test forms and levels. (For example, the third grade gain in both years is based on second grade students taking level 8 of Form K, followed the next year by level 9 of Form L.) For this reason, we begin our analysis of systemwide trends by focusing on the strongest piece of evidence where results are not contingent on the accuracy of the equating study.

Across the board, for all elementary grades three through eight, the 1996 learning gains were substantially greater than in 1994 for both reading and mathematics. (See Figures 14a and 14b respectively.) The same pattern occurs in GE scores. In relative terms, student gains in 1996 represent improvements ranging from 10 to nearly 40 percent over the 1994 levels. (See Figure 14c.) This is an impressive two-year productivity gain by most any standard.[26]

Why We Focused on ITBS instead of IGAP

The ITBS testing program is better suited for answering questions about school productivity than the state assessment system, the Illinois Goals Assessment Program (IGAP). At the elementary level, the state program tests students in reading and mathematics at grades three, six, and eight (with selected other subjects tested at other grades). Given that individual scores on the IGAP were not available until spring of 1993, it was impossible to link student scores to estimate individual learning gains and thereby have a basis to create value-added indicators for schools. Even with individual scores that were made available since 1993, the spacing of the testing remains too far apart because so many students will have changed schools between these grades. The annual ITBS testing program reduces these problems considerably. It provides a basis for measuring learning gains each year for each school and grade.

Figure 14a. 1994 vs. 1996 ITBS Reading Gains

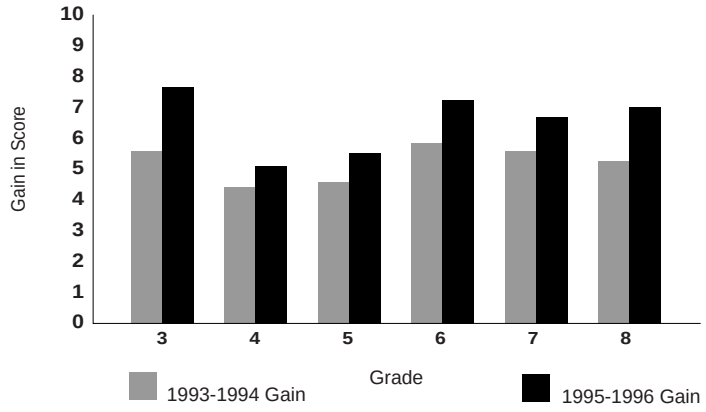


Figure 14b. 1994 vs. 1996 ITBS Mathematics Gains

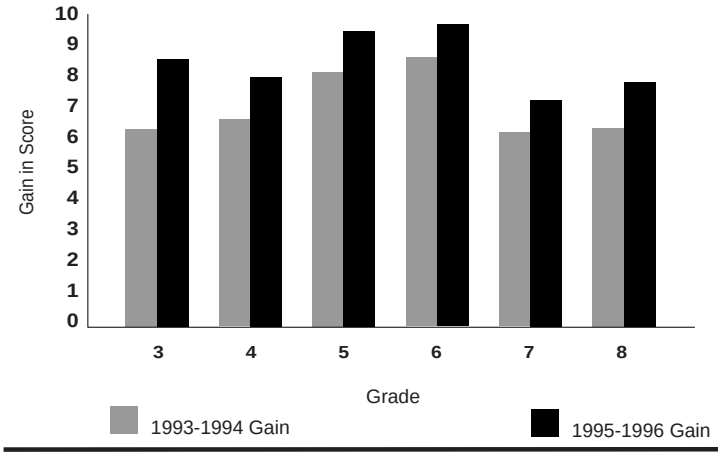
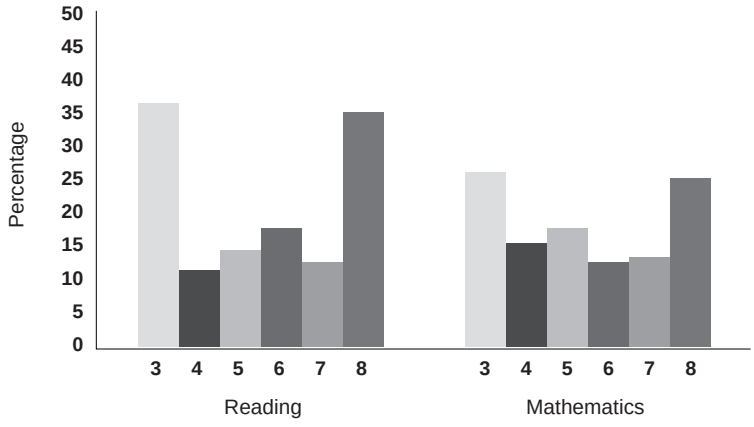
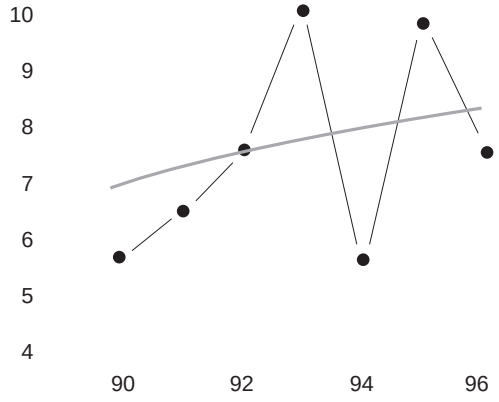


Figure 14c. 1996 Improvements over 1994 Gains

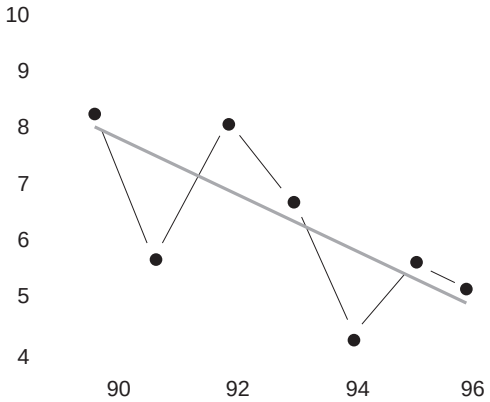


Note: (1996-1994 gain) / average (1988, 1989, 1990 gain)

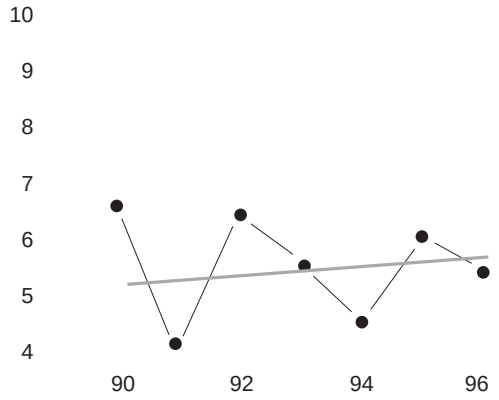
Figure 15. Long-Term Systemwide Gain Trends–Reading



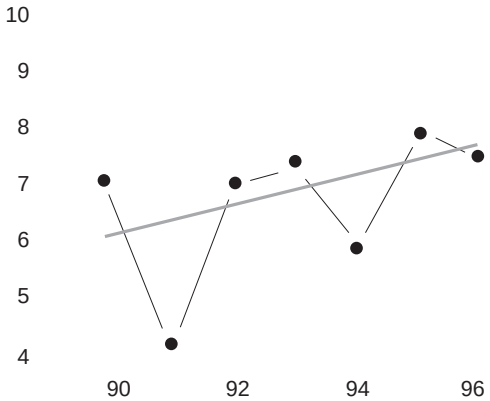
15a . Grade 3



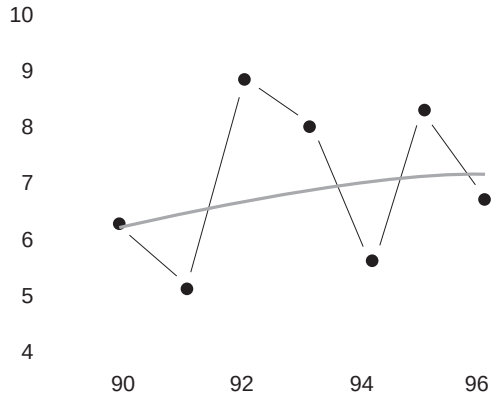
15b. Grade 4



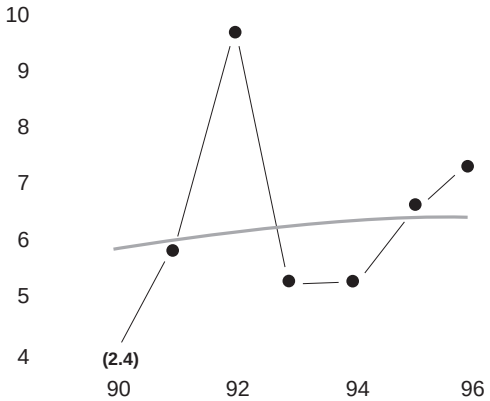
15c. Grade 5



15d. Grade 6

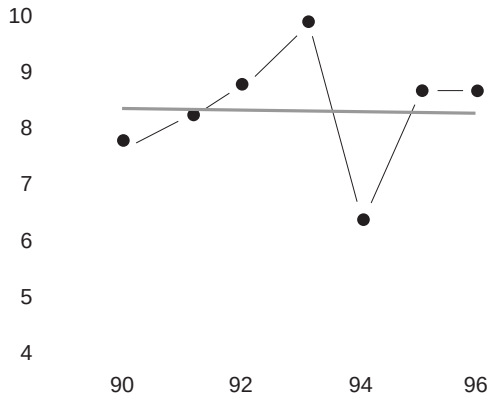


15e. Grade 7

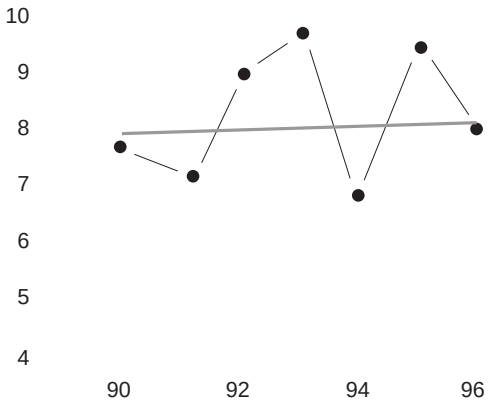


15f. Grade 8

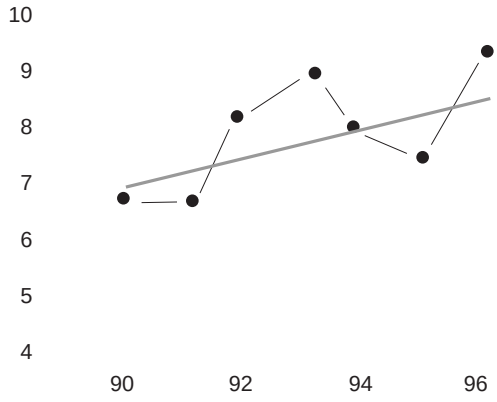
Figure 16. Long-Term Systemwide Gain Trends–Mathematics



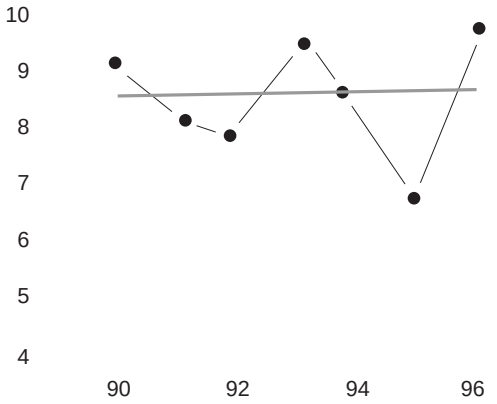
16a. Grade 3



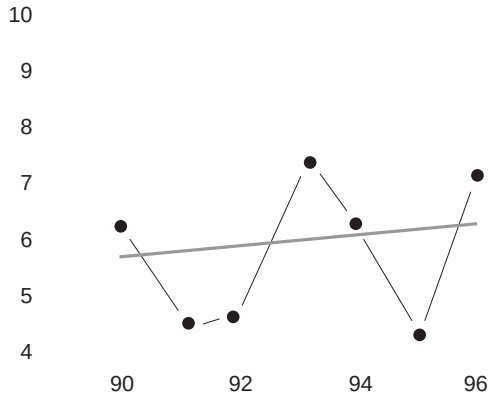
16b. Grade 4



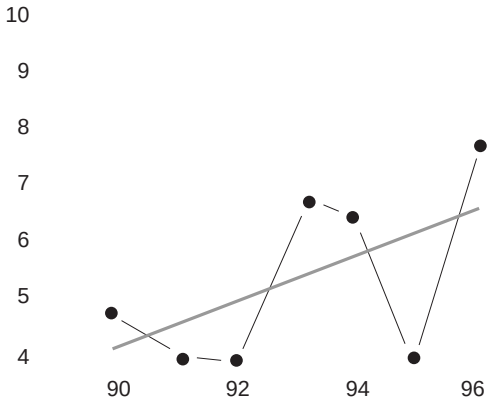
16c. Grade 5



16d. Grade 6



16e. Grade 7



16f. Grade 8

These results immediately raise a second question, "Is this a one-year phenomenon (e.g., something attributable to the administrative reforms of 1995) or rather a signal of a longer-term improvement trend that links back to the reforms of 1988?" To investigate this question, we do have to rely on the equating study results in order to examine the 1996 gains within the larger context of the learning gain trends over the last several years. If the improvement registered in 1996 is a one-year phenomenon, we would expect to find flat or possibly even declining trends in learning gains through 1995 followed by one big upward jump in 1996. In contrast, the 1996 data may look like a natural part of a longer-term trend. The results for reading and mathematics are presented in Figures 15 and 16 respectively.

The first and most immediate observation from scanning the learning gain trends is that these data are quite noisy. While we can see that the 1996 gains are higher than in 1994, a similar pattern occurred for the 1993 gains as compared with 1991. In almost every case, the 1993 results outpaced 1991. Taken overall, the 1996 results appear for the most part to be embedded in longer-term (albeit noisy) trends. To see this more clearly, we estimated trend lines from 1990 through 1996 that discount year-to-year data fluctuations.[27] Generally, the 1996 data look consistent with these overall trends. There is some evidence, however, of an upward jump in 1996 in selected grades, especially in mathematics.

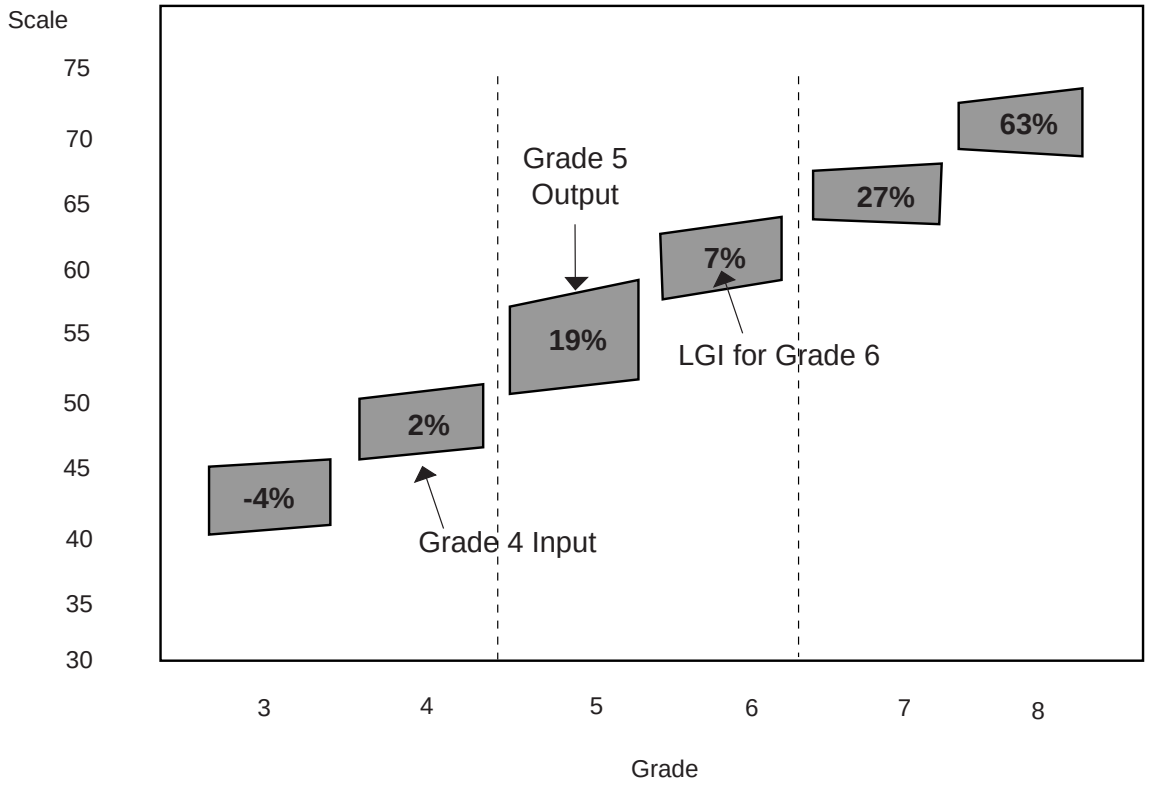
These analyses provide our first evidence about trends systemwide in academic productivity. To probe this further, we now turn to an analysis of the school productivity profiles. This offers more complete information about academic productivity in that it considers simultaneously both output and learning gain trends.

Systemwide Average Productivity Profile

Figure 17 displays the overall ITBS mathematics productivity profile for the CPS. The display aggregates the grade productivity profiles from all individual elementary schools in the system. Notice that the output trends are up for all grades three through eight.[28] The learning gain trends also show marked improvements for the middle and upper grades. For grade five, the systemwide improvement, as summarized in the LGI, was 19 percent over the five-year period from 1992 through 1996. For grades six, seven, and eight the relative improvement in mathematics learning was 7 percent, 27 percent and 63 percent respectively. In contrast, grades three and four show little change in learning gains over this period.

The grade three mathematics data are quite interesting. The estimated LGI for grade three is actually slightly down (by 4 percent), but the output

Figure 17. Mathematics Productivity Profile for CPS, 1987-1996



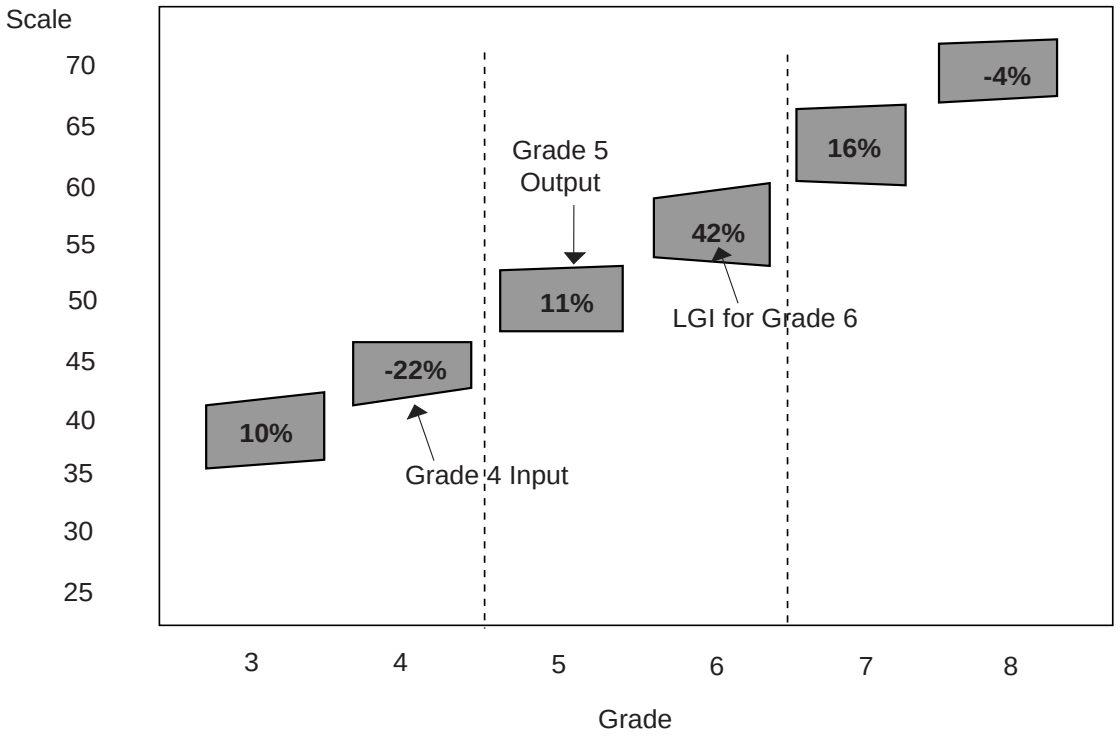
LGI = Learning Gain Index, computed for 1992-1996.

trend is still positive. This is a case where if we just looked at the output trend, as we might under a more traditional accountability approach, we could mistakenly conclude that third grades have been improving systemwide. In fact, the registered gains in achievement at the end of grade three appear largely attributable to major improvements prior to grade three.

Figure 18 displays the systemwide productivity profiles for reading. The results here are a bit more mixed, but still generally positive. The output trends are up at all grades except grade four; the rates of improvement, however, are not as large as in mathematics. Grade four registers a learning rate decline of 22 percent. Grades three, five, six and seven, however, show significant gain trends ranging from 10 to 42 percent; Grade 8 remains basically unchanged.

Overall, our analyses indicate broad-based systemwide improvements in student learning, stronger in mathematics but also in reading. Moreover, we believe that these data, up through 1996, are a reasonably good

Figure 18. Reading Productivity Profile for CPS, 1987-1996



Note: LGI = Learning Gain Index, computed for 1992-1996.

indicator of meaningful changes in instruction and student learning because no high stakes external accountability were associated with them.[29] That is, prior to 1996 the main external accountability force over Chicago Public Schools was the Illinois State Board of Education, which based its school rankings and "academic watch list" on IGAP data. Although ITBS scores still mattered to individual schools, no formal consequences were directly attached to them. Beginning in 1996, the CPS instituted its own high stakes accountability system based exclusively on the ITBS; as a result, the future utility of these data as an indicator of broad instructional improvement has become more questionable.

Distributions of Individual School Productivity Profiles

As we noted in the introduction, we should expect varied outcomes among schools under decentralization reforms such as the 1988 legislation. Some schools were well positioned at the onset of reform to take good advantage of the opportunities and resources it afforded to press for deep changes. Others, starting with a much weaker base of human and social resources,

were likely to progress less rapidly and, in the worst cases, might even move backwards. Clearly, patterns among schools in their academic productivity are likely to be related to these core resources for improvements in student learning.[30]

Thus, in addition to looking at systemwide aggregate trends, we also focus attention on the distribution of productivity trends across the 466 elementary schools that comprise the composite pictures. For this purpose, we rely on the school productivity classification system introduced earlier. We summarize each school's performance in terms of the aggregate productivity profile for the lower, intermediate, and upper elementary units (i.e., grades three and four, grades five and six, and grades seven and eight) and the second grade output trend. The summary productivity profile for each school is classified using a seven-category scheme that ranges from clearly "Up" through "Mixed" and "No Change" categories to clearly "Down." (See page 48 and endnote 22 for a further description.) The second grade output trends are subject to a three-category classification: "Up," "No Change," and "Down." The distribution of school results is presented in Figure 19 for mathematics and Figure 20 for reading.

Figure 19. Distribution of Mathematics Productivity

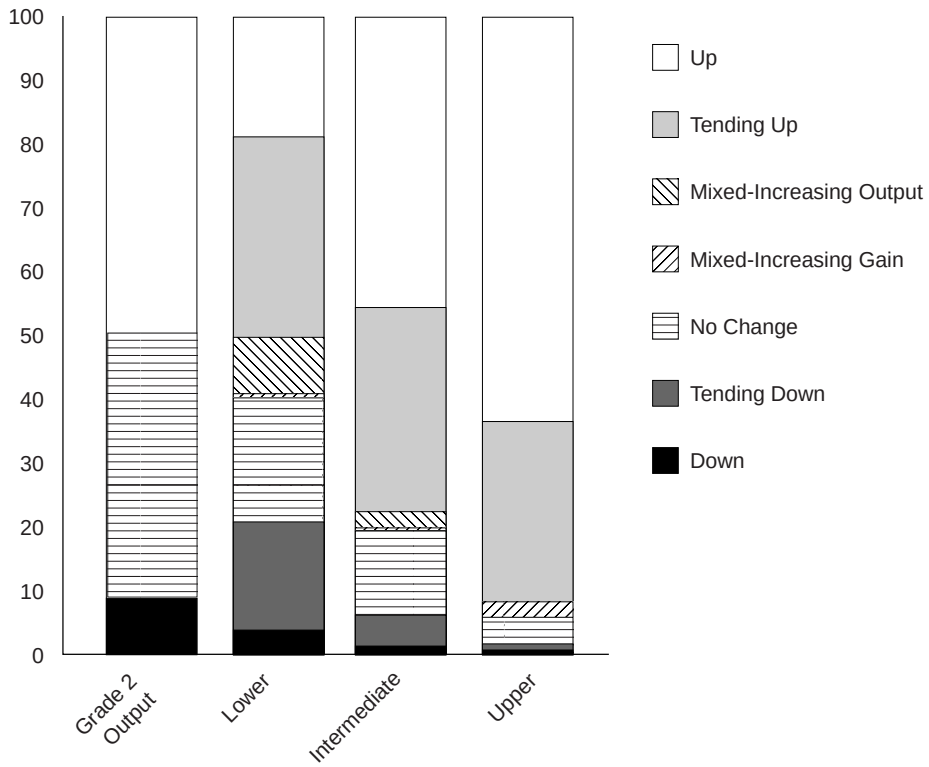
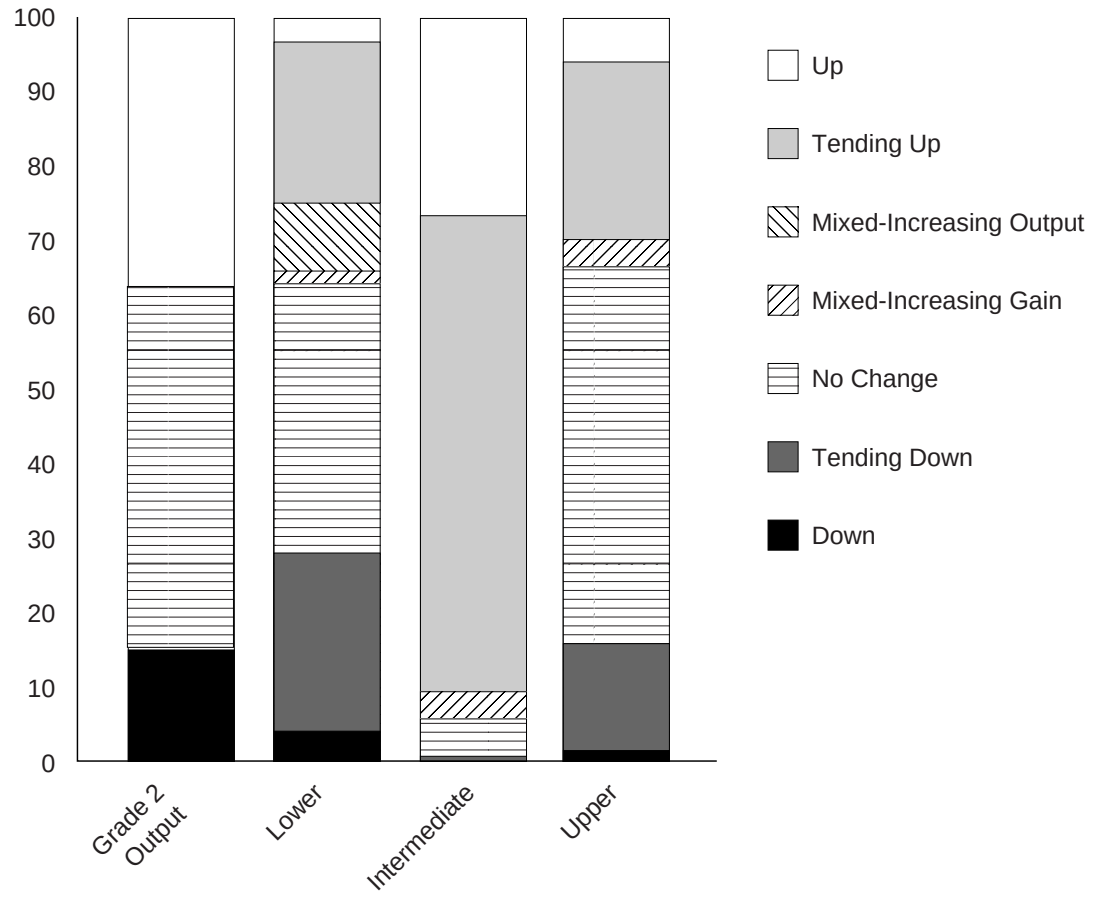


Figure 20. Distribution of Reading Productivity



For mathematics, 50 percent of the schools have positive productivity trends for the lower grades. A substantial group (19 percent) are "Up" in both output and gain trends, and another 31 percent are "Tending Up"—that is, they are up either in output or gain trends. About 28 percent of the schools have "Mixed" or "No Change" productivity ratings. In these schools, there is either no evidence of improved scores or contradictory evidence—that is, either outputs are up and gains are down or the reverse. In 21 percent of the cases, productivity is "Down" or "Tending Down."

The picture is much more positive for the intermediate grade mathematics results where 45 percent of the schools are "Up" and another 32 percent are "Tending Up." About three percent of the cases are "Mixed" and 13 percent show "No Change." A very small number of schools—about six percent—are "Down" or "Tending Down."

The upper grade trends are even better than the intermediate grades in math, where 63 percent of the schools are "Up" and another 28 percent are "Tending Up." The remaining 9 percent of the schools are spread among the other categories.

In reading, the pattern is less clear cut. In the lower grades, there are about equal numbers of schools with positive trends (24 percent) as with negative trends (28 percent). The most prevalent category in these grades is "No Change," (36 percent) and 11 percent of the schools have "Mixed" trends.

The intermediate grades are quite positive, with 27 percent of the schools in the "Up" category and 64 percent "Tending Up." The upper grades are more varied, with the biggest category (51 percent) being "No Change." Thirty percent are "Up" or "Tending Up" and 16 percent are "Down" or "Tending Down."

We also note that grade two output trends suggest some possible improvements in CPS primary programs. For mathematics, 50 percent of the schools show improved trends compared to 9 percent declining. In reading, 35 percent show positive trends compared to 15 percent declining.

Interpretation

Interpreting these results is complicated by the fact that the ITBS has been a tacit rather than an explicit standard for school performance. While the content of the tests, as laid out in the content-referenced scale presented earlier, is certainly reasonable—few would disagree that well-educated students should be able to answers these kinds of questions—this has never been publicly established as a systemwide standard for subject matter content and sequence. Individual schools may well be working on other academic goals. They might, for example, focus more on higher-order thinking skills or deeper engagement of students with projects and some specific subject matter. Some schools may also be teaching to the ITBS content maps, but not in precisely the same sequence. Such schools may record weak test scores in some grades, where the alignment between the test and instruction is poor, and then much better results in other grades where the test better matches with students' actual classroom experiences.

With these caveats in mind, we proceed to offer an interpretation. There appears to have been significant improvement systemwide in primary programs through grade three. This is reflected in the improving second grade output trends in half of the schools in mathematics and a third in reading. We know generally that the student population entering the Chicago Public Schools has been growing gradually more disadvantaged over the recent

10-year period.[31] Thus, although we cannot specifically account for the sources of the improvement (e.g., more state pre-kindergarten; improvement in kindergarten, grades one and two curriculum), something positive appears to have happened here.

For the most part, the early gains prior to grade three, however, are not being further advanced during grades three through four in either mathematics or reading. In fact, some of the early improvements appear to be lost. In contrast, the intermediate and upper elementary grades look quite positive. A word of caution is in order, however. If the primary grades suddenly began to promote children with much higher levels of skills, this would challenge intermediate and upper grade teachers to rethink their instruction to build on the advanced knowledge that students now possess. In the absence of such proactive improvement efforts by intermediate grade teachers, i.e., if they simply repeated past practice, their productivity trends would suddenly start to look worse. In short, the positive trends in the intermediate and upper grades may, in part, be a curious consequence of the lack of productivity improvement in primary grades.

Finally, while the overall thrust of this report has been to design a defensible indicator reporting system for assessing school productivity, we would be remiss not to comment at least briefly on the substantive import of these findings. Notwithstanding the numerous methodological issues documented above, these analyses strongly suggest that Chicago school reform has precipitated substantial improvements in achievement in a very large number of Chicago public elementary schools. The governance reforms of 1988 and 1995 have significantly advanced the learning opportunities afforded to literally hundreds of thousands of Chicago's children. While more improvements are still needed, these results should nonetheless encourage the public that Chicago's schools can substantially improve and that this is, in fact, occurring.

Summary

This study set out to examine the issues involved in using extant standardized test data for school accountability and to devise a more defensible indicator system for monitoring and reporting changes over time. Our efforts met several serious obstacles to a straightforward tallying of the results. We found it necessary to contend with a host of confounding factors including:

- Changing test forms from year to year that rendered both test content and scales not directly comparable;
- Inappropriateness of standard measurement metrics, such as grade

equivalents, for measuring change;
- Significant school-to-school variability in student mobility; and
- Possible effects of changing system policy, such as those regarding retention and bilingual testing.

These concerns pose serious problems for the academic productivity indicator currently used by the CPS—percent of students at or above national norms.

In response to the first two issues, we returned to the original student ITBS item scores and undertook an equating study to derive a *content-referenced scale* for comparing student performance across grades and years. All of the results presented in this report are based on this content-referenced scale. In response to the last two issues, we developed an academic productivity profile based on student learning gains that adjusts for extraneous factors, which also influence reported achievement levels, in order to determine true school effects. Our approach estimates each school's contribution, or *value-added,* to the learning of students enrolled at that school.

We have striven to provide an accurate summary of the progress of Chicago schools over the last decade. The results presented here are the final product of literally thousands of alternative analyses conducted over several years. They represent the best assessment that we can offer, given the limitations of the extant testing system. These inherent limitations are substantial, and we feel less than completely satisfied with the end product. Weaknesses in the data have frustrated our efforts to develop more accurate answers. Again, this is not a criticism of the ITBS per se; rather, we along with the CPS are simply trying to use these tests for purposes for which they were never designed.

Recommendations of the Steering Committee of the Consortium on Chicago School Research[1]

This study has highlighted a number of problems in the standardized testing program conducted by the CPS throughout the 1990s. It demonstrates how these concerns make it difficult to assess accurately the learning gains of students in basic skills, to detail the extent of systemwide improvement, and to make judgments about the productivity of specific schools. While it has deployed sophisticated statistical methods in this work, nonetheless these are at best a weak substitute for a better designed information system. The Consortium Steering Committee concludes that the CPS needs a new standardized testing and reporting system in reading and mathematics in order to have a more reliable basis for judging school and system improvement in these key areas.[2] We offer below some specific recommendations to guide this development.

---

[1]As is customary in reports sponsored by the Consortium, the Steering Committee has reviewed and commented on earlier drafts of this report. Nonetheless, these research findings and the specific methodologies on which they are based are the sole responsibility of the authors. No specific endorsement of this research by individual Steering Committee members should be assumed. The recommendations offered in this section, however, have been specifically endorsed by the Steering Committee. This is only the third time in the eight-year history of the Consortium that its Steering Committee has formally endorsed some aspect of research that the Consortium has sponsored. This decision indicates a consensus among Steering Committee members about the importance of developing an improved student assessment and reporting system for the Chicago Public Schools.

[2]The recommendations presented here derive from a study of problems associated with using the ITBS scores in reading and mathematics for purposes of making judgments about school productivity. These recommendations are limited to how to improve this component of the CPS assessment system. We specifically note that many aspects of the recently approved CPS Learning Standards take us well beyond what can be easily measured with standardized tests of this sort. Similarly, we do not consider at this point the issues involved in improving instructionally embedded assessment, i.e., the data available to teachers to chart student progress within an academic year and to evaluate the effectiveness of instructional lessons and units. The Consortium maintains that new developments in both of these areas are also critical. Specific recommendations in these areas, however, would take us well beyond the research reported in this paper. In addition, we have not considered the relationship between CPS and state assessments in this paper.

Alignment with CPS Learning Goals

New CPS standardized tests in reading and mathematics should be directly aligned with the recently approved CPS Learning Standards. Only if the assessments are specifically developed to achieve this aim and have been demonstrated to be valid in this regard will teachers, students, and parents know whether they are making genuine progress on these important goals. The content of the standards should dictate the content of the tests. A "back-in" solution (i.e., choosing among existing tests the one that comes closest to matching the standards) is inadequate. Under such an approach the test publishers rather than local leaders get to decide the accountability standards for judging schools.

Score Reporting on a Content-Referenced Scale

In addition to a reference to national norms, test score results should be reported in terms of the specific knowledge and skills that students have demonstrated. The content-referenced scales used in this paper illustrate this kind of approach. The results for a sample of actual test questions that were administered to students each year should also be made public. This can help to promote in each school community, and across the city, an educational, as well as a numeric conversation about what students actually know and can do.

A Stable Measurement Ruler for Assessing Academic Progress

For purposes of making judgments about school and system improvement over time, it is essential to maintain a stable testing system that consistently measures students against the same content standards from year to year. This will require that rigorous equating procedures be fully integrated into the overall testing system design. As views about appropriate learning goals shift over time, this will entail developing new assessments and starting new trend lines about progress toward these new standards. The changes in goals should not be obscured by changing forms of the assessment, as has been the case in the past.

An Accountability Focus on Schools' Value Added to Student Learning

Each Chicago school should be held accountable for the amount of learning acquired by students enrolled in that school. While the current CPS school accountability indicator—the percentage of students at national norms—provides information about the overall attainment of students, it

does not tell us how much they have learned in any particular school. Whether gains in student learning are improving over time should play a major role in the school accountability system. Thus, for purposes of reporting on individual school accountability, the CPS should add a focus on gain trends in addition to its current data on output trends.

An Inclusive Orientation

The design of new assessments should incorporate a strong commitment to measuring the learning progress of all students. Similarly, schools should be held accountable for the progress of all students, and procedures should be established to minimize the exclusion of students from the accountability system.

The accountability system must take better account of the learning gains of mobile students (i.e. students who have changed schools during the 12-month period prior to spring testing.) The attribution of these students' learning gains to schools, however, must be adjudicated in a fair manner.[3] Procedures can be devised that are both fair to individual schools (i.e., including in their accountability report only the students who have been enrolled long enough for the school to have a measurable impact on their learning) and inclusive in orientation (i.e., where learning gains for over 90 percent of the CPS students are counted.)

Reporting test scores and learning gains for retained students is also an important issue. Incentives implicit in the current system are to retain students because their scores will then count against a lower grade level. Since the progress of retained students is especially important, the reporting system should give them explicit treatment. One possible method would be to analyze the progress of retained students separately (that is, disaggregated from the total); another possibility would be to report scores by age level of students rather than by grade level.

---

[3]Although we were unable to do this in the analyses presented in this paper, because of limitations in the test data information files available to us, the CPS can accomplish this by merging the test files with information from the student administrative history files on the time and places of student enrollments. One possible rule would be to count for each school the learning gains for all students who were enrolled in the school continuously from some date, for example October 1, through spring testing. Another possible rule would be to assign the learning gain to the school where the student spent the most time in the past academic year, regardless of entry and exit dates.

Because of the importance of the first years of schooling for developing a solid foundation of literacy and numeracy skills, Chicago schools need better assessment tools for evaluating the progress of primary grade students toward the CPS Learning Standards. Extant group administered standardized tests, however, are not adequate. A more comprehensive approach, using reliable and valid techniques, needs to be adopted for young students. We specifically caution that assessments in kindergarten and first grade have often been misused in the past to label and track students into weak instruction with low expectations for learning. While it is important that Chicago schools have better data for judging their efforts to educate students during the critical early years of schooling, such data should never be used to limit students' opportunity for learning.

The number of students in the CPS whose home language is not English is now one in six and rapidly growing. We lack good data on the academic progress of these students, and many are currently excluded from annual school accountability reports. The design of a new basic skills assessment should make provisions to meet the linguistic needs of students who are in the process of learning English. Similarly, special education students should be included. In this regard, new tests should be designed consistent with the guidelines for testing and reporting set out in the recent settlement agreement on the education of special needs students in the CPS.

Concluding Comment

These recommendations are offered as a set of guidelines to the CPS to assist its efforts to strengthen data collection and reporting on school academic accountability. We view such developments as necessary, in fact essential, to the continued improvement of Chicago schools. Maintaining core academic standards and vitalizing a capacity to report more accurately to each school community on its progress toward the standards is key to making school reform work.

Endnotes

[1]For simplicity of the illustration, we have presented results in terms of percentile scores. Normally we do not compute means on such rank order statistics. The use of means is highly appropriate with the equated measures developed later in this paper.

[2]Our point here focuses exclusively on choice of any appropriate statistical indicator. Regardless of the indicator chosen, the accuracy of our inferences will depend on whether all eligible students are actually tested within each school. Unless the accountability system is crafted properly and carefully implemented, it can create incentives for schools not to test some students. This problem has to do with the design of the testing system as well as with the choice of an appropriate analysis indicator.

[3]These testing experiments were part of our common person equating study. The entire test equating design involved 24 different situations where students took two different forms and/or levels of the ITBS. The three cases presented here were chosen to illustrate the kinds of problems that can occur.

[4]Formally, we are concerned about evidence of bias in the testing system—are students as a group likely to perform better or worse on one test as compared to the other? In each of the cases reported here, the mean difference in student results between test administrations was statistically significant beyond the 0.01 level.

[5]Fillmore is a fictitious school name; the data are, however, from a real Chicago school. We have adopted the convention of pseudonyms for real school names throughout this report. Since the purpose of the report is to examine issues in the CPS testing and indicator system, directing attention to specific schools can distract the reader from our primary aim. Separate individual school reports are being released at the same time.

[6]This is based on the results of the item calibration undertaken as part of the equating study described in a subsequent section.

[7]A second major advantage of the equated test score metric is that it produces a goodness of fit statistic for each student's test score responses. Since the items have now been arranged in difficulty order, we expect most students to get the easier ones correct and to miss the harder ones. The misfit statistic tells us whether students are responding to the scale in this manner. If, for example, some students are just guessing randomly, they will have large misfit statistics because they will have missed easy items and gotten some hard ones correct just by chance. The misfit statistic tells us that we just don't know much about the true competencies of these students. We use this information in the school productivity profile to compute more accurate estimates of school trends.

A third major advantage of the Rasch equating is that, in theory, it produces a "linear test score metric." This is an important prerequisite in studies of quantitative change. This allows us to compare directly the gains of individual students or schools who start at different places on the test score metric.

[8]In these comparisons, a student is said to have performed differently ("better" or "worse") on the two tests if the difference between test scores is more than twice the square-root of the sum of the squares of the standard error of measurement (that is, the standard error of the difference) of each of the two test scores. A student is classified as "likely better" or "likely worse" if this difference is larger than one s.e.diff. On the Rasch metric, the standard error of measurement was calculated for each test score and includes an inflation for unusual response patterns. Standard errors of measurement are provided by the ITBS test publisher

for developmental standard scores but not grade equivalents. An approximate standard error of measurement for GEs was deduced for each test level based on the relationship of the GE with the developmentalstandard score, as found in the conversion table provided for Form M (1996).

[9]As noted earlier, "on grade level" is test form and level dependent. As a result, the national norm points for any given grade in the equated metric vary some across the forms. For purposes of creating some cross walk between the two scales we chose to average the 1987 to 1996 scores to create the anchoring points for our content referenced metric. That is, a scale score of 20 is equivalent to a GE of 1.8 on level 7, based on these averages, and a scale score of 80 is equivalent to a GE of 8.8 on level 14.

[10]The difficulty score for a passage is based on the median difficulty of the items associated with that passage. See Luppescu (1996) for an in-depth analysis of the factors that contribute to both item and passage difficulty in the ITBS reading series.

[11]We adopt a conventional definition for our level of mastery. A student is said to master an item if he has greater than a 75 percent chance of responding correctly to that item. On the logit scale, this translates to a student ability estimate that is 1.1 higher than the calibrated difficulty of the item.

[12]Lee (1992), Lee and Wright (1992), Luppescu (1996).

[13]See Kerbow (1996).

[14]Meyer (1993) reported on efforts to characterize school improvement from testing programs that employ annual school aggregate scores. The results were startling. Individual schools can actually be making improvements (i.e., individual student achievement gains are increasing), but these increasing gains can be totally obscured by factors such as those discussed in this report.

[15]An important consequence of our definition of school productivity is that only students who are tested in the same school for two consecutive years are included in our analysis. In grades four to seven from 1994 through 1996, we averaged about 80 percent of students who are tested in reading. Of these students, only 84 to 87 percent have useable test strings. Of this group of students, about 5 percent were further excluded because on one or both occasions a student was in a special education category or had taken an off-level test. Thus, using reasonable criteria, our accountability system employs only about 60 percent of the children tested on average. This raises a highly sensitive issue of who among the diverse population of children served by the system actually enters our analyses of school accountability.

[16]We were constrained to follow this procedure in our analyses because of the data then available in the CPS test score files. Because of the analysis presented in this report, the CPS has now merged its test score and administrative history files. As a result, it is now possible to assign the learning gains of a student to a school if he or she was enrolled by some date (e.g., October 1 of the test year) or spent a minimum amount of time there (e.g., at least 100 institutional days). Since the majority of school changes occur over the summer, the use of such a procedure would result in more than 90 percent of the students entering into the accountability analysis each year.

[17]This choice was driven primarily by testing design considerations. The gains recorded in 1991 are based on two new and different forms of the ITBS. The pattern of different test forms persists in subsequent years. In contrast, the gains through 1989 are based on repeated

administrations of Form 7 of the ITBS that the CPS used throughout the 1980s. The 1990 gain also involves Form 7 data as the input status. Because this test form had been used for a very long period of time, concerns had been raised about the security of the test items and whether some teachers and schools might actually be teaching the test. We first considered using the 1990 data as the base year as this is formally the first implementation year for the 1988 reform. The estimated base year results in 1990 looked spurious, however, in many schools. Since subsequent productivity trends are anchored in the base year results, it seemed prudent to us to switch to 1991 instead.

[18]Our adjustment strategy reflects two principal concerns affecting our analysis. First, we control for possible residual test-form effects on trend estimates. Recall that from 1990 forward, the CPS changed test forms each year. Form 7 was so familiar prior to 1990 that, in many schools, student gains for this period could well harbor long-term practice effects. In contrast, in 1990 we observe a systemwide plunge in gains as a new and unfamiliar test was given for the first time.

Thus, we chose to introduce in our analyses model-specific "form effects" for the test data used prior to 1991. As a result, differences in test administration prior to 1991 will not affect our overall trend estimates.

Second, we adjust each school's true gain by how its student composition (such as minority status, bilingual status, retention rate, and percentage old for their grade) compares with the prereform status defined as the average of the school's 1988 and 1989 level for the grade. Differences among adjusted and unadjusted gain trend estimates are minor. They correlate at .98 or better irrespective of grade.

[19]Formally, the LGI is computed as follows:

LGI = 100 x (5 x School Gain Trend Estimate) / (1991 System Gain Estimate)

An LGI compares the gains made in a grade for the five years (1992 through 1996) as a percentage of the observed gain systemwide for that grade in the 1991 base year.

[20]Meyer (1993) made a strong case for measuring school performance using gain trends. See Bryk, Deabster, Easton, Luppescu, and Thum (1994) for an early perspective focusing on the Chicago Public Schools.

[21]Technically, the value added to student learning is different from the observed learning gains in that the value-added estimate is adjusted for other factors affecting observed outcomes. For simplicity of presentation, we use the terms interchangeably here. All of the results presented in this paper are adjusted, i.e., value-added estimates.

[22]School improvement is defined by two factors. A clearly improving school should exhibit an increase in productivity over time as well as increasing outputs over time. For each educational sub-unit, therefore, we classify each school's improvement as:

Up: Total gain trend (across grades in grouping) is +.02 logits or higher. Over the five-year time span (from 1992 to 1996), this amounts approximately to a 15 percent improvement in learning gains (LGI). At the same time, the output trend at the end of the grade group is +.02 or higher.

Tending Up: Schools in this group show a +.02 output trend for the last grade in the grade group and no change in their total gain trends (-.02, +.02), or a total of +.02 or higher in gain trends without noticeable changes in the final output trends (again, from -.02 to +.02).

Mixed Output: Output trends at the last grade in the grade group is equal to .02 (or better), but the total gain trend is -.02 or less.

Mixed, Productivity: The total gain trend is .02 or better, but the output trend at the last grade in the grade group is equal to -.02 or less.

No Change: No discernable output trend or total gain trends beyond the -.02 to .02 range.

Tending Down: Schools in this group show a -.02 or lower output trend at the last grade in the grade group and no change in their total gain trends, or a total of -.02 or less in gain trends without noticeable changes in the final output trends.

Down: Total gain trend (across grade-groups) is -.02 or lower. This amounts approximately to a 15 percent loss in the LGI over the grades for the five years (from 1992 to 1996) approximately. At the same time, the output trend at the end of the grade group is -.02 or lower.

[23]Correlations of gain trend estimates for adjacent grades tend to be negative. For example, the ITBS reading estimates for grades 3 and 4 (-.16, $p<.001$), grades 4 and 5 (-.12, $p<.001$), grades 5 and 6 (-.22, $p<.001$), and grades 6 and 7 (-.12, $p<.02$).

[24]For a further discussion of the problems of school-based professional community see Kruse, Louis, and Bryk (1995). For an analysis of these issues in Chicago schools, see Sebring, Bryk, Easton, Luppescu, Thum, Lopez, and Smith (1995).

[25]Standardized testing in the CPS is optional in grades one and two. A substantial number of eligible students are not tested at grade one. By grade two, however, there is almost universal administration of the ITBS. Thus, grade three becomes the first grade at which entry and exit data exist and, as a result, a grade productivity profile can be computed for almost all schools.

In principle, other choices of grade groupings could be used such as 2, 3; 4, 5, 6; and 7,8. The splits at grades 6 and 8 seemed natural given that these are also the grades for state mandated reading and math assessments. In this way, the local accountability system would parallel the state approach. The treatment of grade four was a main issue for us. Our first inclination was to group it with grades five and six, for the reasons about the state testing just articulated. Our analysis of auto-correlations across grade levels found relatively strong associations between grades three and four, however. These results suggest that these two grade levels may work as a system in many schools and that we should treat them together as a unit.

For schools with grade level structures other than kindergarten through eight, the school productivity summary consists of only the relevant components. For example, in a kindergarten through five school, the summary would include only the second grade output trend and the grades three and four summary. Data from grade five would not be used. If we wanted to undertake a more detailed analysis of kindergarten through five schools, it would be possible to reconfigure the summaries, e.g., a grade grouping of three, four, and five, and examine these. In general, any application of a standard reporting system to a school system that has as many different grade structures as in Chicago, confronts the dilemma that no system works best for all. We are convinced, however, that the one used here works best for most.

[26]The systemwide gains for 1988, 1989, and 1990 were averaged separately for each grade level to form the baseline (i.e., used as the denominator) for estimating these percent improvement statistics.

[27]Any kernel smoother would suffice. Here, we used a symmetric *k*-nearest neighbor linear least squares procedure that is implemented in the S-PLUS© function *supsmu*.

[28]We note that the output trend from a given grade is not always identical to the input trend for the following grade. This occurs for several reasons. First, year to year mobility, as noted earlier in the report, means that somewhat different student populations exist for each grade profile. Typically, about 20 percent of the students exit a school each year and are replaced by a new 20 percent. Second, the population of schools shifts across grades, as not all schools are kindergarten through eighth grade. Since the system profile is an average of individual school profiles, this, too, can effect a difference, especially at the upper grades (five and above), where most of these changes occur. Third, the output trend for a grade includes 1996 data, whereas the corresponding input trend only goes through 1995. Similarly, the input trend starts with 1987, but the first output information is 1988. Since the time series are relatively short, these data differences at "the ends" can create leverage points that affect trend estimates. We examined these alternative explanations for the differences observed in Figures 17 and 18 and concluded that no one factor dominates. Depending upon the particular input-output combination, any one of the three may be the causal agent.

[29]Linn, Koretz, and Baker (1996), for example, raise questions about the validity of standardized test score trends as indicators of broad instructional improvements when these same data are used for high stakes accountability.

[30]This will be the focus of a subsequent report in the Examining School Productivity series. From the preliminary analyses to date, it is clear that such patterns exist.

[31]Storey, Easton, Sharp, Steans, Ames, and Bassuk (1995).

REFERENCES

Bryk, Anthony S., Paul E. Deabster, John Q. Easton, Stuart Luppescu, and Yeow Meng Thum (1994). Measuring Achievement Gains in the Chicago Public Schools. *Education and Urban Society, 26,* 306-319.

Bryk, Anthony S., and Steven W. Raudenbush (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage Publications.

Bryk, Anthony S., and Herbert I. Weisberg (1977). Use of the Non-equivalent Control Group Design When Subjects Are Growing. *Psychological Bulletin, 84* (5):950-62.

Harris, Chester W. (Ed.). (1963). *Problems in Measuring Change.* Proceedings of a conference sponsored by the Committee on Personality Development in Youth of the Social Science Research Council. Madison, WI: University of Wisconsin Press.

Kerbow, David (1996). *Pervasive Student Mobility: A Moving Target for School Improvement*. Chicago: Chicago Panel on School Policy.

Kruse, Sharon D., Karen S. Louis, and Anthony S. Bryk (1994). Building Professional Community in Schools. *Issue Report No. 6*. Madison, WI: Center on Organization and Restructuring of Schools, pp. 3-6.

Lee, Ong Kim (1992). *Measuring Mathematics and Reading Growth*. Unpublished doctoral dissertation, University of Chicago.

Lee, Ong Kim, and Benjamin D. Wright (1992). *Mathematics and Reading Test Equating*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Linn, Robert L., Daniel M. Koretz, and Eva L. Baker (1996). *Assessing the Validity of the National Assessment of Educational Progress: Final Report of the NAEP Technical Review Panel*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Luppescu, Stuart (1996). *Virtual Equating: An Approach to Reading Test Equating by Concept Matching of Items.* Unpublished doctoral dissertation, University of Chicago.

Meyer, Robert (1993). *Can Schools Be Held Accountable for Good Performance? A Critique of Common Educational Performance Indicators.* Working paper. Harris Graduate School of Public Policy Studies, University of Chicago.

Meyer, Robert (1996). Value-added Indicators of School Performance. In Eric A. Hanushek and Dale W. Jorgenson (Eds.), *Improving America's Schools: The Role of Incentives.* Washington, DC: National Academy Press, pp. 197-223.

Rogosa, David R., D. Brand, and Michele F. Zimowski (1982). A Growth Curve Approach to the Measurement of Change. *Psychological Bulletin, 90*: 726-48

Rogosa, David R., and John B. Willett (1985). Satisfying a Simplex Structure Is Simpler Than It Should Be. *Journal of Educational Statistics, 10* (2): 99-107.

Sebring, Penny B., Anthony S. Bryk, John Q. Easton, Stuart Luppescu, Yeow Meng Thum, Winifred A. Lopez, and BetsAnn Smith (1995). *Charting Reform: Chicago Teachers Take Stock.* Chicago: Consortium on Chicago School Research.

Storey, Sandra, John Q. Easton, Thomas C. Sharp, Heather Steans, Brian Ames, and Alicia Bassuk (1995). *Chicago's Public School Children and Their Environment.* Chicago: Chicago Public Schools with Chicago Urban League and the Latino Institute.

Willett, John B. (1989). Questions and Answers in the Measurement of Change. *Review of Research in Education, 15 (1988-1989),* pp. 345-422.

Appendix: Estimating Trends in School Productivity

Estimation Model

We applied the basic analysis model described below for each grade.

1. We denote the Rasch estimated scale score (i.e., the equated metric) obtained for student *j* from school *k* at time point *t* by y_{jkt}. Also available is an estimate of the precision, $1/s_{jkt}$ associated with each scale measure. This is based on the real standard error of measurement, s_{jkt}, which is the nominal standard error inflated for scale misfit. Students in school *k* are subscripted $j = 1, 2, \cdots, n_k$. Schools are indexed $k = 1, 2, \cdots, N$. Each time point, *t*, may run from *1988* to *1996.* Only students with at least two consecutive time points, a test score (y_{jkt}) at time *t* and an input score ($y_{jk(t-1)}$) at time $t - 1$, are included in our analysis.

2. Given the student's test scores, we proceed with a parameterization at level-1 of a measurement model that estimates a student's true input ability and true gain for year *t:*

   (1) $y^*_{jkt} = a^*_{1jkt} \cdot \pi_{1jkt} + a^*_{2jkt} \cdot \pi_{2jkt} + e^*_{jkt}$.

   Note that: $y^*_{jkt} = \frac{y_{jkt}}{s_{jkt}}$. Furthermore,

   $$a^*_{1jkt} = \frac{1}{s_{jkt}} \quad \text{and} \quad a^*_{2jkt} = \begin{cases} 0 & \text{if } t \text{ is the input data.} \\ \frac{1}{s_{jkt}} & \text{if } t \text{ is the test year data.} \end{cases}$$

   Thus $e^*_{jkt} \sim N(0,1)$ and π_{1jkt} is the student's *true input ability* while π_{2jkt} is an estimator of the *true gain*.

3. For all students in a given grade $j = 1,2,\cdots, n_k$ in all *k* schools, we estimate the school input and gain trends at level 2 by

   (2) $$\begin{matrix} & \\ \textit{input} \\ \textit{gain} \end{matrix} \begin{bmatrix} \pi_{1jkt} \\ \pi_{2nkt} \end{bmatrix} = \begin{bmatrix} 1 & (t-1991) & 0 & 0 \\ 0 & 0 & 1 & (t-1991) \end{bmatrix} \begin{bmatrix} \beta_{11k} \\ \beta_{12k} \\ \beta_{21k} \\ \beta_{22k} \end{bmatrix} + \begin{bmatrix} r_{1jkt} \\ r_{2jkt} \end{bmatrix}.$$

In this model, we center *t* at 1991, so that β_{11k} estimates the average input in 1991. Similarly, β_{21k} is the average value-added in 1991. β_{12k} is the input trend for the grade and β_{22k} is the gain trend for the grade. We assume that the errors in (2) are correlated, $\mathbf{r}_{jkt} \sim N(\mathbf{0}, \mathbf{T}_{\pi})$. Equation (2) is sometimes called a "cross-domain" growth model because it tracks two closely related short time-series simultaneously.

Student-level covariates. In addition, we adjust the school effect vector, β_k , for five student composition characteristics that might be changing over time. The five student characteristics are bilingual "C" status, too old for a grade, if the student is retained, if the student is white, and if the student is African-American. Each of these covariates is deviated around the school mean for 1988 and 1989, or generally $(X_{sjk,t} - \overline{X}_{s\cdot k,0})$ where $\overline{X}_{s\cdot k,0}$ denotes the 1988–1989 baseline. The associated coefficients estimate a time-varying school composition effect. We assume that these level 2 coefficients are constant across years and across schools.

Form effect adjustments. The final version of this model adjusts for form effect associated with 1988 and 1989 tests (Form 7) and for 1990. This is accomplished by incorporating two additional dummies. The first takes on a value of "1" if the test year for a student is 1988 or 1989 and "0" at all other times. Similarly, a second dummy variable is coded "1" if the test data are for 1990 and is coded "0" for all other times. In summary, we adopted the following coding scheme for the trend component (slope) and the form effects (Form 7 and CPS 90):

	Year								
	88	**89**	**90**	**91**	**92**	**93**	**94**	**95**	**96**
Slope (1991)	-2	-2	-1	0	1	2	3	4	5
Form 7	1	1	0	0	0	0	0	0	0
CPS 90	0	0	1	0	0	0	0	0	0

This results in a final level 2 model for π_1 and π_2 of:

$$\pi_{jkt} = \beta_{21k} + (t - 1991) \cdot \beta_{22k} + FORM\,7 \cdot \beta_{23} + CPS90 \cdot \beta_{24} + \sum_{s}(X_{sjk,t} - \overline{X}_{s\cdot k,0}) \cdot \beta_{2s} + r_{2jkt}. \quad (3)$$

4. School input and gain trend estimates, β_k , are expected to vary from one school to the next according as:

$$(4)\qquad \begin{matrix} \text{1991 } \textit{input} \\ \textit{input trend} \\ \text{1991 } \textit{gain} \\ \textit{gain trend} \end{matrix} \begin{bmatrix} \beta_{11k} \\ \beta_{12k} \\ \beta_{21k} \\ \beta_{22k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{bmatrix} + \begin{bmatrix} u_{11k} \\ u_{12k} \\ u_{21k} \\ u_{22k} \end{bmatrix}$$

Here, the quantities in γ are the systemwide input and gain trend estimates. Across schools, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{T}_\beta)$, and future attempts to identify the correlates of school performance entails adding plausible school-level covariates to this basic level-3 model in a fashion analogous to the level-2 model in (3).

Equation (4) yields *empirical Bayes* estimates of β_k for each school. They are weighted composite estimators that take into account information about school *k* relative to other schools in the system. If the information for school *k* is relatively weak, β_k is *shrunk* towards the system average γ. This estimator thus efficiently utilizes all of the available information to provide predictions for each school.

Discussion of the Estimation Model

Appropriate techniques for research on change have long perplexed behavioral scientists (see, for example, Harris, 1963). The methodological studies of Rogosa, Brand, and Zimowski (1982), Rogosa and Willett (1985), and Willett (1988) have greatly clarified these problems. Briefly, they demonstrate that if individual growth is linear (or approximately so), then the gain score is the unbiased estimator of the instantaneous growth rate. Even if the underlying growth demonstrates some curvature, the gain score will estimate the average growth rate over the time period of study. In contrast, the covariance model can be seriously biased in studies of school and program effects on individual growth (Bryk and Weisberg, 1977).

For these reasons, the analysis model employed in this research is based on gain scores rather than a covariance adjustment approach. It develops out of the growth modeling strategy explicated in Bryk and Raudenbush (1992). Since we are studying students' academic development over a one-year period of time, the use of a gain score seems quite appropriate. We note that over the full 100-point metric, individual growth displays some deceleration at the upper ability levels. Within any grade slice, however, we found no evidence

of non-linearity in an analysis of a subset of eight-year longitudinal data on students in the same schools.

A key to use of a gain score strategy is an appropriate quantitative metric for measuring change. The content-referenced scale developed in this research is critical in this regard. Unlike grade equivalents, the equated test score metric yields an interval measurement scale based on the relative difficulty of the items. Such interval measurement is necessary for quantitative studies of change.

We also note that our analysis is based on the latent initial status and gain scores rather than the observed data. This is accomplished by the use of a measurement model at level 1. The problems of statistical artifacts due to correlated errors in observed input status and gains are thus eliminated. Also, unlike a covariance model, all of the test data appears on the left hand side of the equations; as a result, we also avoid the problems of key adjustor variables that are fallible covariates. This is another strength of this modeling approach.

See the Consortium's world wide web site for productivity profiles of each Chicago Public Elementary School. These data are considered public information.

http://www.consortium-chicago.org

Appendix: Estimating Trends in School Productivity

Estimation Model

We applied the basic analysis model described below for each grade.

1. We denote the Rasch estimated scale score (i.e., the equated metric) obtained for student *j* from school *k* at time point *t* by y_{jkt}. Also available is an estimate of the precision, $1/s_{jkt}$ associated with each scale measure. This is based on the real standard error of measurement, s_{jkt}, which is the nominal standard error inflated for scale misfit. Students in school *k* are subscripted $j = 1, 2, \cdots, n_k$. Schools are indexed $k = 1, 2, \cdots, N$. Each time point, *t*, may run from *1988* to *1996.* Only students with at least two consecutive time points, a test score (y_{jkt}) at time *t* and an input score ($y_{jk(t-1)}$) at time $t - 1$, are included in our analysis.

2. Given the student's test scores, we proceed with a parameterization at level-1 of a measurement model that estimates a student's true input ability and true gain for year *t:*

$$(1)\ y^*_{jkt} = a^*_{1jkt} \cdot \pi_{1jkt} + a^*_{2jkt} \cdot \pi_{2jkt} + e^*_{jkt}.$$

Note that: $y^*_{jkt} = \frac{y_{jkt}}{s_{jkt}}$ Furthermore,

$a^*_{1jkt} = \frac{1}{s_{jkt}}$ and $a^*_{2jkt} = \begin{cases} 0 & \text{if } t \text{ is the input data.} \\ \frac{1}{s_{jkt}} & \text{if } t \text{ is the test year data.} \end{cases}$

Thus $e^*_{jkt} \sim N(0,1)$ and π_{1jkt} is the student's *true input ability* while π_{2jkt} is an estimator of the *true gain*.

3. For all students in a given grade $j = 1,2,\cdots, n_k$ in all *k* schools, we estimate the school input and gain trends at level 2 by

$$(2)\ \begin{matrix} input \\ gain \end{matrix} \begin{bmatrix} \pi_{1jkt} \\ \pi_{2nkt} \end{bmatrix} = \begin{bmatrix} 1 & (t-1991) & 0 & 0 \\ 0 & 0 & 1 & (t-1991) \end{bmatrix} \begin{bmatrix} \beta_{11k} \\ \beta_{12k} \\ \beta_{21k} \\ \beta_{22k} \end{bmatrix} + \begin{bmatrix} r_{1jkt} \\ r_{2jkt} \end{bmatrix}.$$

In this model, we center *t* at 1991, so that β_{11k} estimates the average input in 1991. Similarly, β_{21k} is the average value-added in 1991. β_{12k} is the input trend for the grade and β_{22k} is the gain trend for the grade. We assume that the errors in (2) are correlated, $\mathbf{r}_{jkt} \sim N(\mathbf{0}, \mathbf{T}_{\pi})$. Equation (2) is sometimes called a "cross-domain" growth model because it tracks two closely related short time-series simultaneously.

Student-level covariates. In addition, we adjust the school effect vector, β_k, for five student composition characteristics that might be changing over time. The five student characteristics are bilingual "C" status, too old for a grade, if the student is retained, if the student is white, and if the student is African-American. Each of these covariates is deviated around the school mean for 1988 and 1989, or generally $(X_{sjk,t} - \overline{X}_{s\cdot k,0})$ where $\overline{X}_{s\cdot k,0}$ denotes the 1988–1989 baseline. The associated coefficients estimate a time-varying school composition effect. We assume that these level 2 coefficients are constant across years and across schools.

Form effect adjustments. The final version of this model adjusts for form effect associated with 1988 and 1989 tests (Form 7) and for 1990. This is accomplished by incorporating two additional dummies. The first takes on a value of "1" if the test year for a student is 1988 or 1989 and "0" at all other times. Similarly, a second dummy variable is coded "1" if the test data are for 1990 and is coded "0" for all other times. In summary, we adopted the following coding scheme for the trend component (slope) and the form effects (Form 7 and CPS 90):

	Year								
	88	**89**	**90**	**91**	**92**	**93**	**94**	**95**	**96**
Slope (1991)	-2	-2	-1	0	1	2	3	4	5
Form 7	1	1	0	0	0	0	0	0	0
CPS 90	0	0	1	0	0	0	0	0	0

This results in a final level 2 model for π_1 and π_2 of:

(3)

4. School input and gain trend estimates, β_k , are expected to vary from one school to the next according as:

$$
\begin{array}{lllllllll}
(4) & 1991\ \textit{input} & \beta_{11k} & 1 & 0 & 0 & 0 & \gamma_{11} & u_{11k} \\
 & \textit{input trend} & \beta_{12k} & 0 & 1 & 0 & 0 & \gamma_{12} & u_{12k} \\
 & 1991\ \textit{gain} & \beta_{21k} & 0 & 0 & 1 & 0 & \gamma_{21} & u_{21k} \\
 & \textit{gain trend} & \beta_{22k} & 0 & 0 & 0 & 1 & \gamma_{22} & u_{22k}
\end{array}
$$

Here, the quantities in γ are the systemwide input and gain trend estimates. Across schools, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{T}_\beta)$, and future attempts to identify the correlates of school performance entails adding plausible school-level covariates to this basic level-3 model in a fashion analogous to the level-2 model in (3).

Equation (4) yields *empirical Bayes* estimates of β_k for each school. They are weighted composite estimators that take into account information about school *k* relative to other schools in the system. If the information for school *k* is relatively weak, β_k is *shrunk* towards the system average γ. This estimator thus efficiently utilizes all of the available information to provide predictions for each school.

Discussion of the Estimation Model

Appropriate techniques for research on change have long perplexed behavioral scientists (see, for example, Harris, 1963). The methodological studies of Rogosa, Brand, and Zimowski (1982), Rogosa and Willett (1985), and Willett (1988) have greatly clarified these problems. Briefly, they demonstrate that if individual growth is linear (or approximately so), then the gain score is the unbiased estimator of the instantaneous growth rate. Even if the underlying growth demonstrates some curvature, the gain score will estimate the average growth rate over the time period of study. In contrast, the covariance model can be seriously biased in studies of school and program effects on individual growth (Bryk and Weisberg, 1977).

For these reasons, the analysis model employed in this research is based on gain scores rather than a covariance adjustment approach. It develops out of the growth modeling strategy explicated in Bryk and Raudenbush (1992). Since we are studying students' academic development over a one-year period of time, the use of a gain score seems

quite appropriate. We note that over the full 100-point metric, individual growth displays some deceleration at the upper ability levels. Within any grade slice, however, we found no evidence of non-linearity in an analysis of a subset of eight-year longitudinal data on students in the same schools.

A key to use of a gain score strategy is an appropriate quantitative metric for measuring change. The content-referenced scale developed in this research is critical in this regard. Unlike grade equivalents, the equated test score metric yields an interval measurement scale based on the relative difficulty of the items. Such interval measurement is necessary for quantitative studies of change.

We also note that our analysis is based on the latent initial status and gain scores rather than the observed data. This is accomplished by the use of a measurement model at level 1. The problems of statistical artifacts due to correlated errors in observed input status and gains are thus eliminated. Also, unlike a covariance model, all of the test data appears on the left hand side of the equations; as a result, we also avoid the problems of key adjustor variables that are fallible covariates. This is another strength of this modeling approach.

Academic Productivity of Chicago Public Elementary Schools

A Technical Report Sponsored by the Consortium on Chicago School Research

Mission

The Consortium on Chicago School Research is an independent federation of Chicago area organizations that conducts research activities designed to advance school improvement in Chicago's public schools and to assess the progress of school reform. The Consortium aims to encourage:

- Broad access to the research agenda-setting process;
- Collection and reporting of systematic information on the condition of education in the Chicago Public Schools;
- High standards of quality in research design, data collection, and analysis; and
- Wide dissemination and discussion of research findings.

Researchers from many different settings who are interested in schooling and its improvement come together under the umbrella of the Consortium. Its deliberate multipartisan membership includes faculty from area universities, leadership from the Chicago Public Schools, the Chicago Teachers Union, education advocacy groups, the Illinois State Board of Education, and the North Central Regional Educational Laboratory, as well as other key civic and professional leaders.

The Consortium views research not just as a technical operation of gathering data and publishing reports, but as a form of community education. The Consortium does not argue a particular policy position. Rather, it believes that good policy results from a genuine competition of ideas informed by the best evidence that can be obtained. The Consortium works to produce such evidence and to ensure that the competition of ideas remains vital.

Consortium on Chicago School Research
1313 East 60th Street Chicago, Illinois 60637 (773) 702-3364
http://www.consortium-chicago.org