4546

When the Test Fails:

The Invalidity of Assumptions of Normative

Stability with Above-Average Populations

A Dissertation Submitted to the Graduate Division of the University of Hawai'i
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in

Educational Psychology

December 2004

By:
Katherine A. Tibbetts

Dissertation Committee:

Morris K. Lai, Chairperson
Frederick T. Bail
Paul R. Brandon
Mary I. Martini
W. James Popham

ABSTRACT

This dissertation examines the validity of inferences from longitudinal test data about the effectiveness of educational programs serving high-achieving students. Declines in test scores were noted over the span of Grades 1 to 4 across multiple cohorts and achievement tests. Four potential sources of change that are exogenous to the school were examined for their contribution to the changes in scores: regression toward the mean, test ceilings, changes in the content of the tests across grades, and the instability of the normative performance of young children. The study relies on informational resources available from the test publishers and basic descriptive and inferential statistics. Thus, the methods are transportable to other educational program evaluations.

The context for the study is a school with a population of above-average achievement. The relevance of the study extends to other schools serving similar populations, programs for the academically gifted, and to schools serving students of below-average achievement.

This dissertation provides insights into the reasonableness of our expectations of stability of students' performance relative to a norm group. It advances the methodology for estimation of regression effects by combining two strategies: one to estimate the effect of test reliability, and the other to estimate the effect of imperfect correlations between assessments.

The findings indicate that all four sources of change studied could contribute to the decline in scores. First, statistical regression can largely explain the observed changes. Second, the impact of the test ceilings and the inability of the tests to adequately measure differences in the achievement of the students is documented. Third, evidence of the changes in the content of the tests across grade levels is discussed. Finally, a proxy for learning curves suggests that the pattern of change in scores at the school may be typical for students like those at this school.

The findings from this study make obvious the dangers of reliance on test scores as the main indicator of academic effectiveness. The findings bring home, in a very immediate way, the need to move beyond descriptive studies to more rigorous methods when making evaluative judgments.

CONTENTS

TABLES

Table Page

FIGURES

Figure Page

CHAPTER 1: INTRODUCTION

> Educational and psychological testing and assessments are among the most important contributions of behavioral science to our society....The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route to broader and more equitable access to education and employment. The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 1999, p. 1).

In 1997, a private school trustee announced to readers of a local newspaper of wide circulation that the school with which she was associated was harming the children enrolled there. In her words, "the longer they stay, the worse they do."[1]

While this statement was based in part on misrepresentation of some data, the fact remains that longitudinal data for elementary school students at this school did show a pattern of declining test scores from Grade 1 through Grade 4. Were the claims and concerns valid? Should those responsible for education of children in other schools with similar patterns be concerned? The pattern of scores observed at this school is hardly unique (personal communication, H. D. Wainer, January 1999).

There is a wide spread assumption that successful students and schools maintain or improve their position on norm-referenced achievement tests. This assumption is so pervasive that it underlies nearly all evaluations of educational programs that include these types of assessments as indicators of school or program effectiveness. At times, the

[1] Source omitted to protect the privacy of the school.

assumption that we have tools and methods to adequately measure growth with young children is challenged. For example, in its recommendation to the National Educational Goals Panel, the Goal 1 Early Childhood Assessments Resource Group (1998) stated

> Before age 8, standardized achievement measures are not sufficiently accurate to be used for high-stakes decisions about individual children and schools. Therefore, high-stakes assessments intended for accountability purposes should be delayed until the end of third grade (or preferably fourth grade). (p. 29)

Notably, the reauthorization of the Elementary and Secondary Schools Education Act in 2001 (more commonly known as the "No Child Left Behind Act of 2001" or "NCLB") requires annual testing of student progress beginning at third grade (U.S. Department of Education, 2001, 2002). While the rationale for starting at third grade is not provided in the Act, it is consistent with the recommendations to the National Education Goals Panel cited above.

Purpose of This Study

The purpose of this study is to provide a coherent, accessible explanation of what the observed changes in test scores do, and do not, indicate about students' achievement over time and the quality of the school's curriculum and instruction. This understanding is important for the school in question, for other schools serving students with comparable achievement profiles (e.g., with average scores in stanines 7 to 9), and for schools with average scores in stanines 1 to 3. It also provides information that can be used to address the concerns parents and students may have about their scores. Most specifically, the study presents information important for those involved with evaluation and school improvement in these contexts.

Study Context

Admissions Selection and Test Scores

Selection Process

The school examined in this study admits students at Kindergarten and Grades 4, 7, and 9. Admission is highly competitive at all entry points. For example, in a typical year, only 1 of every 14 applicants to Kindergarten is admitted, and 97% of those admitted accept the invitation to enroll. The admissions selection process begins with families self-selecting into the process by requesting and completing an application form. The Kindergarten applicants are then invited to participate in a test of cognitive achievement that includes verbal and quantitative skills. (The school requires that the name of the test remain confidential to ensure no applicants have an unfair advantage from prior exposure to the test). The test is individually administered by trained examiners, includes a number of hands-on tasks appropriate for use with preschool-age children, and uses age-based norms with intervals of three months. The applicants are ranked based on their prereading and quantitative scores, and a pool of applicants that is twice the size of the group to be admitted is identified. This group is stratified by gender and other socio-demographic characteristics that the school considers important to creating a diverse student population.

The members of this smaller pool of applicants are invited to participate in the observation phase of the process. In this phase they spend two hours at the school on a Saturday with up to 19 other applicants. During the observation, they engage in a variety of individual and small- and large-group activities that typically occur in early childhood education settings (e.g., morning circle, story time, outdoor play, and home center). They are observed concurrently by four raters who evaluate each child holistically, looking at a range of behaviors that includes, but is not limited to, the children's interactions with adults and peers, interest and engagement in the activities, creativity, and problem-solving. Based on the observers' ratings, half of the children are invited to enroll in the following school year, and half are placed on a waiting list. As with invitations to the observation phase, invitations to enroll are stratified by gender and other socio-demographic characteristics that the school considers important to creating a diverse student population.

This study is based on the three cohorts of entering kindergartners that completed Grade 6 in 1999, 2000, and 2001 (cohorts A, B, and C, respectively). These cohorts were selected because they had the most complete data relevant to the study.

Admissions Test Scores

The students included in this study had admissions test scores that ranged from the 13th to the 99th percentile in prereading with a mean score at the 85th percentile (in the 7th stanine). Their quantitative concepts scores ranged from the 19th to the 99th percentile, and the mean score in this domain was equivalent to the 73rd percentile (in the 6th stanine). The median scores for prereading and quantitative concepts were at the 84th and 74th percentiles respectively. The distributions of their admissions test scores are shown in Table 1. To simplify presentation, test scores are presented using stanines.

Table 1
Distribution of Admissions Test Scores by Stanine (Percentage of Students)

Test	Percentile Rank of Mean	Stanine								
		1	2	3	4	5	6	7	8	9
Prereading	85			1	1	8	23	30	31	8
Quantitative Concepts	73			2	4	19	37	24	12	3

Note. Percentages do not sum to 100 due to rounding.

Prior School Experience.

Another significant characteristic of the children admitted to Kindergarten at this school is their previous school experience. Because these data are not available for the cohorts included in this study, I offer proxy data from the children completing Grade 6 in 2003. In that cohort, 96 percent of the children had experience in a formal educational setting prior to entry to Kindergarten. There are no significant differences in age at admission or admissions test scores between the 2003 cohort and those in this study. This suggests the data on prior school experience is also generalizable to the earlier cohorts.

To be admitted to Kindergarten, girls must be five years old by September 30 and boys must be five by June 30 of the year in which they enter. The local public schools require that children entering Kindergarten be five by December 31. Thus, on average, girls at the school are about three months older than public school students in the same grade, and boys are about six months older. Fifty-five percent of the students admitted to Kindergarten had birthdays that would have allowed them to enroll in Kindergarten in the local public schools for the school year prior to entry to Kindergarten at this school.[2]

[2] The 55 percent comprises 63 percent of the boys admitted and 46 percent of the girls.

Achievement Test Scores: Kindergarten Through Grade 6

All students were tested at the beginning and end of their Kindergarten year with the Peabody Picture Vocabulary Test-Revised (PPVT-R). Like the admissions test, this test of receptive vocabulary uses age-based norms. The norm intervals are two months wide through age six years and three months wide through age 17. The PPVT-R is designed to be used with children as young as two years and six months and has norms through age 40 (Dunn, 1981). This is significant because the test has no floor or ceiling for the study population.

Children who completed Grade 6 in 1999 (Cohort A) were tested on reading and mathematics in the spring of Grades 1 through 6. Those who completed Grade 6 in 2000 (Cohort B) were tested in spring of Grades 1 through 5. Those who completed Grade 6 in 2001 (Cohort C) were tested in spring of Grades 1 through 4. At all these points, they were tested in reading and mathematics. The Comprehensive Testing Program–Version III (CTP III) was used through 1996, and the Stanford Achievement Test–Ninth Edition (SAT-9) was used from 1997 through 1999. No standardized achievement tests were administered in spring 2000. When standardized testing was resumed, the CTP III was administered in the fall (beginning with fall 2000). The tests administered to each cohort are shown in Table 2.

Table 2
Tests Used With Each Cohort by Grade Level

	Test Administered						
Cohort	K	1	2	3	4	5	6
A (1999)	PPVT-R	CTP III	CTP III	CTP III	SAT-9	SAT-9	SAT-9
B (2000)	PPVT-R	CTP III	CTP III	SAT-9	SAT-9	SAT-9	—
C (2001)	PPVT-R	CTP III	SAT-9	SAT-9	SAT-9	—	—

Note. No testing was conducted in spring 2000 (Grade 6 for Cohort B and Grade 5 for Cohort C). Testing for Cohort C in Grade 6 was conducted in the fall using the CTP III.

Changes in the content of the tests from one grade level to the next within a single test battery are a source of variability in the longitudinal test scores and confound the interpretation of change scores (Hoover, 1984; Paris & Lindauer, 1982). The change from the CTP III to the SAT-9 introduces another source of variability, as each test defines the domains tested somewhat differently. In addition, each test is normed independently, and variation due to differences in the norm groups affects the comparability of the results.

In this study, the results of the two tests are combined by relying on normal curve equivalent (*NCE*) scores as the metric for analysis. Although less than optimal, this is not unreasonable because the results on the two tests, using the national norms, are comparable. The results of analyses of the effect of combining scores from the two tests are presented in Chapter 3.

Kindergarten PPVT-R Scores.

At the beginning of the school year, the PPVT-R scores for the combined cohorts ranged from the 14th to the 99th percentile, and the mean score was equivalent to the 75th percentile (in the 6th stanine). Over the course of the Kindergarten year, as a group the students made significant gains in receptive vocabulary. The spring test scores ranged from the 7th to the 99th percentile, and the mean score was equivalent to the 84th percentile (in the 7th stanine). The median scores for the fall and spring administrations were at the 77th and 82nd percentiles respectively. The distributions of scores at the beginning and end of kindergarten, by stanines, are shown in Table 3. Note that if children's gains match those expected based on maturation alone their scores would remain the same.

Table 3
Distribution of PPVT-R Scores at the Beginning and End of Kindergarten, Combined Cohorts

	Percentile	Stanine								
	Rank of									
Test	Mean	1	2	3	4	5	6	7	8	9
Fall	75			4	9	19	16	22	20	10
Spring	84		1	0	2	11	23	26	19	19

Note. Percentages do not sum to 100 due to rounding.
A one-sample *t* test for change in standard score equivalents produced the following results: *t* (171) = 6.7, p < .0001 (two-tailed), Cohen's *d* = 0.5.

Grades 1 Through 4 Achievement Test Scores

Test results in spring of Grade 1 were very high: the mean Reading score was at the 95th percentile, and the mean Mathematics score was at the 98th percentile. Fifty-two percent of the students had scores in stanine 9 in Reading, and 83% had Mathematics scores in stanine 9. Longitudinal analysis of these test scores reveals that the children's percentile ranks on standardized achievement tests declined from Grade 1 through Grade 4. The mean Reading score in spring of Grade 4 was at the 84th percentile. The mean Mathematics score in Grade 4 was at the 85th percentile. This pattern was observed with all cohorts tested beginning in Grade 1, with both the CTP III and the SAT-9, and on the reading and mathematics portions of both test batteries. The longitudinal patterns, aggregated across the three cohorts, are shown in Figures 1 and 2. The distributions of scores are shown in Tables 4 and 5.

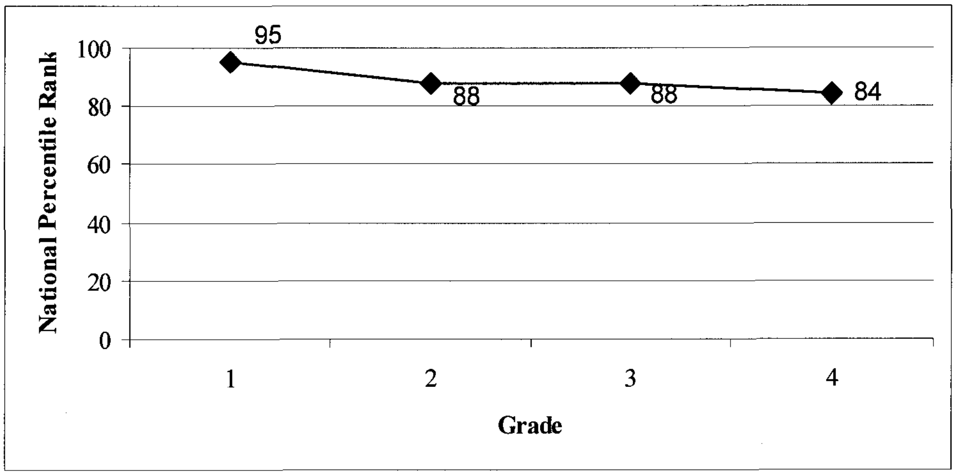

Figure 1. Longitudinal reading related standardized test scores, combined cohorts.

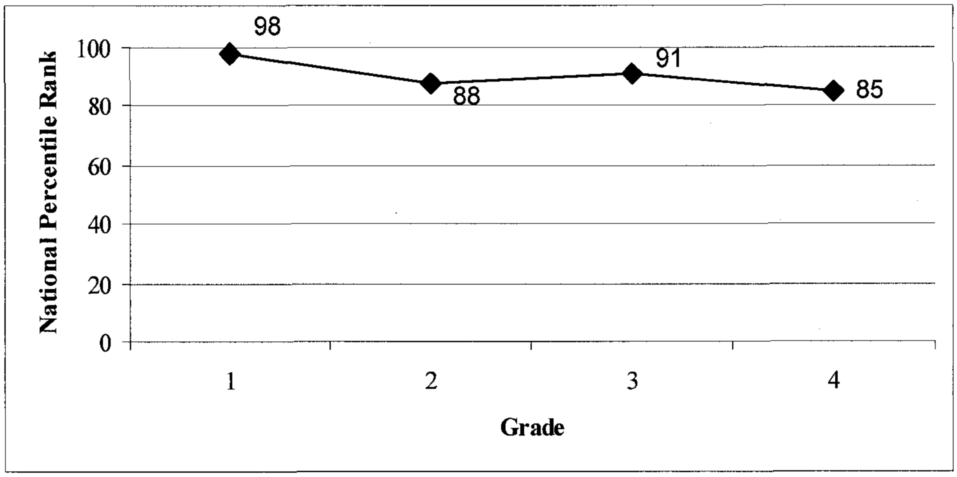

Figure 2. Longitudinal mathematics standardized test scores, combined cohorts.

Table 4
Distribution of Reading Scores by Grade, Combined Cohorts

Test	Percentile Rank of Mean	Stanine								
		1	2	3	4	5	6	7	8	9
Grade 1	95				2	4	5	17	20	52
Grade 2	88			1	1	6	19	20	25	28
Grade 3	88				1	4	13	33	30	19
Grade 4	84				1	11	23	25	24	16

Table 5
Distribution of Mathematics Scores by Grade, Combined Cohorts

Test	Percentile Rank of Mean	Stanine								
		1	2	3	4	5	6	7	8	9
Grade 1	98				1	0	5	5	6	83
Grade 2	88					9	19	23	24	26
Grade 3	91				1	8	13	21	20	36
Grade 4	85				4	8	18	28	30	12

In an early effort to understand these data, the results and preliminary analyses were shared with a number of recognized leaders in the fields of educational evaluation and measurement. One of these responded, “Obviously the test doesn’t capture the range of achievement at [the school].” (T. D. Cook, personal communication, April 1997).

Research Questions and Approach

Research Hypothesis and Questions

Are the standardized tests used at this school (and other tests with similar characteristics) valid indicators of the achievement and growth of high-achieving students at these grade levels? It is my contention that the tests used at this school are *not* valid for the purpose of monitoring student growth and program impact with the typically high-achieving students enrolled at this school, neither are they appropriate for use in other schools with comparable student populations.

This contention is supported by two measurement experts who reviewed the data. They both expressed the opinion that the phenomenon typically labeled "regression toward (or to) the mean" or the "regression effect" is operating here (personal communication, H. D. Wainer, April 1997 and W. J. Popham, May 1999).

To understand and evaluate Popham's and Wainer's responses, we need to understand the nature of regression toward the mean and under what conditions it is a concern. We need to know if those conditions are present in this and similar circumstances. And, we need to identify any other systematic factors operating that are relevant to the interpretation of the data.

I hypothesize that the change, or regression, observed in the test scores can be largely explained by four factors: (a) the traditionally defined regression toward the mean that occurs as a result of measurement error and imperfect correlations between measures; (b) related to the above but distinct is the use of grade-level norms during a period of rapid cognitive growth, which creates a ceiling for these students; (c) the extent to which changes in the content of the tests across grade levels affects test scores longitudinally; and (d) the documented instability of young children's performance relative to the norm groups. A simplified representation of this model is provided in Figure 3.

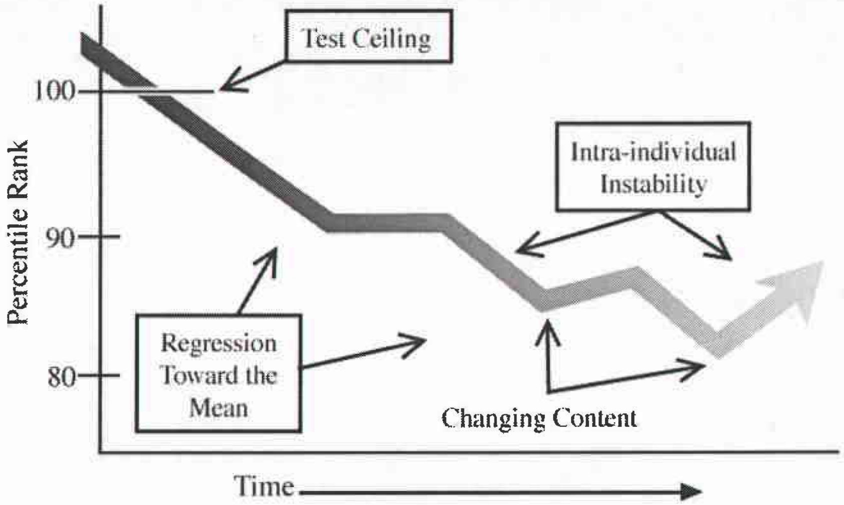

Figure 3. Model of factors that influence the normative stability of test scores over time.

To test this hypothesis, I explore the factors identified above through the following research questions:

1. To what extent can the change in test scores reasonably be explained by regression toward the mean?
2. Do test ceilings for children who demonstrate markedly high levels of achievement at young ages lead to inappropriate expectations about their later performance?
3. How does the knowledge and skills assessed in these test batteries change across grade levels, and, in turn, do the changes in test content affect students' normative ranks?
4. Is it reasonable to expect children demonstrating high levels of achievement relative to the norm group at young ages to maintain that relative position over time?

Approach

To address these questions, I developed a synthesis of theory, best practices, and research findings from the fields of evaluation and research design, psychometrics, and developmental psychology. I used the convergence of these disparate sources of knowledge to explore the research questions in a way that is not currently available to evaluation practitioners or others responsible for the selection of educational assessments and the use of test results for educational program monitoring, evaluation, and improvement.

By design, I used the materials and statistical analyses that, I believe, should be part of the "tool kit" of any professional with responsibility for student assessment and program evaluation at or above the school district level. All reference information came from the technical reports and norms books routinely published by the commercial test developers. The data analyses were conducted using the information that is standard in the electronic files of test results from the scoring services (with the exception of the raw scores for the CTP III which were added to the files after they were received from the scoring service). The statistical methods used are primarily descriptive statistics, with just a few forays into correlations and ANOVA. These statistics can all be calculated with nothing more sophisticated than Microsoft Excel, although using these Excel functions can be more cumbersome than using statistical software programs. Thus, my approach is widely replicable in daily practice across the country.

In addition to its applied value, this work makes contributions to the continued development of the theory of regression toward the mean. The careful analysis of data from this school provides a relatively rare opportunity to test and elaborate upon the theory of regression toward the mean with a high-achieving population.

In Chapter 2, I provide a review of the literature relevant to each of my four questions.

CHAPTER 2. REVIEW OF LITERATURE

A Question of Validity

The fundamental issue in this study is one of validity–making valid inferences about the effectiveness of an educational program. Validity is not an intrinsic characteristic of tests or other psychometric tools. Validity is situational; it is present when we achieve congruence between characteristics of the measurement tools, methods, the context we are studying, and the questions we are asking. Validity is defined in the *Standards for Educational and Psychological Testing* as follows:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed users of tests....The process of validation involves accurately accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores that are evaluated, not the test itself....Identifying the propositions implied by a proposed test interpretation can be facilitated by considering rival hypotheses that may challenge the proposed interpretation."
> (American Educational Research Association et al., 1999, pp. 9–10)

Developers of educational assessments have an ethical responsibility to accurately describe the contexts and purposes for which their tests are valid (American Educational Research Association et al., 1999, pp. 67–70). Test users have a responsibility to ensure that the tests they select are appropriate to the context and questions at hand and that all inferences drawn from the results are valid. Standard 11.15 reads as follows:

> Test users should be alert to potential misinterpretations of test scores and to possible unintended consequences of test use; users should take steps to minimize or avoid foreseeable misinterpretations and unintended negative consequences.
>
> *Comment:* Well-meaning, but unsophisticated, audiences may adopt simplistic interpretations of test results or may attribute high or low scores or averages to a single causal factor. Experienced test users can sometimes anticipate such misinterpretations and should try to prevent them. Obviously, not every unintended consequence can be anticipated. What is required is a reasonable

> effort to prevent negative consequences and to encourage sound interpretation. (American Educational Research Association et al., 1999, pp. 116–117).

Standard 11.18 expands upon this in ways that are particularly relevant to the context of the current study. "When test results are released to the public or policymakers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretation of the data." (American Educational Research Association et al., 1999, p. 117).

Threats to Validity

In any research or evaluation project, the researcher needs to identify and consider potential threats to the validity of the conclusions or inferences that result from the study. There are many threats, and the potential significance of each varies from one study to another. The threats are typically grouped into two broad classes: (a) threats to internal validity which includes factors that relate to the integrity of the study itself and the ability of the researcher to draw conclusions that the treatments did or did not have an effect in the particular instance, and (b) threats to external validity which includes factors related to the generalizability of the findings beyond the particular circumstances of the study (Campbell & Stanley, 1963).

Internal threats "could plausibly have caused an observed relationship *even if the treatment had never taken place*" (Shadish, Cook, & Leviton, 1991, p. 126). Could the observed changes in test scores have occurred despite the quality of the school?

The following sections explore four potential threats to the validity of the most facile interpretation of the regression in the longitudinal test scores; that is, that the regression in the test scores was proof that the school program had a negative effect on the children's cognitive development. The threats addressed here are (a) regression to the mean ("statistical regression" in Campbell & Stanley, 1963), (b) the test ceiling, (c) the changing nature of what was tested at each grade level, and (d) the assumptions of stability in the normative achievement of young children.

Regression Toward the Mean

Definitions

"Regression toward the mean" or the "regression effect" is the label given to the tendency of persons or groups with extreme scores on a measure to score closer to the mean value when measured on another occasion on the same or a related construct. It is a pervasive phenomenon, observed in education, health, business, and many other fields. Regression toward the mean is easily one of the most ubiquitous and overlooked threats to the validity of educational evaluation research. It has long been recognized in textbooks on research design and applied educational and psychological statistics (Campbell & Stanley, 1963; Glass & Hopkins, 1984; Hartman & George, 1999; Huck, Cormier, & Bounds, 1974; Pedhazur & Schmelkin, 1991). It is a not an uncommon term in popular culture. For example, a Google search for the phrase "regression to the mean" returned over 32,500 hits while "regression toward the mean" and "regression effect" produced thousands more. However, it is a complex phenomenon and often misunderstood.

The first recorded attempt to identify and explain the phenomenon of regression toward the mean appeared in the work of Francis Galton in 1886. He was trying to explain data showing that children of tall parents were, on average, not as tall as their parents and, at the same time, the children of short parents were, on average, not as short as their parents, even though the average values for each generation remained the same (Clemons, 2000; Glass & Hopkins, 1984).

In *Experimental and Quasi-Experimental Design for Research*, Campbell and Stanley (1963) discuss regression toward the mean using the term "statistical regression." They explain that

> If in a remediation experiment, students are picked for a special experimental treatment because they do particularly poorly on an achievement test…, then on a subsequent testing using a parallel form or repeating the same test… this group will almost surely average higher…. This dependable result is not due to any genuine effect of *X* [the treatment], any test-retest practice effect, etc. It is rather a tautological aspect of the imperfect correlation between [the measures used]. (p. 10)

They go on to provide a number of examples, including one that does not involve pretesting and posttesting with the same or comparable forms of a test.

> The principal who observes that his highest-IQ students tend to have less than the highest achievement-test scores (though quite high) and that his lowest IQ students are usually not right at the bottom of the achievement-test heap (though quite low) would be guilty of the regression fallacy if he declared that his school is understimulating the brightest pupils and overworking the dullest. (p. 11)

Regression to the mean occurs when either of two conditions exists: measurement error or the lack of perfect correlation between variables. The magnitude of regression toward the mean increases as the observed values move away from the mean; the more extreme the value, the greater the potential regression effect. Therefore, threats from regression toward the mean to the validity of our inferences are particularly critical when people or groups are chosen for study because of their extreme standing on some variable, as in the case of educational programs serving students with below- or above-average achievement (Campbell & Stanley, 1963; Furby, 1973; Pedhazur & Schmelkin, 1991).

In "Interpreting Regression toward the Mean in Developmental Research," Furby (1973) provides one of the more comprehensive explanations of this phenomenon. She explains how measurement error contributes to regression toward the mean. Positive error occurs when examinees provide correct responses to items where they really don't know the answer. This could be the result of informed guessing or good luck. Negative error occurs when examinees provide incorrect responses to items where they do know the correct answer. This could be the result of poor attention or other bad luck. Typically, positive and negative errors balance each other out, but this is not always the case. At the extremes positive errors are not likely to be offset with negative errors and vice versa. Just as some gamblers leave Las Vegas as big winners and some leave with large losses, some of those students with high scores will have large positive error, while some of those with low scores will have large negative error.

By definition, measurement error is unrelated to the construct being measured. Thus, it is highly unlikely that students with large positive error in one score will also have large positive error when retested or tested with another tool. Their second scores will tend to contain a balance of positive and negative errors and will regress toward the mean (i.e., on average, those students with previously high scores will have lower scores on the other occasion). The same will occur for students with very low scores on one measurement (i.e., on average, their scores on the other occasion will be higher).

When the population or sample represents the full range of achievement, these patterns of errors would be occurring at all measurement occasions and so would not result in changes to the group mean. However, when groups are selected because of their extreme standing (above- or below-average) and measured on a second occasion or with a second tool, the large positive (or negative) error obtained in the first measurement is not likely to recur. In these situations the scores of the group will move closer to the mean of the population from which they were drawn.

The second element in regression toward the mean is the lack of perfect correlation between the measurement tools. Even when our variables can be measured without error, an extreme score obtained with one tool is likely to have been the result of a rare combination of factors. It is highly unlikely that this student or group of students will have the necessary rare combination of factors to produce as extreme a score with the second tool and, therefore, their second scores will regress toward the mean (Furby, 1973, pp. 174–176).

Nesselroade, Stigler and Baltes (1980) further develop the concept of regression toward the mean. They build upon the work of Furby and others and apply the concept to longer series of data collection points. They begin with what they refer to as the "two-occasion model" used by Furby and others and calculate expected score at Time 2 as the product of the original score and the correlation coefficient between the two measures.

Nesselroade et al. (1980) suggest that to adequately understand the impact of regression toward the mean on change, it is necessary to move beyond the two-occasion models previously published to the study of multiple-occasion change. For example, some might be inclined to assume that to calculate the expected score at Time 3 one would compute the product of the Time 2 score and the correlation between the measures used at Time 2 and Time 3. Instead, the authors assert that the expected score at each time is a function of the correlation between the *first* and each subsequent measurement.

Incorrect expected sequence:
x
corr $(X_1, X_2)*x$
corr $(X_1, X_2)*$corr $(X_1, X_3)*x$
corr $(X_1, X_2)*$corr $(X_1, X_3)*$corr $(X_1, X_4)*x$, and so forth.

Correct expected sequence:
x
corr $(X_1, X_2)*x$
corr $(X_1, X_3)*x$
corr $(X_1, X_4)*x$, and so forth.

(Nesselroade et al., 1980, p. 627)

As a result of applying the correct sequence in the situation in which the correlation between the original measurement and subsequent measurements decreases over time, the scores would continue to regress toward the mean. In the situation in which the correlation between the original measurement and subsequent measurements increases over time, the scores will begin to move away from (egress from) the mean after the initial regression.

Multiple-occasion change models must also incorporate consideration of the nature of the processes assumed to generate the change. For example, if the true score is *not* static and tends to evolve over time as a result of treatment or maturation, the effect of treatment or maturation at each measurement point would contribute to the expected score.

Applications

As demonstrated above, given the lack of error-free assessment tools or methods or perfect correlations in educational measurement, regression toward the mean is inevitable. It is a serious and pervasive threat to the validity of inferences we make about individuals and the effects of programs.

The ubiquity of the "regression fallacy" might be inferred from the continued trickle of publications on this topic from a methodological perspective where the authors warn us against ignoring the regression effect. A sample of these includes Beath and Dobson (1999), Brown and Jackson (1992), Cole (1998), Furby (1973), Hills (1993), Kruger, Savitsky and Gilovich (1999), Mee and Chua (1991), Nesselroade et al. (1980), and Wainer (1999).

However, recognition of this phenomenon appears infrequently in the applied educational research and evaluation literature. For example, a search of the contents of the former ERIC Assessment and Evaluation archives (www.ericae.net, closed by the U.S. Department of Education in December, 2003) for the terms "regression toward the mean," "regression to the mean," "regression effect," and "statistical regression" revealed two articles where this phenomenon was considered in the interpretation of change scores for an educational program. A search of all articles in the online journal Practical Assessment, Research & Evaluation (PARE, http://pareonline.net) for these terms resulted in no returns that involved use of the concept in educational research or evaluation. A search of the Ingenta database (www.ingenta.com), which indexes more than 18,000 journals, resulted in 65 relevant hits, 31 of which were in health-related journals, 11 in business and economics, 10 in statistics, 6 in psychology, 3 in education, and 4 in other assorted content areas. All of these searches included all years of indexed articles and other publications.

That this threat is rarely considered in evaluation studies and research is evident in the fact that so few reports of research and evaluation in education settings involving the measurement of change in extreme populations make reference to the phenomenon. Although there are undoubtedly articles and reports I overlooked, my search for this application covered a broad swath of the educational evaluation publications. Yet, in total, I discovered only a handful of articles in which regression toward the mean was applied in the evaluation of an educational program.

Ingebo and Doherty (March 16, 1982) made a bold statement when they undertook to explore the "magnitude of untruth that is represented in evaluations of Title I programs" via their examination of the effects regression toward the mean. Their methodology is unusual. They used *T*-scores created independently for fall and spring based on the distribution of scores at both those points at each grade level. The students were enrolled in Grades 3 through 8. Ingebo and Doherty assumed that any change in *T*-scores between fall and spring testing was regression toward the mean (p. 3). Their process assumes the program should have resulted in students maintaining their position relative to their peers (i.e., having the same *T*-score at the beginning and end of the school year) and that any other change was considered evidence of the regression effect.

Tallmadge (April, 1988) reported on a system of selecting students for Title I services. The initial selection includes achievement testing. At the beginning of the services the students were tested with a different tool to reduce regression to the mean related to measurement error. The program with which Tallmadge was involved also applied a statistical correction to posttest scores based on the correlation between the pretest and posttest to address the component of regression toward the mean that is a function of the less than perfect correlation between the tests.

Tallmadge (April, 1988) cited previous literature (Linn, 1980; Roberts, 1980; and Tallmadge, 1982) that suggested that the magnitude of regression toward the mean for Title I populations is 2 *NCEs* in Grades 1 and 2 and 1 *NCE* above Grade 2. In his study, using carefully simulated data but without the floor or ceiling effects often found in actual data from extreme populations, he found a regression effect of 5 *NCEs* between the selection test and the pretest and an additional 1 *NCE* between the pretest and posttest (p. 15).

Note that if Tallmadge's scenario had included test floors the magnitude of the regression effect would have been larger. As Mills and Jackson (1990) noted in their study of precocious readers, when some scores are at or near the test ceiling, the measures are less reliable and this contributes to increased measurement error and reduces the correlation between the measures.

Furlong and Feldman (1992) suggested that regression toward the mean can explain the discretionary decisions by multi-disciplinary teams to place students referred for special education services in a level of service that is less intensive than what would be indicated based on a review of the ability-achievement score discrepancies alone. Furlong and Feldman used estimated scores with a correction for regression toward the mean to recalculate the ability-achievement discrepancies. They found that between 25% and 50% of the discretionary decisions were consistent with the "corrected" scores.

Venezky (1994) attempted to control for regression toward the mean between pre- and posttest in scores of participants in an adult literacy program by correcting pretest scores for guessing on a multiple choice test.

Miller-Whitehead (1999) explored the possibility that schools with high test scores would not show continuous improvement from year to year over a period of several years due to regression toward the mean. Her analyses of the scores of the five highest and lowest achieving school systems in Tennessee showed an overall pattern of increasing test scores among both groups of schools. From these results, she concluded that "even top systems can and do continue to improve in achievement from year to year…" (p. 5). However, she noted that "for identified subgroups of students, classes, or schools which have indeed 'topped out' this [regression toward the mean] would, of course, continue to be an issue." (p. 4).

More recently, Camilli and Bulkley (2001) critiqued an evaluation of the Florida A-Plus program as flawed because the evaluator failed to adequately control for regression toward the mean. They followed a model suggested by Glass and Hopkins

(1996) and employed an analysis of residual scores in an attempt to control for regression effects.

Although not identified as an application of regression toward the mean, the work by Berliner and others (for example, Amrein-Beardsley & Berliner, August 4, 2003) examining the effects of high-stakes testing on scores obtained from non-high stakes examinations of the same content area could be construed as an example of regression toward the mean. These authors contend that while the results on the local high-stakes tests have improved, results on other tests (e.g., NAEP) have not improved, have not improved at the same rate, or can be explained by factors other than the introduction of high-stakes testing. *If* the improvements observed on the high-stakes tests are reflective of a special circumstance (e.g., the convergence of the test and the local curriculum), this is an example of the "rare combination of factors" that does not generalize to another test of the same construct—one of the causes of statistical regression according to Furby (1973).

The literature on the education of the gifted and talented also makes references to the importance of regression toward the mean in evaluating programs for that population. These articles use regression toward the mean to explain the tendency of norm-referenced test scores for students in these programs to decline over time. They strongly suggest that growth is occurring even in the face of declining test scores (e.g., Callahan, 1992; Kelly & Peverly, 1992).

An Additional Application: Comparison of Test Scores from the Admissions Process and at Entry to the School.

Another example of regression toward the mean comes from the school that provides the data for this study. Applicants for admission to Grade 7 take the CTP in late Fall of Grade 6. Those who are admitted are tested again at the beginning of the school year. The conditions under which we would expect to see the regression effect are clearly present. First, there is selection based, in part, on test scores. This intrinsically favors students with larger positive error. Second, there is an imperfect correlation between the two tests. Third, there are special circumstances which increase positive error: when applicants took the admissions test, they were engaged in what is inarguably a high-stakes testing situation, and anecdotal evidence suggests that many of the applicants engage in extensive test preparation in anticipation of the admissions test. The distributions of scores at each occasion are displayed in Figures 4 and 5. The regression effect is evident in the figures.

Using the formula provided by Nesselroade et al. (1980), I calculated the expected scores after adjusting for regression toward the mean stemming from the lack of perfect correlation. Again, the formula is expected score = corr $(X_1, X_2)*x$. I used two approaches with this task to create upper and lower bounds on the regression toward the mean estimates.

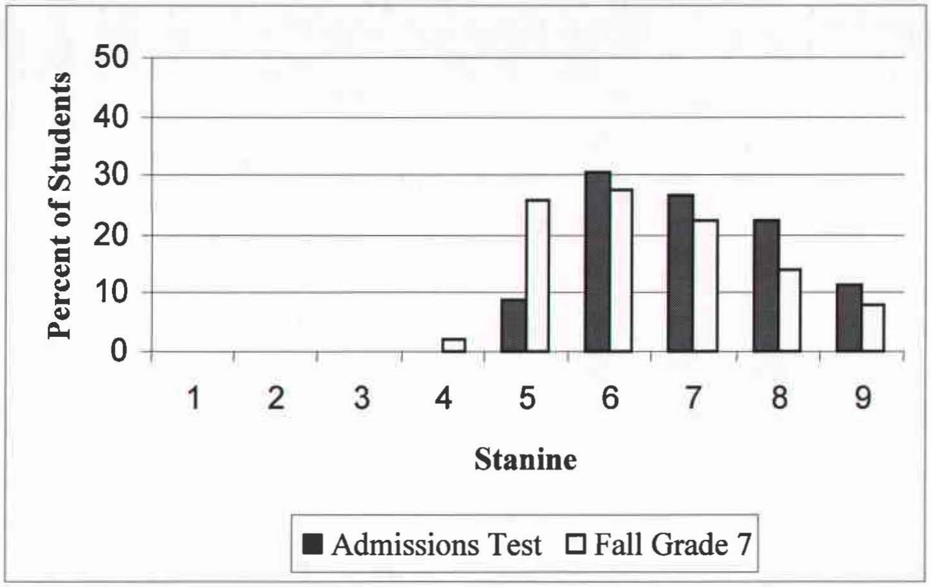

Figure 4. Comparison of distributions of reading test scores for school admittees during the admissions process and at entry to Grade 7.

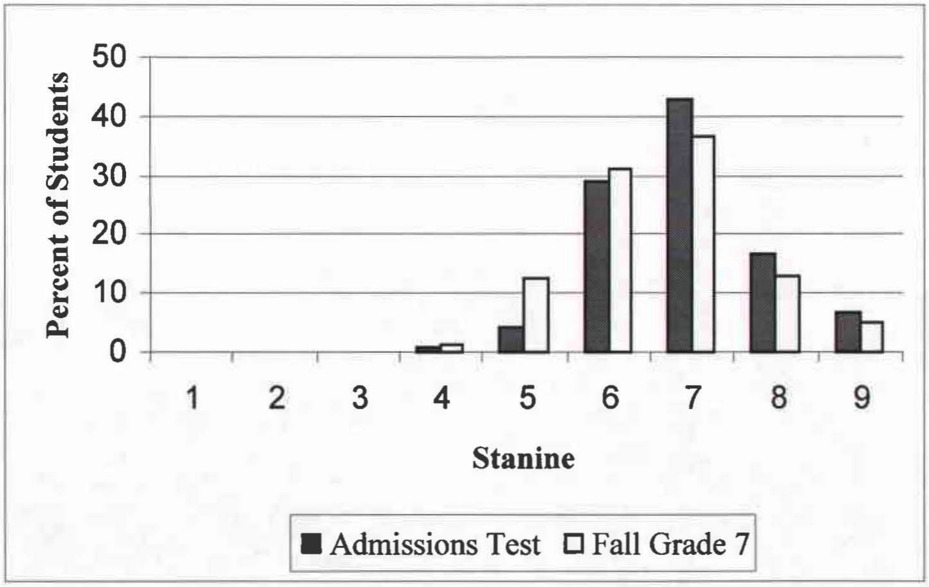

Figure 5. Comparison of distributions of mathematics test scores for school admittees during the admissions process and at entry to Grade 7.

First, since the correlation between levels of the CTP III are not provided in the technical manual (Educational Testing Service, 1995), I used the highest value for the K–R 20 for each pair of tests. This value was consistently higher than the available indicator of reliability—the alternate forms correlation. Because the K–R 20 is the larger of the two values, it will produce the smallest estimated regression effect. The maximum value of K–R 20 is .91 for Reading Comprehension and .90 for Mathematics (Educational Testing Service, 1995, pp. 59-60).[3] The results of these calculations are shown in Table 6.

Table 6
Expected and Actual Scores From Fall Grade 7: Based on K–R 20

Domain	Mean Admissions Test Score		*	K–R 20	=	Expected Fall Test Score		Actual Fall Score
	NCE	RS				RS	*NCE*	
Reading	70	34	*	.91	=	31	64	64
Mathematics	69	65	*	.90	=	59	64	66

Note. RS = raw score. *NCEs* are converted to raw scores for computation because K–R 20 values are computed using raw scores, and the transformation from raw scores to *NCEs* is not linear.

Second, I used the correlation between the tests for the admittees (r_{12}). For this group of students, the actual correlations between the admissions and Grade 7 test scores are .63 for Reading Comprehension and .72 for Mathematics. However, the range of

[3] The information on test reliability for mathematics was presented separately for each of the two parts of the test. I applied the formula provided by Nunnally (1970, p. 557) to estimate the reliability coefficient for a change in test length to estimate the reliability coefficient for the total score for the mathematics test: $r_{kk} = \frac{kr_{11}}{1+(k-1)r_{11}}$
(k = factor by which the number of items is changed).

scores is severely truncated by the admissions selection process. When the formula suggested by Nunnally to correct for attenuation[4] is applied to the observed correlations, the adjusted correlation coefficients are .70 and .79, respectively. I used these coefficients in the same calculations reported above to produce a more liberal estimate of the expected fall test scores (see Table 7).

Table 7
Expected and Actual Scores From Fall Grade 7: Based on Sample Correlations, Corrected for Attenuation

Domain	Mean Admissions Test Score		*	Adjusted r_{12}	=	Expected Fall Test Score		Actual Fall Score
	NCE	RS				RS	*NCE*	
Reading	70	34	*	.70	=	24	54	64
Mathematics	69	65	*	.79	=	51	59	66

Note. RS = raw score. *NCEs* are converted to raw scores because reliability coefficients are computed for raw scores and the transformation from raw scores to *NCEs* is not linear.

The results reported in Table 6 indicate that the actual Reading scores in fall of Grade 7 exactly match the prediction based on the test reliability coefficients. The actual Mathematics scores did not regress quite as far as the calculations suggested they might. The actual Reading and Mathematics scores are substantially higher than expected when using the correlation between the admissions and fall scores, even after these correlations are corrected for attenuation.

[4] $r' = \frac{r_{12}}{\sqrt{r_{11}}\sqrt{r_{22}}}$ (Nunnally, 1970, p. 117)

Test Ceilings

Definitions

When, after reviewing the pattern of test scores at this school, T. D. Cook stated that "Obviously the test doesn't capture the range of achievement at [the school]" (personal communication, April 1997), he was referring to the impact of the test ceiling. In "Ten Psychometric Reasons Why Similar Tests Produce Dissimilar Results," Bracken (1988) identified ceiling effects as one of the causes of disparate results from related measures.

A test has a ceiling for a given population when it cannot accurately measure achievement at the upper levels of the examinees. Test ceilings occur when an instrument does not have a sufficient number of difficult items to distinguish between a very able child and a child who is average or high average in the skill areas assessed.

A test floor is, as one might expect, the same construct applied to the lowest level of achievement at which the test is accurate. The "effective range" of a test is the range of achievement (or ability) over which it returns accurate information.

The ceiling reduces the reliability and validity of the test scores. The truncation of scores increases the probability that some portion of the observed regression toward the mean is attributable to the reduced reliability of the instrument (Mills & Jackson, 1990). In addition to creating a situation where the test cannot adequately discriminate between examinees, scores at or near the test ceilings represent the most extreme scores. Often, scores derived from the raw scores, such as percentile ranks or scale scores, are based on mathematical extrapolation from observed data and thus tend to be less accurate than more moderate scores (Bracken, 1988).

Tests ceilings or floors also affect the vertical equating, or linking, of multiple levels. Yen (1983) demonstrated that tests with ceiling, or floors, for groups of examinees will produce systematic error in percentile ranks and that percentile ranks across equated tests of unequal difficulty are not equivalent, even for scores at or near the 50th percentile.

In program evaluation, the ceiling effect, when present, is a potent source of invalid inferences about the effectiveness of programs for the gifted. When the content of the curriculum offered to the students is accelerated, the range of test items presented to the students from a traditional on-grade standardized assessment does not assess the goals and objectives of the program (Callahan, 1992). It should be noted that the students do not need to be in the range of achievement considered gifted for the ceiling effect to be present. They merely need to have tested achievement at a level that exceeds the range over which the test is effective.

To determine if a test ceiling (or floor) is likely to be of concern, Bracken (1990) suggests that test users examine the relationship between the raw scores and standard scores. If the standard score equivalent to the highest (or lowest) raw score is less than two standard deviations from the mean the test may have a ceiling or floor effect.

A related concept is item gradient, or how much change in standard scores or percentile occurs as a result of minor changes in raw scores. Tests with steep item gradients are less sensitive to differences in children's achievement or ability than those with more gradual gradients. Bracken (1988) observes that "It is frequently the case that tests with poor ceilings and floors also evidence steep item gradients." (p. 159).

In "Maximizing Content Relevant Assessment," Bracken (1990) suggests that when an increase or decrease of a single raw score alters the corresponding standard score by more than one third of a standard deviation the gradient at that point of the score distribution is too steep. Tests with item gradients of this magnitude or higher cannot accurately assess individual differences in achievement or ability.

Applications

Bracken (1988) provides the following example of a child whose score is at the ceiling of one test but not another and the potential impact of that ceiling on the inferences drawn from the test scores.

> As an example, assume that two different M-team [multi-disciplinary team] members have screened a 7 1/2 year-old girl for gifted placement, using two different measures of receptive vocabulary. The first examiner has used the Bracken Basic Concepts Scale (BBCS; Bracken, 1984), which is appropriate for children in the age ranges of 2–6 through 7–11. The second examiner has assessed the child with the Peabody Picture Vocabulary Test-Revised (PPVT-R; Dunn & Dunn, 1981). The PPVT-R is appropriate for individuals from 2 1/2 years through 40 years. When the two diagnosticians confer they find disparate results – results that can be explained by examining the norm tables of the two instruments.
>
> At 7 1/2 years, a child who answers all 258 items correctly on the BBCS earns a total test standard score of 136. However, should that child fail as few as five items, the score drops precipitously to 117 (no longer in the gifted range). The same child can earn a total test PPVT-R score of 160 (nearly two standard deviations higher than the highest attainable BBCS score at this age level) with all items answered correctly; and failure of only five PPVT-R items reduces the score minimally to 153 (still well into the gifted range) (p. 158).

Because the achievement of the child in this example exceeds the effective range of the BBCS, her true achievement may be underestimated, and a small change in scores produces an apparent but erroneous large change in her rank relative to the norm group.

Stanley (1990) provides an example of a college admissions test with a ceiling. In the 1987–88 school year, approximately 45,000 (four percent) of more than a million college-bound high school seniors earned the top score on the Test of Standard Written English. These students were rendered indistinguishable from each other due to the ceiling inherent in the test.

The National Educational Longitudinal Survey of 1988 (NELS:88) was designed to assess student growth from Grade 8 through Grade 12. The authors describe the assessment dilemma inherent in developing appropriate tools for that project:

> the potentially large variation in student growth trajectories argues for a longitudinal "tailored testing" approach to assessment.... If the **same** test, in say, Mathematics or Reading Comprehension were administered to the same student at the eighth, tenth, and twelfth grades, the potential for observing "floor effects"' at grade eight and "ceiling effects"' at grade twelve is greatly increased. (Rock, Pollack, & Quinn, 1995, p. 5).

The challenge for Rock and his colleagues was to accurately assess the wide range of achievement they expected to observe while working within constraints on the length of the tests and without the opportunity to use computerized adaptive testing. Their solution for the assessment of reading comprehension and mathematics was to create three levels of tests for use at the tenth and twelfth grades which were tailored to the achievement level demonstrated by individual students on the prior test. For example, students with relatively high scores on the Grade 8 test (i.e., those scoring in the top 25 percent) were given a more difficult test when they were retested in Grade 10. Those students who scored in the bottom 25 percent were given an easier form. Those scoring in the middle range received a test form of average difficulty. The same process was repeated when the tenth graders were retested in Grade 12.

To determine who would get which test in tenth and twelfth grades, it was important that the eighth grade test provide accurate results across the full range of achievement. This was accomplished by building a test with a rectangular distribution of item difficulty (as contrasted with the typical test, which tends to follow a normal distribution of item difficulties with most items of moderate difficulty).

Changes in the Nature and Content of Assessments of Reading and Mathematics in the Early Grades and Vertical Equating

Definitions

Changing Content. It is clear that the nature of reading and mathematics changes dramatically during the early school years. In the domain of reading, children move through letter recognition, letter-sound correspondence, word recognition, decoding, and eventually reading comprehension. In the domain of mathematics, children move from quantitative concepts such as more and less, to rote counting, to numerical representations of numbers, to manipulation of number in arithmetic operations, and beyond.

As the skills and knowledge related to reading and mathematics evolve, so does the content of achievement tests of these domains. Two tests, or levels of the same test, may sample the universe of reading or mathematics but the tests may not be parallel in their sampling of the content domain.

In "Questions and Answers in the Measurement of Change," Willett (1988) notes that if the variable we are measuring changes over time (e.g., through instruction) the construct validity of the measurement tool changes and results collected at one point may not be equatable to those collected on other occasions.

Bracken (1988) includes content differences across tests as one of the factors that can result in apparently discrepant results from similar tests. He notes that

> Some tests include a higher incidence of items that assess the examinee's knowledge of concepts (e.g., half, dozen, twice, more than), whereas others assess whether the examinee's ability to rote and place count, and other tests immediately assess whether the examinee can add, subtract, multiply and divide. (p. 163)

While Bracken is describing different tests of the same construct administered concurrently, he could just as well be describing changes in tested content across levels of standardized achievement test batteries.

The major commercially published standardized achievement tests all have previous versions in existence, and, for some tests, more than one version may be in use at any one time. They have multiple forms of the same test for each version (the SAT-9 has forms S and T). They also have multiple levels for students in different grades (the SAT-9 has 13 levels) (Harcourt Brace Educational Measurement, 1997c, p. 35).

Scores from these multi-level standardized achievement tests are often used in longitudinal studies that examine changes over time for one or more cohorts of students, as in the case of this study. Even more common, is their use in cross-sectional studies that compare the scores of different groups of students (e.g., students in Grades 3, 5, and 8) collected at one point in time or scores at one grade level across multiple cohorts (e.g., scores for students in Grade 6 in 1999, 2000, and 2001). These uses require that the scores be equivalent in some way. The need for equivalence is addressed through the creation of developmental scales that span the multiple levels of the test via the process of vertical equating (Peterson, Kolen, & Hoover, 1989).

Equating. The equating process for standardized achievement tests typically operates in two directions. Horizontal equating is used to ensure that multiple forms of the same test, if they exist, yield equivalent scores. It is also used to ensure that results from old and new versions of a test are comparable.

Vertical equating (or linking) to create developmental scales makes it possible to study change across multiple levels of one test battery. This resulting continuous series of scores represents increasing levels of achievement or ability (Linn, 1993; Peterson et al., 1989). It is vertical equating that is of most concern in this study.

The first step of vertical equating is the rank-ordering of examinees on a single developmental continuum. There are four commonly used methods to accomplish this.

1. Single-Group Design. The two tests to be linked are given to the same group of examinees. This is a simple design but it may be impractical to implement because testing time will be long. Moreover, practice and fatigue effects (if the two tests are administered one after the other) may have an effect on parameter estimation and, hence, on the linking results.

2. Equivalent-Groups Design. The two tests to be linked are given to two equivalent but not identical groups of examinees, chosen randomly. This design is more practical and avoids practice and fatigue effects.
3. Anchor-Test Design. The tests to be linked are given to two different groups of examinees. Each test has a set of common items that may be internal or external to the tests. This design is feasible and frequently used, and, if the anchor items are chosen properly (see, for example, Klein & Jarjoura, 1985), it avoids the problems in the single group or equivalent group designs.
4. Common-Person Design. The two tests to be linked are given to two groups of examinees, with a common group of examinees taking both tests. Because the testing will be lengthy for the common group, this design has the same drawbacks as the single-group design.

(Hambleton, Swaminathan, & Roger, 1991, p. 128)

Peterson et al. (1989) point out that each of these methods generally yields somewhat different outcomes. Their studies have indicated that the choice of linking method has a more significant impact on the resulting scale than the choice of statistical method used to assign actual scale scores. Mittman (1958) conducted an experiment to compare the results of the linking methods. He found that the anchor-test design results in scales that possess greater grade-to-grade overlap than those created using the equivalent-groups design. Greater grade-to-grade overlap decreases the sensitivity of the scale to differences in developmental achievement among examinees (cited in Peterson et al., 1989).

Peterson et al. (1989) prefer an alternative to the four most common methods of rank-ordering examinees. They use the scaling-test design to develop the grade equivalent scales for the Iowa Test of Basic Skills. First a scaling test is created. This test contains a representative sample of the content from all the test levels to be equated on the developmental scale. The scaling test is administered to a representative sample of examinees from each of the grade- or age-levels that the scale is intended to cover. This method yields the least amount of grade-to-grade overlap, resulting in a scale with the greatest sensitivity to differences among examinees and to changes in achievement over time.

Developmental scale scores should be used with caution and are not used in this study. The results of vertical equating are often not as elegant as the theory. These types of scores can produce apparently erratic patterns of growth that some consider educationally unreasonable (Burket, 1984; Hoover, 1984a, 1984b). Test ceilings and floors affect the scaling process and the reasonableness of results (Yen, 1983, 1986). In addition, it has been found repeatedly that test equating is not "population invariant;" that is, the characteristics of examinees (e.g., race, gender, and geographic region) affect the outcomes of the linking studies (Doran, 2004; Kolen, 2004; Yang, 2004). In short, if the results appear suspect, they may well be.

Applications

Changing Content. The standardized achievement test batteries are designed to reflect the changing nature of the domains they assess. For example, consider the subtests that contribute to the SAT-9 Total Reading scores (displayed in Table 8). The levels of the test designed for use from fall of Kindergarten and through fall of Grade 1 (SESAT I and 2) include the subtest Sounds and Letters – assessing students understanding of letter-sound correspondence. In the test levels designed for use between spring of Grade 1 and fall of Grade 3 (Primary 1 and 2) this subtest evolves to Word Study Skills, which assesses children's understanding of structural and phonetic analysis of words. This entire subtest is deleted in Primary 3 and beyond (Harcourt Brace Educational Measurement, 1997c). Presumably most students will have mastered these skills by that point. (The changes in test content across levels are presented in more detail in Tables 17 through 20.)

Table 8
SAT-9 Subtests Included in Total Reading for Selected Levels

SESAT 1	SESAT 2	Primary 1	Primary 2	Primary 3	Intermediate 1
Sounds and Letters (48 items)	Sounds and Letters (40 items)	Word Study Skills (36 items)	Word Study Skills (48 items)		
Word Reading (30 items)	Word Reading (40 items)	Word Reading (30 items)	Reading Vocabulary (30 items)	Reading Vocabulary (30 items)	Reading Vocabulary (30 items)
	Sentence Reading (30 items)	Reading Comprehension (40 items)	Reading Comprehension (40 items)	Reading Comprehension (54 items)	Reading Comprehension (54 items)

Note. Source: Harcourt Brace Educational Measurement (1997c, p. 35).

Tramontana, Hooper, and Seltzer (1988) conducted a meta-analysis of the prediction of early school achievement. They report that the pattern of effective predictors of reading varied according to the grade level in which outcomes were assessed and summarize their findings as follows:

> Some studies have suggested that developmental factors may be associated with a shift in the pattern of measures predicting reading over successive grade levels. Specifically, measures of cognitive or verbal abilities were not found to be predictive until about the second or third grade in some studies… possibly because reading at a beginning level depends more upon perceptual recognition abilities. (p. 131–132)

These conclusions were tested and affirmed in research conducted by Storch and Whitehurst (2002). They studied changes in the predictive power of a number of precursors of reading. They examined longitudinal data for children from Kindergarten through Grade 3. They found that different predictors come into play at different points in time, as the emphasis in reading shifted from mastering the representation of sounds and words in print to comprehension.

> As children begin formal schooling, the relationship between oral language and code related domains diminishes. In Grades 1 and 2 the relationship between oral language and reading ability is non-significant…. In Grade 2, though the tasks become more complex, reading is still heavily determined by the code-related skills children acquired by the end of their Kindergarten year. Furthermore, reading comprehension at this stage is highly correlated with word and nonword reading tasks, reinforcing the position that at least during the early stages of reading development, reading comprehension is primarily a function of word reading abilities…. By Grades 3 and 4, the pattern of influence changes significantly as reading accuracy and reading comprehension tasks can be reliably separated. (p. 942)

Mills and Jackson (1990) studied changes in the students' exceptionality as readers from Kindergarten to Grades 5 or 6. They found that, as a group, the precocious readers in their study remained "good readers but not necessarily extraordinary" and attribute this, in part, to changes in cognitive components of reading over the elementary school years (p. 417).

Vertical Equating. The SAT-9 developmental scores (the "scale score") were created using a modification of the equivalent-groups design for linking. Students completed the test level for their grade and one level lower. The order of administration was randomly assigned by classroom. The content domain for the two tests was calibrated as one long test. The actual scales were created using the Rasch model. This model places items on a common scale of difficulty and, by extension, places scores on a common scale of achievement or ability (Harcourt Brace Educational Measurement, 1997c).

The equivalent-groups design was also used to create the CTP III developmental scores (also named the "scale score"). The CTP III development differed from the SAT-9 in that students typically completed the test level for their grade, one level lower, *and* one level higher. Thurstone's absolute scaling method was used to derive the scale scores (Educational Testing Service, 1995).

The equivalent-groups design used in the development of scale scores for the SAT-9 and CTP III produces a result that is more sensitive to differences between the grade levels than the anchor-test design. However, it is still likely to underestimate the real differences (Mittman, 1958, cited in Peterson et al, 1989). After rank ordering was completed, the different methods used to assign scale scores for each battery have unique assumptions that are often problematic in practice.

The Rasch model used to develop the SAT-9 assumes that achievement is unidimensional. Clearly, this is not the case, as shown in Table 8 (see also Hoover, 1984a; Yen, 1985, 1986). The multidimensionality of reading achievement and the variation in the importance of different dimensions over the course of learning to read are discussed in detail by Storch and Whitehurst (2002).

Thurstone scaling assumes that within grade level achievement is normally distributed. According to Peterson et al. (1989), "this assumption of within-grade normality seems to be unwarranted in most areas of achievement. For example, in a subject matter area such as reading, the achievement of children beginning school is apt to be markedly skewed to the right, because most children cannot read at all, whereas a few can read very well." (p. 236).

Violations of assumptions underlying both the Rasch and Thurstone approaches to scaling are likely to contribute substantial error to the scale scores. In turn, this can create consternation, if not havoc, in the longitudinal study of growth in achievement, particularly when the students have reached or approached the test ceiling.

Instability of Growth in Early Achievement

Definitions

Fundamental to the interpretation of the observed trend in achievement scores are our expectations about what they should be. It is commonly believed that successful students and schools maintain or improve their position relative to the norm group over time. However, these assumptions are *not* supported by the research on the stability of cognitive growth, particularly during early childhood.

The paradoxical concept of stability of growth refers, in part, to the extent to which growth is predictable, that is, the extent to which individual differences are preserved over time. This phenomenon is typically represented from a quantitative perspective as the stability coefficient: the correlation between measures at two or more ages (Wohlwill, 1980).

The foundational research on stability of cognitive growth comes out of the field of developmental psychology. Much of the research on stability of cognitive growth was conducted between 1920 and about 1980. The bulk of the data used in this line of research came from a handful of major longitudinal projects conducted in the 1930s and 1940s. Prominent among these studies are the Fels Longitudinal Study, the California Growth Study, the Brush Foundation Study, the Berkeley Guidance Study, the Berkeley Growth Study, and the Oakland Growth Study (all cited in Wohlwill, 1980). In all these studies, the indicator of choice for cognitive growth was IQ scores. A number of researchers have used the findings from these studies to analyze the extent to which individual children maintain their standing relative to their peers and the factors that increase and reduce stability (Wohlwill, 1980).

Three types of stability are particularly relevant to the present study: absolute invariance, regularity of form, and constancy of relative position. Absolute invariance refers to the step-like pattern of growth—recurring periods of rapid qualitative and quantitative change surrounded by apparent plateaus. Regularity of form refers to predictable patterns and sequences of change, without which normative comparisons are meaningless. Understanding invariance and assuming some level of regularity of form are necessary conditions for the element of most interest to this study: constancy of relative position (Wohlwill, 1980).

An example of absolute invariance is the development of the concept of conservation as described by Piaget. Picture a young child presented with two glasses of water, both filled to the same height, and one wider than the other. Before the emergence of the concept of conservation, children consistently believe that the amount of water in the two glasses is the same. After the emergence of conservation, children consistently understand that the wider glass contains more water. If the children's understanding of conservation is assessed on two occasions, with both occasions occurring either before or after its attainment, little demonstrable change occurs and the child's behavior is perfectly predictable. We can observe changes in the child's understanding of the concept only if our measurement occasions straddle the acquisition of the concept. Absolute invariance is important to the study of stability of growth because it reinforces our understanding that growth does not occur in a continuous linear fashion and that the timing of assessment relative to these periods of invariance affects outcomes and apparent stability.

Constancy of relative position is normative and relies on reference to a regular form of change (or norm). This is easily the most intractable part of the measurement challenge. Wohlwill (1980) notes that some of the researchers of his time (e.g., Thurstone and Ackerman) relied on cross-sectional data. Others used longitudinal data but without an analytic model and so had "nothing more than an impressionistic representation of an individual curve's deviation from the group trend" (p. 369). He warns that

> Stability coefficients provide information on the stability of intellectual performance across age for a group of individuals rather than for a given child... The extensive literature on correlations between IQs obtained over a given age span amounts to a laborious (and for the most part surprisingly successful) effort to establish the robustness of this index as a measure of intellectual attainment relative to a norm that is largely unaffected by gross changes in actual cognitive functioning. But interpreting the magnitude of such coefficients in terms of the relative immutability of intelligence or its impermeability to environmental influences is an entirely different matter (p. 370).

As noted above in the discussion of vertical equating, the developers of the CTP III and SAT-9 still use cross-sectional data to develop their version of the cognitive growth curve – the developmental scale score. The question of the reasonableness of the assumption that cognitive growth for individuals or for groups of students should parallel that of the norm group goes unasked and, hence, unanswered.

Applications: Research on IQ

In their landmark report based on Fels data for middle-class children with a normal distribution of ability, McCall, Appelbaum, and Hogarty (1973) provide concrete evidence of the instability of IQ over time. The IQ of the average child in their analyses shifted 28.5 points (1.78 *sd*) between the ages of 2 1/2 and 17 years. While 21% had changes of less than 20 points, 43% had changes of 21–30 points, 36% had changes of 31–40 points, and 14% had changes of more than 40 points (p. 63).

McCall et al. (1973) also found strong evidence that there is no single prototype growth curve. Their analyses revealed five distinct patterns of change over time. Overall, 45% of their sample evidenced a relatively constant pattern of growth while the remaining children displayed "marked changes in IQ that were not simply random fluctuations about a constant value" (p. 64). They noted that

> clusters of IQ changes were associated with global assessments on parental behavior which could be sensibly interpreted and were reasonably consistent with other literature. For example, one of the two parental behaviors found to distinguish between increasing and decreasing preschool trends was the acceleration attempt of parents for intellectual achievements (p. 64).

The research reviewed above is about change in IQ over time. The data available for this study are achievement test scores. Because ability and achievement are so intertwined, we might reasonably generalize the findings related to the stability of IQ to the expected stability of achievement test scores. Additional support for this generalization of the findings comes from the contention of many experts in educational assessment, including Popham (2000), who make effective arguments that much, if not most, of what is measured by the norm-referenced achievement tests is a function of aptitude and knowledge acquired (or not acquired) outside the school.

Applications: Research on Achievement

Recent research on the prediction of early school achievement is related to the question of normative stability. While typically not of the same depth as the longitudinal studies of ability described above, these studies produce findings that are consistent with those related to stability of IQ.

La Paro and Pianta (2000) and Tramontana, Hooper, and Seltzer (1988) conducted meta-analytic reviews of the longitudinal research literature on prediction of achievement in the early elementary grades. Their findings are consistent with the hypothesis that normative achievement at this age tends to be unstable.

Tramontana et al. (1988) reviewed 74 studies conducted between 1973 and 1986. They reported that the pattern of effective predictors of reading varied according to the grade level in which outcomes were assessed. Their findings are summarized as follows:

> There was a convergence of evidence suggesting that overall prediction is actually better for *later* grades, at least from the second or third grade on.... This suggests that academic skills sometimes may not be sufficiently developed or stable by the end of first grade to be assessed reliably (p. 134).

Twelve years later, in an effort to inform the debate about school readiness assessments, La Paro and Pianta (2000) also reviewed the literature on the preschool or Kindergarten prediction of later school achievement. They found "there are no published quantitative estimates of the stability of individual differences in children's skills and abilities from preschool or Kindergarten to the early grades that are based on available published studies and include nonsignificant as well as significant findings." (p. 445)

La Paro and Pianta (2000) undertook a meta-analysis of the research published between 1985 and 1998. This period picked up about where Tramontana et al. (1988) left off and included the years after the 1989 publication of the National Educational Goals which emphasized children's readiness for school. La Paro and Pianta's work focused on the extent to which indicators of children's development at preschool or Kindergarten predicted performance in Kindergarten through Grade 2. They analyzed results from 70 published reports of longitudinal studies that used data from 62 independent samples.

They summarize their findings as follows:

> Individual differences in children's academic/cognitive and socio/behavioral development, as assessed through a variety of methods before or just after the children enter school, account for a small to moderate portion of the variance in similar outcomes assessed in the early school years....
>
> The results suggest that although individual differences in child characteristics, particularly in the academic/cognitive domain, are moderately stable across the period, there is substantial variability within the two major domains of child school performance that is not accounted for by prior assessment of similar constructs (pp. 472–473).

They go on to conclude that:

> The small to moderate effect sizes we found for academic/cognitive and social/behavioral outcomes clearly indicate that children's rank order changes over the preschool to second grade period.... Instability or change may be the rule rather than the exception during this period (p. 476).

Additional support for the conclusions of Tramontana et al. (1988) and La Paro and Pianta (2000) is found in the results of longitudinal studies conducted by Lindquist

(1982); Mills and Jackson (1990); Schmidt and Perino (1985); Storch and Whitehurst (2002) and Williamson, Appelbaum, and Epanchin (1991).

In a study of the Denver Developmental Screening Test, Lindquist (1982) found that 20% of children identified as at-risk for educational problems when they registered for Kindergarten were later achieving in the above average range and 22% of those who were below average had not been identified by the Denver. Mills and Jackson (1990) studied young children who had been identified as precocious readers in Kindergarten and found that in later elementary school grades these children were good readers but not necessarily exceptional. Schmidt and Perino (1985) studied the prediction of school achievement at the end of Grade 2 from screening tests administered at entry to Kindergarten using the Vane Kindergarten Test and Vane Test of Language. They found the Kindergarten screening correctly identified about 77% of the low-achieving students and 73% of the high-achieving students.

The work on the development of mathematical models of individual growth continues to progress. In their longitudinal analysis of academic achievement, Williamson et al. (1991) applied developments in mathematical models of growth to the study of change for a cohort of public school students as they progressed from Grade 1 to Grade 8. In their analyses, Williamson et al. used the tracking index developed by Foulkes and Davis (1981, cited in Williamson et al.). The tracking index γ (gamma), "is the empirical probability that two randomly chosen growth curves do not cross in the observed range of time" (p. 71). If the growth curves do not cross, students would have maintained their relative position to each other over the period of the study.

Williamson et al. (1991) found values of gamma for reading of .76 for female students and .74 for male students. They found values of gamma for mathematics of .72 and .69, again for female and male students, respectively. Thus, while many student pairs maintained their relative rank order (but not necessarily a consistent magnitude of distance in rank) between 24 and 31% of student pairs did *not* maintain the same relative ranks.

While these values of gamma may seem to suggest a tendency toward stability of rank order, the study population included the full range of achievement in the district (presumably, not including special education students exempt from testing). The average ability test scores for the participants in the study were just slightly above the 50th percentile (mean Cognitive Abilities Test scores collected in Grade 2 were approximately 105 on the Verbal, Mathematics, and Non-Verbal scales). If Williamson et al. (1991) had been modeling growth for a subset of students with more extreme scores, it is reasonable to expect that the rate of crossing growth lines would be higher due to regression toward the mean and any ceiling effect for those students.

Summary

As this review of the literature has demonstrated, there are a number of factors that could affect the stability of results between one test and another. Regression toward the mean and the ceiling effect, when present, impact the scores of students, particularly those with extreme scores. Reading and mathematics are complex, multidimensional skills and the critical elements of each change over time. These changes are reflected in the corresponding achievement tests and affect the comparability and equatability of test levels used in longitudinal research. In addition, widespread assumptions about the stability of ability and achievement, particularly during early childhood, are not supported by research. All of these present plausible, if partial, explanations for the observed changes in achievement test scores at the school that is the subject of this study.

Recall that to draw validity inferences, we need to rule out plausible competing hypotheses. To accept the hypothesis, that the school was failing the students, requires that we reject the hypothesis that the observed changes can be explained by any one or a combination of regression to the mean, the ceiling effect, the changing nature of reading and mathematics, and the inherent instability of normative achievement in young children.

In Chapter 3, I describe how I applied the findings and/or methods from the literature review to understand the regression in the test scores for the school contributing data to this study.

CHAPTER 3. METHODS

Subjects and Data

The subjects for this study are students who completed Grade 6 between 1999 and 2001. These cohorts were selected because they were tested every year between Kindergarten and Grade 4. Test data are not available at all these points for previous and succeeding cohorts. Before attrition due to disenrollment or missed tests, there were 192 students in this group; the number of students with complete data is 172 (90% of the entering students). The selection process and achievement profile of these students at entry to the school are described in some detail in Chapter 1.

For the purposes of this study, I combined data from the three cohorts. This provided improved generalizability of the findings by reducing cohort effects, and the larger N provided more statistical power. However, this required aggregating scores from the CTP III and the SAT-9.

Comparability of CTP III and SAT-9 Results

Prior to making the decision to combine the data, I examined the comparability of the results of the CTP III and the SAT-9 (using *NCE* scores for the national percentile ranks of the students). I compared cross-sectional data from the two tests from consecutive years of tests. For example, I compared the CTP III results for Grade 1 in 1996 with those from the SAT-9 for Grade 1 in 1997. I repeated this process for each the tests at all levels included in this study (i.e., all levels of both tests for students in Grades 1 through 4).

Cross-sectional results on the CTP III administered in 1996 and the SAT-9 administered in 1997 are shown in Figures 5 and 6. I computed *t* tests of the means (using *NCE* scores) and effect sizes for each pair of points in Figures 6 and 7. The differences in means for three of the eight pairs were statistically significant: Grade 4 Reading, *t* (185) = 3.0, *p* = .003 (two-tailed), Cohen's *d* = 0.5; Grade 1 Mathematics *t* (131) = 4.8, *p* < .0001 (two-tailed), Cohen's *d* = 0.76; and Grade 3 Mathematics *t* (110) = 2.8, *p* = .006 (two-tailed), Cohen's *d* = 0.54. (Results for all pairs of data points are presented in the Appendix in Tables A1 and A2.)

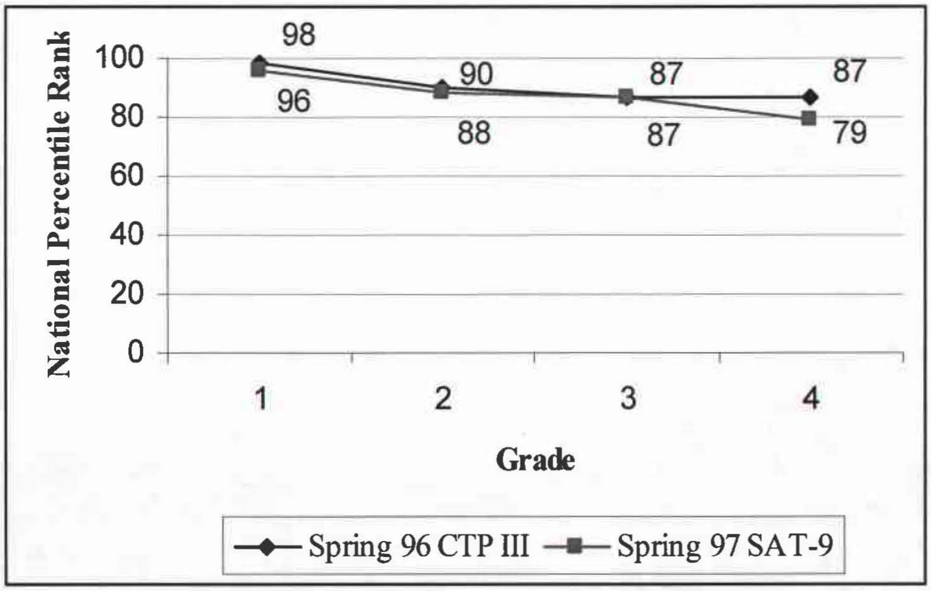

Figure 6. Comparison of national percentile ranks for reading on the CTP III and the SAT-9, based on cross-sectional data for two consecutive years.

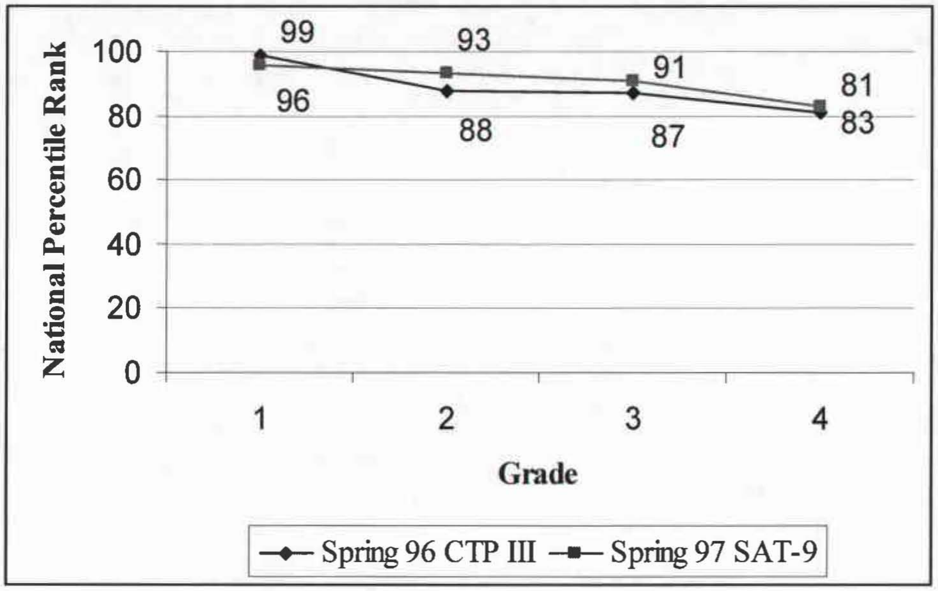

Figure 7. Comparison of national percentile ranks for mathematics on the CTP III and the SAT-9, based on cross-sectional data for two consecutive years.

The results seemed reasonably comparable, but the observed differences, or similarities, could be related to differences between the cohorts. To understand how the effects observed above compare to the typical cohort effect, I examined changes in mean scores between contiguous cohorts on the *same* test battery: CTP III scores from 1995 and 1996 and SAT-9 scores from 1997 and 1998, for reading and mathematics at each of the grade levels, for a total of 14 comparisons (the SAT-9 was only administered at Grade 1 in 1997). The differences in *NCE* scores ranged from 0.2 to 8.7; 6 of the 14 differences were statistically significant. For these groups of students, the impact of the change from the CTP III to the SAT-9 was no greater than the typical variation between cohorts. See Tables A3 through A6 in the Appendix for more information.

The above analyses strongly suggest that—at least for this population—the results of the CTP III and SAT-9 are comparable when using national norms. Therefore, the results from these tests are treated as equivalent in this study, and their common metric, *NCE* scores, is used in computations involving the combined data.

Analytical Methods

Regression Toward the Mean

The literature reviewed in Chapter 2 identifies two primary factors in regression toward the mean: (a) the lack of perfect correlation between variables, and (b) measurement error. Wainer (1999) used test reliability coefficients with the subset of the data available to him to derive true score estimates for Grades 1 and 4 which he then compared to actual scores.

In my analysis of the effects of regression to the mean, I built on Wainer's work. Using the same formula employed by Wainer (Kelley's formula), I estimated true scores for Reading and Mathematics at Grade 1. Next, I computed expected scores after regression due to imperfect correlations for Grades 2, 3, and 4. To do this, I used the estimated true scores for Grade 1 in the formula used by Nesselroade et al. (1980).

Lacking actual data on the correlation between the tests, I used two proxies in my calculations. The first is the reliability coefficients for the tests (K–R 20). Because the K–R 20 is a measure of internal consistency and not a direct indicator of consistency of student performance over time, it is less desirable than more direct information on the consistency of student performance across test levels. However, the K–R 20 statistics are available from both test publishers while correlations across levels are not. Another available alternative is the alternate forms correlation. Both the CTP III and the SAT-9 had two test forms at each grade level and provided information on the correlations between the scores for these two forms from their standardization and horizontal equating studies. However, these statistics are also indirect estimates of what the expected correlations might be *across* levels. Because the values of K–R 20 are consistently higher that the alternate forms correlations I chose to use the K–R 20 to estimate the upper bounds of the regression due to less than perfect correlations between tests.

The second proxy is the correlation between the tests for the study population, adjusted for attenuation. Because the Grade 1 scores for the sample are highly skewed (see Tables 4 and 5) and because of the modest sample size, the observed correlation is likely to be smaller than what would be obtained if it could be measured with a population comparable to the students in this study or without the measurement error caused by the test ceiling. The use of the reliability coefficients will predict a minimum amount of regression, and the use of the correlation coefficients will predict the maximum amount expected.

The Test Ceiling

Bracken (1990) suggested that if the highest item on a test was not at least two standard deviations above the mean, the potential for a ceiling effect existed. I suggest that, when the frame of reference is performance relative to a norm group, the raw score equivalent to the 99th percentile is the effective ceiling because raw scores beyond this point do not increment any additional change in the scores. I used Bracken's guideline to determine the potential ceiling for each of the tests at each grade level, with the score equivalent to the 99th percentile defining the effective ceiling. I then determined the percentage of students in this study who reached or approached the effective test ceilings.

In the review of the literature, I noted that the issue of item gradients is also a concern in looking at scores that approach the limit of the test. Bracken (1990) also offered a guideline for determining when item gradients might be too steep to return valid or reliable results (i.e., when an increase or decrease of a single raw score alters the corresponding standard score by more than one third of a standard deviation). I used Bracken's guidelines to examine the item gradients for the observed scores.

The Changing Content of the Tests and Vertical Equating

For tests to be valid indicators of achievement, they need to respond to the changing nature of the knowledge and skills in each domain. To describe how the tests changed, I examined the changes in subtests, strand, and the number of items in each for the CTP III Reading Comprehension and Mathematics tests and for the SAT-9 Total Reading and Total Mathematics tests.

I also looked at the apparent reasonableness of the outcomes of the vertical equating. To identify any apparent discontinuities in scale scores, which are related to the distributions of the percentile ranks, I plotted the growth in scale scores between spring of Grade 1 and Grade 4 for the CTP III and SAT-9.

Instability of Growth

One of the definitions of stability offered by Wohlwill (1980) is the preservation of individual differences over time. Wohlwill notes that to estimate this we need prototypic models of growth. Wohlwill and others (e.g., McCall et al., 1973), emphasize the plurality of growth curves and the need to select the curves most appropriate to the individuals (or groups) being studied.

Lacking an appropriate empirical model of expected growth curves for the students in this study, I resorted to another proxy: the comparison of mean scores from alternative norm groups. In addition to the national norm group, the CTP III and SAT-9 both offer alternative norms. The CTP III includes norms for suburban and independent Schools. The SAT-9 includes norms for high socioeconomic status (SES) and private schools. The alternative norms for both tests are based on data from all users of those tests and, therefore, cannot be construed to be nationally representative of the populations from which they are drawn. In fact, the level of performance represented by the SAT-9 Private School norm group is more comparable to that represented by the CTP III Suburban norm than it is to the Independent School norm.

To analyze the patterns of the means of the various norm groups, I identified the mean scale score for each of the alternative norm groups and then determined the national percentile rank of those scores. I compared these national percentile ranks to get a sense of how stable the performance of these groups might be in relation to each other. These group means would, of course, be more stable than those of individual students, or even smaller groups of students, like the moderate number of students in this study. The generalizability of the group means to the current study is also limited by the fact that they are cross-sectional, rather than longitudinal.

CHAPTER 4. RESULTS

Regression Toward the Mean

The literature reviewed in Chapter 2 identifies the lack of perfect correlation between variables and measurement error as primary factors in regression toward the mean. In the analyses reported here, I build on analyses originally conducted by Wainer (1999).

Given his audience and purpose, Wainer chose to apply hypothetical (but reasonable) test reliability coefficients to the data from one cohort of students to derive true score estimates. Based on his comparison of the estimated true scores for Grades 1 and 4, he concluded that "we shouldn't expect a regression effect, *due to the unreliability of the test* [emphasis added], to shrink the performance of fourth graders toward the mean as much as was observed…" (p. 28). Note that his computations accounted for only one of the factors contributing to regression toward the mean.

Estimated True Scores

Following Wainer's example, I used Kelley's formula to estimate the true score of the typical (average) examinee from average of the observed scores:

$$\hat{\tau} = \rho\bar{x} + (1 - \rho)\mu$$

where $\hat{\tau}$ is the estimated true score, ρ is the reliability of the test, $\bar{x}$ is the mean observed score, and μ is the mean of the population of the group to which the student belongs (Kelley, 1947, cited in Wainer, 1999). I used *NCE* scores in all the calculations.

As explained in Chapter 3, lacking a good estimate of test reliability for populations similar to the one at this school, I used the reliability coefficients (K–R 20) for the CTP III as reported by the test publishers (Educational Testing Service, 1995). The results of these calculations are presented in Tables 9 and 10.

Table 9
Estimated True NCE Scores After Correcting for Unreliability: Grade 1

Domain	Mean *NCE*	K–R 20	Estimated True Score
Reading	85	.92	82
Mathematics	93	.82	85

I chose to use the reliability coefficients (K–R 20) from the CTP III rather than those from the SAT-9 (K-R 21). Because the CTP III is designed for use with students with higher levels of achievement, the test includes more items of greater difficulty (Educational Testing Service, 1995). The increased variation in item difficulty reduces the test reliability. The reliability is also affected by the length of the test. With the exception of Mathematics in Grades 3 and 4, the CTP III has fewer items than does the SAT-9. As a result of the differences in item difficulty and test length, the reliability coefficients for the CTP III are lower than those for the SAT-9 (Harcourt Brace Educational Measurement, 1997c) and may be more representative of the actual reliability of these tests when used with above-average students, like those in this study. However, it should be noted that this choice had little effect on the outcome. The differences between the values of the K–R 20 for the CTP III and the K–R 21 for the SAT-9 ranged from 0 to .04.[5]

Estimated Regression Effects

Starting with the estimated true scores for Reading and Mathematics from Table 9, I calculated the expected scores at Grades 2 through 4 using the formula suggested by Nesselroade et al. (1980, p. 627): Expected Score = corr $(X_1, X_n)*x$. As with the example provided in Chapter 2 (see Tables 6 and 7) and as described in Chapter 3, I used two strategies to calculate expected scores after regression due to imperfect correlations.

In the first set of calculations (shown in Tables 10 and 11), I used the published reliability of the tests (K–R 20) as a proxy for the correlations between the tests. Again, I chose to use the reliability coefficients from the CTP III. The resulting expected scores for Reading are very close to the actual scores at all grade levels (a difference of 1 or 2 *NCE*s). For students in Grades 2 and 3, the actual scores for Mathematics are higher than the expected scores for students in Grades 2 and 3 (a difference of 6 *NCE*s). The expected and actual scores for Grade 4 are both 72.

[5] The K–R 21 is reported in the technical manual for the SAT-9 in place of the K–R 20. The K–R 21 is computationally simpler than the K–R20 but yields a lower valued. This discrepancy between the results of the two formulas is most pronounced on tests that are both short and relatively easy (Nitko, 1983).

Table 10
Expected Reading NCE Scores After Correcting for Imperfect Correlations: Based on K-R 20

Grade	Estimated True Score	K–R 20	Expected Score	Actual Score
1	82			85
2		0.89	73	75
3		0.89	73	75
4		0.88	72	71

Table 11
Expected Mathematics NCE Scores After Correcting for Imperfect Correlations: Based on K-R 20

Grade	Estimated True Score	K–R 20	Expected Score	Actual Score
1	85			93
2		0.84	69	75
3		0.88	72	78
4		0.88	72	72

In the second set of calculations, I used the actual correlation coefficients from the longitudinal data for the students at this school after adjustments for attenuation.[6] Using this methodology, the observed scores are much higher than the expected scores (see Tables 12 and 13). Observed scores for Reading are 17 to 27 *NCE*s higher than predicted with this method. Observed scores for Mathematics are 25 to 30 points higher than predicted. These findings might suggest that the correlation coefficients from the school are unreasonable estimates for the population. However, these correlations are of the

[6] $r' = \frac{r_{12}}{\sqrt{r_{11}}\sqrt{r_{22}}}$

(Nunnally, 1970, p. 117)

same magnitude as those typically reported from longitudinal studies of achievement over this span of grades (see meta-analysis by La Paro & Pianta, 2000).

Table 12
Expected Reading NCE Scores After Correcting for Imperfect Correlations: Based on Sample Correlations

Grade	Est. True Score	Adj. Corr. with Grade 1	Expected Score	Actual Score
1	82			85
2		.71	58	75
3		.59	48	75
4		.63	52	71

Table 13
Expected Mathematics NCE Scores After Correcting for Imperfect Correlations: Based on Sample Correlations

Grade	Est. True Score	Adj. Corr. with Grade 1	Expected Score	Actual Score
1	85			93
2		.60	50	75
3		.61	50	78
4		.49	41	72

Test Ceilings

As demonstrated above, regression toward the mean can explain the changes in test scores observed at this school. However these analyses do not address the question of whether or the achievement test scores observed in Grade 1 (the 95th percentile for Reading and the 98th percentile for Mathematics) were accurate representations of the children's achievement. If these scores were inflated due to the test ceiling or related attributes of the test, they would provide misleading information and could lead to inappropriate expectations for subsequent performance.

Effective Ceiling Index

Bracken (1990) suggests that the potential for a ceiling (or floor) exists if the most extreme scores for any given test are less than two standard deviations from the mean. I examined the norms tables for the CTP III and SAT-9 to determine the distance between the mean and maximum scores. Since the frame of reference for interpreting the test scores in this study is the national percentile rank, I also looked at the distance between the mean and the raw score equivalent to the 99th percentile (referred to as "Total Effective Items" in Tables 14 and 15). I suggest that this number is the effective length of the test (when using the national percentiles) because raw scores above this provide no additional increment to the percentile rank or the *NCE*.

I computed the Effective Ceiling Index as follows:

(Total Effective Items – Mean Raw Score)/*sd* of Raw Scores

An Index value of less than 2.0 falls below Bracken's guideline, indicating a potential ceiling effect.

As shown in Table 14, all the reading tests have potential ceilings. The index values for the reading tests ranged from 1.2 to 1.8, with the most severe ceilings expected for reading at Grades 1 and 2 when measured with the CTP III.

Table 14
Effective Ceiling Index: Reading Tests

	CTP III Reading Comprehension				**SAT-9 Total Reading**			
	Grade				Grade			
	1	2	3	4	1	2	3	4
Raw Score Mean	24.3	28.5	27.6	24.9	69.8	73.6	54.5	56
sd	9.5	9.1	8.7	8.6	20.6	22.9	16.3	16.7
Total Items	40	45	45	45	106	118	84	84
Total Effective Items	36	40	40	38	104	114	81	81
Effective Ceiling Index	1.2	1.3	1.4	1.5	1.7	1.8	1.6	1.5

Note. Test statistics are from the technical manuals for the CTP III (Educational Testing Service, 1995, pp. 42–43) and SAT-9 (Harcourt Brace Educational Measurement, 1997c pp. 68–71). Effective Items are from the norms tables for each test (Educational Testing Service, 1997 and Harcourt Brace Educational Measurement, 1997b, respectively).

All the mathematics tests also have potential ceilings (see Table 15). These are especially low for mathematics at Grades 1 and 3 using the CTP III (with Effective Ceiling Index values of 0.5 and 1.1, respectively).

While both the CTP III and the SAT-9 demonstrate potential ceilings for students with above-average achievement, the ceiling is lower on the CTP III for test takers at Grades 1, 2, and 3. The Total Effective Items for the CTP III is well below the total number of items on the tests.[7] Therefore, when relying on the national norms as the reference group for interpretation, the CTP III is actually a shorter test (at some grade

[7] This is because more of the items on the CTP are relatively difficult. The purpose of including more items of greater difficulty is to discriminate among the above-average students for whom the test is designed. The more difficult items do increment changes in the scores when using the suburban and independent school norms. These findings suggest that CTP III test users who employ the national norms or use the NCE scores in computations are working with a shorter test with reduced reliability and are discarding much of the variation this particular test was designed to measure.

levels, a *much* shorter test). It has been repeatedly demonstrated that the reliability of test scores decreases as the number of items decreases. This reduced reliability affects the validity of test-based inferences and the stability of tests scores whether viewed from a longitudinal or cross-sectional perspective.

Table 15

Effective Ceiling Index: Mathematics Tests

	CTP III Mathematics				**SAT-9 Total Mathematics**			
	Grade				Grade			
	1	2	3	4	1	2	3	4
Raw Score Mean	27.3	28.9	54.5	44.6	44.1	49.4	49.5	49.1
sd	7.1	7.5	15.1	14.1	12	13	14	15
Total Items	45	45	100	100	69	74	76	78
Total Effective Items	31	40	71	69	66	71	73	74
Effective Ceiling Index	0.5	1.5	1.1	1.7	1.8	1.7	1.7	1.7

Note. Test statistics are from the technical manuals for the CTP III (Educational Testing Service, 1995, pp. 42–43) and SAT-9 (Harcourt Brace Educational Measurement, 1997c pp. 68–71). Effective Items are from the norms tables for each test (Educational Testing Service, 1997 and Harcourt Brace Educational Measurement, 1997b, respectively).

Impact of the Effective Ceiling

The impact of a test ceiling is of most concern when students' scores are relatively high although the ceiling affects scores at all levels (Yen, 1983). One of the challenges with combining data from two batteries of tests in this set of analyses is that the number of items in each test is different. To represent the extent to which the students in this study had scores that were approaching the test ceiling using one metric, I computed the percentage of correct responses (the raw score divided by the Total of Effective Items for each test). The distribution of the percentage correct statistic at each grade level is presented in Table 16.

The percentage of students with scores at the 99th percentile in Reading ranges from 39% at Grade 1 to 6% at Grade 4. The tests in use are not able to discriminate between the varying levels of achievement for these students. If we expand this to include those whose scores are within 5% of the effective ceiling (one or two items below the ceiling depending on test length), the test may not be able to accurately assess variations in performance for up to 46% of the students in Grade 1.

The lack of discrimination in Mathematics is even more severe. Fully 71% of the students in Grade 1 and 18% of the students in Grade 3 had scores at or above the effective ceiling. If we again include those students whose scores are within 5% of the ceiling, the test may not be effective for up to 79% of the students in Grade 1 and 37% in Grade.3

I also investigated item gradients as suggested by Bracken (1990). He suggested that if a raw score change of 1 results in a standard score change of more than one third standard deviation the gradient is too steep. I used the mean raw scores for reading and mathematics for the CTP III and SAT-9 at each grade level to assess the item gradients. No gradients close to one third of a standard deviation were found. Results of these analyses are presented in the Tables A7 and A8 in the Appendix.

Table 16
Percentage of Students Approaching the Test Ceilings

Grade	Correct Responses as a Percentage of the Effective Test Ceiling	Reading	Mathematics
	<.95	54	22
1	<1.00	7	8
	1.00	39	71
	<.95	76	78
2	<1.00	13	14
	1.00	11	8
	<.95	76	63
3	<1.00	16	19
	1.00	8	18
	<.95	67	78
4	<1.00	27	16
	1.00	6	6

Changes in the Content of the Tests and Vertical Equating

Changes in Content

For tests to be valid indicators of achievement, they need to reflect the changing nature of the knowledge and skills in each domain. The changes in subtests and the number of items in each for SAT-9 Total Reading and Total Mathematics are provided in Tables 17 and 18 below. The same information for the CTP III is provided in Tables 19 and 20.

In the SAT-9, the assessment of phonemic skills (Word Study Skills) and sight vocabulary (Word Recognition) comprise 34 and 28% of the Grade 1 test, respectively.

Comprehension contributes 38% of the test score at this grade. Similarly, the Grade 2 test consists of phonemic skills (41%), vocabulary (25%), and comprehension (34%). Direct assessment of phonemic skills drops out of the tests at Grade 3. At this level comprehension moves to the forefront with 64% of the items, with the remainder of the items going to Reading Vocabulary.

The shifts in SAT-9 Total Mathematics are not as obvious, but some changes are evident. For example, Number Sense and Numeration declines from 17% of the test at Grade 1 to 8% in Grades 3 and 4. In addition to these quantitative shifts, there are qualitative shifts, with items and associated calculations becoming increasingly sophisticated and complex.

Table 17
SAT-9 Subtests, Strands, and Number of Items in Each by Test Level: Total Reading

Subtest	Test Level and Grade Administered			
Strand	Primary 1	Primary 2	Primary 3	Intermediate 1
	Grade 1	Grade 2	Grade 3	Grade 4
Total Items	106	118	84	84
Word Study Skills	36	48		
Structural Analysis	12	12		
Phonetic Analysis – Consonants	12	18		
Phonetic Analysis - Vowels	12	18		
Word Reading	30			
Reading Vocabulary		30	30	30
Synonyms		18	18	16
Multiple Meanings		6	6	7
Context		6	6	7
Reading Comprehension	40	40	54	54
Two–sentence stories (Riddles)	5			
Short Passages (Cloze)	15			
Short Passages with Questions	20			
Initial Understanding	9	14	14	12
Interpretation	11	20	24	24
Critical Analysis		3	8	9
Process Strategies		3	8	9

Note. Number of items per subtest and strand (source: Harcourt Brace Educational Measurement, 1997c, pp. 139–151).

Table 18
SAT-9 Subtests, Strands, and Number of Items by Test Level: Total Mathematics

Strand	Test Level and Grade Administered			
	Primary 1 Grade 1	Primary 2 Grade 2	Primary 3 Grade 3	Intermediate 1 Grade 4
Total Items	69	74	76	78
Problem Solving	44	46	46	48
Number Sense and Numeration	12	10	6	6
Concepts of Whole Number Computation	3	4	4	4
Fraction and Decimal Concepts	3	3	4	6
Patterns and Relationships	5	5	3	3
Statistics and Probability	5	6	6	6
Geometry and Spatial Sense	5	5	6	6
Measurement	8	10	10	10
Estimation			3	3
Problem Solving Strategies	3	3	4	4
Procedures	25	28	30	30
Computation in Context	6	8	9	12
Computation Using Symbolic Notation	11	12	12	12
Number Facts	8	8	6	3
Rounding			3	3

Note. Number of items per subtest and strand (source: Harcourt Brace Educational Measurement, 1997c, pp. 139–151).

The CTP III uses a different taxonomy to classify items, yet the change in emphasis is still evident (see Tables 19 and 20.) For example, in the Reading Comprehension test, items related to vocabulary disappear in Grade 4. At the same time, items related to explicit information are cut roughly in half while those requiring inference and deduction nearly double. In Mathematics, the trend is also clear with items related to number systems and number theory, measurement, statistics and probability, and pre-algebra increasing while the weight given to number sense and whole number operations declines.

Table 19

CTP III Subtests, Strands, and Number of Items in Each by Test Level: Reading Comprehension

Subtest	Test Level and Grade Administered			
Strand	Level A Grade 1	Level B Grade 2	Level C Grade 3	Level D Grade 4
Total Items	40	45	45	45
Vocabulary	8	6	5	
Generalization	5	4	3	
Application	3			
Breadth of Meaning		2	2	
Recognize Explicit Information	11	11	13	14
Supporting Details	8	6	11	8
Ordered Information	3	5	2	6
Identify Explicit Main Ideas	8	8	2	3
Analyze	6	8	9	9
Cause and Effect or Condition	3	5	3	4
Motive or Intention			2	4
Contrast and Compare	1		2	
Pronoun Referent	2	3	2	2
Hypothesize	7	8	14	17
Implicit Main Idea	1	2	5	
Infer Details and Ordered Information	2	2	2	
Infer Character Motives		1	2	3
Infer Purpose	1	1	1	3
Draw Conclusions	3	2	4	11
Summarize		4	2	2

Note. Number of items per strand (Educational Testing Service, 1993, pp. 16–25). Categories containing eight or more items are reported separately.

Table 20
CTP III Subtests, Strands, and Number of Items in Each by Test Level: Mathematics

Strand	Test Level and Grade Administered			
	Level A Grade 1	Level B Grade 2	Level C Grade 3	Level D Grade 4
Total Items	45	45	100	100
Communications	10	6	14	16
Number Sense	3	3	4	3
Whole Number Operations	2	1		
Geometry/Spatial Sense	1	1	3	7
Measurement	2		3	3
Statistics/Probability	1		2	2
Fractions/Decimals	1	1	2	
Pre-Algebra				1
Reasoning	7	11	13	10
Number Sense	1	5	3	
Numbers and Number Relationships				2
Number Systems and Number Theory				3
Whole Number Operations	4	5	5	
Geometry/Spatial Sense	1		2	2
Measurement	1	1		1
Statistics/Probability			2	1
Fractions/Decimals			1	
Pre-Algebra				1
Problem Solving	11	11	28	41
Number Sense	2		1	
Numbers and Number Relationships				5
Number Systems and Number Theory				7
Whole Number Operations	6	6	11	
Geometry/Spatial Sense			5	3
Measurement	2		2	9
Statistics/Probability	1	2	3	11
Fractions/Decimals		1	6	
Pre-Algebra				6

Table 20 (continued).

CTP III Subtests, Strands, and Number of Items in Each by Test Level: Mathematics

Computation/Estimation	8	9	24	23
Whole Number Operations	6	4	19	
Numbers and Number Relationships				2
Number Systems and Number Theory				10
Measurement	1	4	1	9
Fractions/Decimals	1	1	4	1
Pre-Algebra				1
Patterns and Relationships/Functions	9	8	10	10
Number Sense	3	2	4	
Number Systems and Number Theory				2
Whole Number Operations		3	2	
Geometry/Spatial Sense	1	1	4	3
Measurement	2	2		
Fractions/Decimals	3			
Statistics/Probability				3
Pre-Algebra				2

Note. Number of items per strand (Educational Testing Service, 1993, pp. 61–81). Categories containing eight or more items are reported separately.

These analyses of the content of the tests clearly demonstrate that there are substantial quantitative and qualitative changes in the content of the tests over the span of Grade 1 to Grade 4. This is true in the areas of reading and mathematics and for both the SAT-9 and the CTP III. While the changes in content are sensible, the fact of the changes suggests that users should expect the performance of students relative to that of the norm group is likely to change as the alignment of the tested content and their personal strengths changes over the grade levels. The same would also pertain at the school level as the match between the tested content and curriculum varies across grade levels.

Vertical Equating

Both Burket (1984) and Hoover (1984a and 1984b) noted that vertical equating can produce results that appear unreasonable. In Figures 8 through 11, I plotted the growth in scale scores from spring of Grade 1 through spring of Grade 4 for reading and mathematics for both tests used in this study. The plots reveal some of the anomalies in the pattern of expected growth. (Sources: Educational Testing Service, 1997, 2001; Harcourt Brace Educational Measurement, 1997a, 1997b.)

For high-achieving students, the largest change in scale scores in SAT-9 Total Reading occurs *over the summer* between Grades 1 and 2, a period of about 10 weeks of instruction interrupted by three months of summer vacation.[8] At the same time, very little growth is expected for students with scores at the 89th or 96th percentiles over the course of Grade 2. In fact, the scale score values of 662 and 666 shown for students at the 96th percentile *do not correspond to any raw score*. The closest value in the norms booklet is 664, corresponding to a raw score of 110. Thus, there is *no difference* in scores for the highest achieving students between fall and spring of Grade 2. The change in scale score from 634 to 642 for students at the 89th percentile is equivalent to a two-point change in raw score, which is well within the standard error of measurement of 4.3.

Results for SAT-9 Total Mathematics are more consistent with what we expect—little change over the summer months, followed by more change during the school year,

[8] The testing window for the SAT-9 is September 15–October 15 for fall (i.e., 5–8 weeks into the school year) and April 1–April 30 (i.e., 30–34 weeks into the school year) for spring administrations (Harcourt Brace Educational Measurement, 1996, p. 2).

and with growth noted in scores at all levels of the distribution. However, we might expect "summer loss" to produce lower scores in fall than in spring.

The changes in scale scores on the CTP III also appear grossly consistent with our expectations, again with the exception of no evidence of summer loss. However, the test publishers explain that "many CTP-user schools test their students late in the fall, after one or two months (or more) of instruction, while many others test early in the spring, one or two months (or longer) before the end of the school year" (Educational Testing Service, 2004, p. 29).[9]

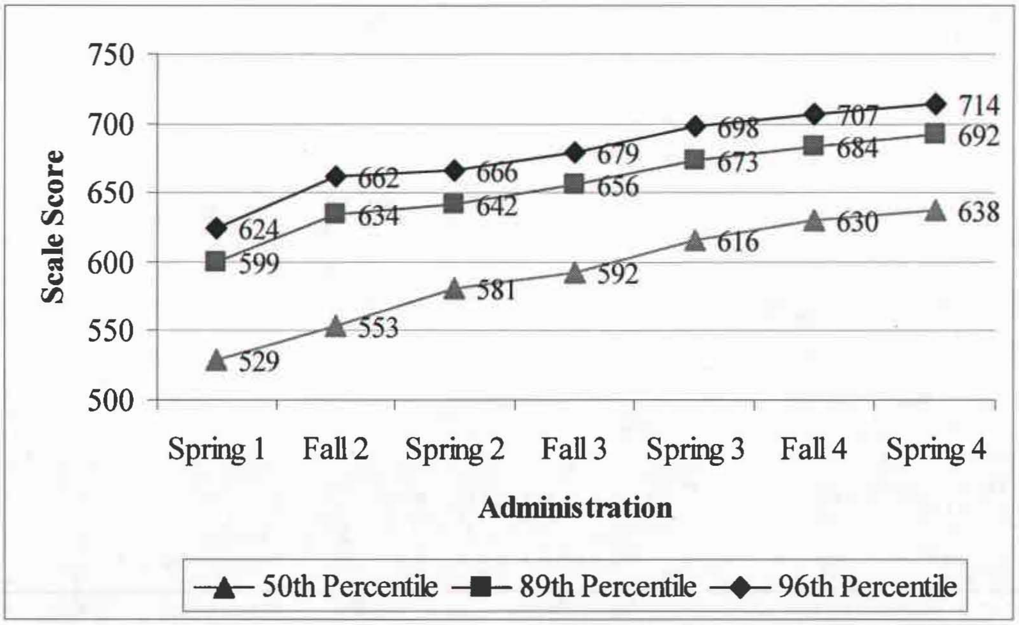

Figure 8. Changes in scale scores by test administration period: SAT-9 Total Reading.

[9] The testing window for the CTP III is September 1–February 28 for fall and March 1–August 31 for spring administrations (Educational Records Bureau, 2001, p. 6)

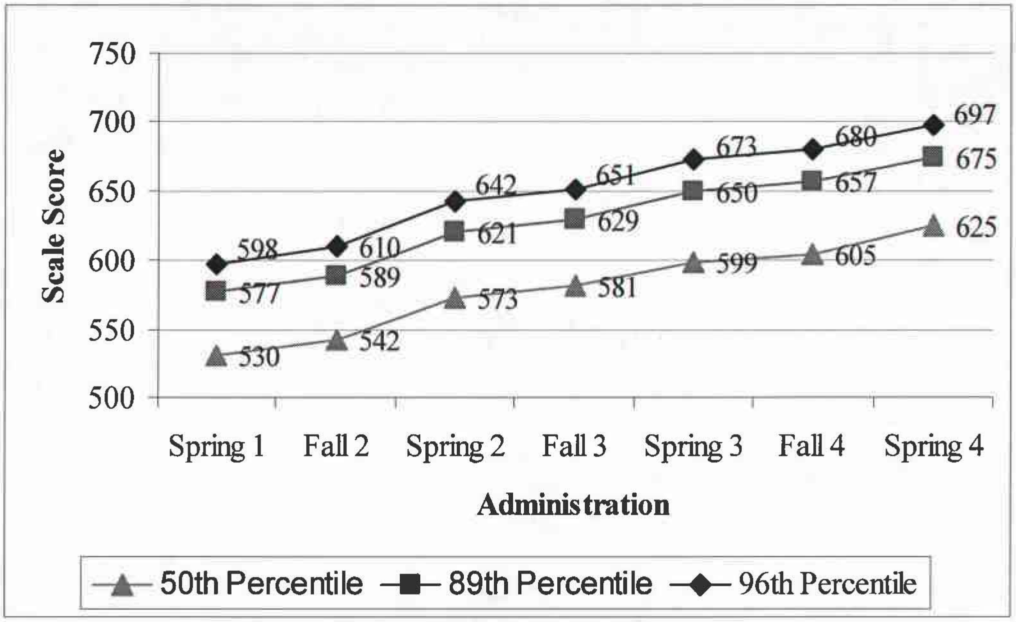

Figure 9. Changes in scale scores by test administration period: SAT-9 Total Mathematics.

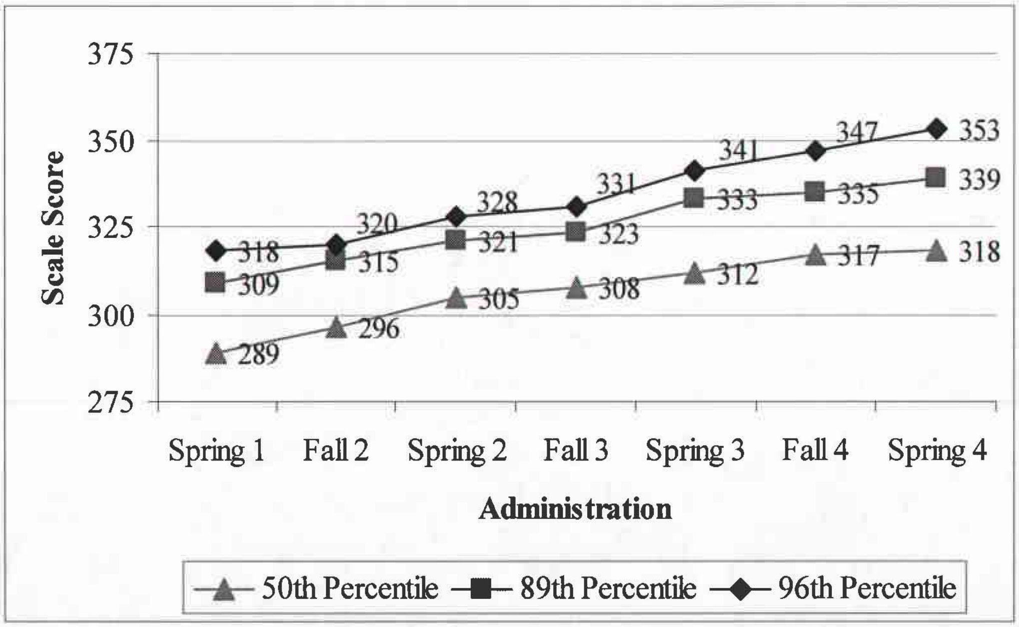

Figure 10. Changes in scale scores by test administration period: CTP III Reading Comprehension.

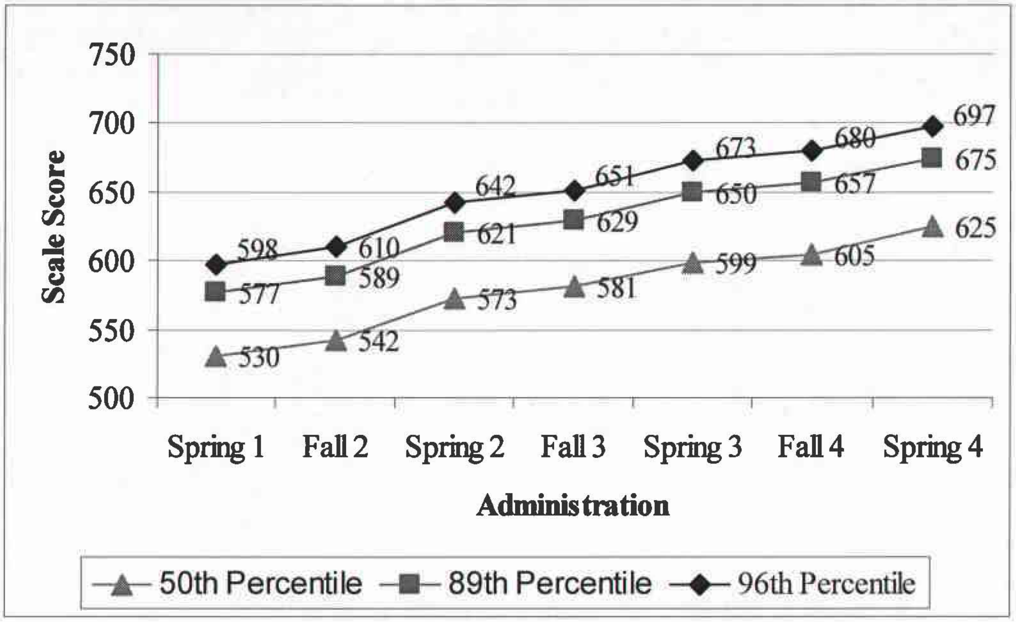

Figure 11. Changes in scale scores by test administration period: CTP III Mathematics.

Instability of Growth

As I noted in Chapters 2 and 3, prototypic learning curves relevant to the group of students in this study have not yet been developed. A limited, but available, proxy for the missing learning curves might be found in the comparison of the performance of the multiple norm groups for the CTP III and the SAT-9.

In Figures 12 and 13, I plotted the national percentile ranks of the mean scores for each of the alternative norm groups for the CTP III: the Independent School and Suburban School norm groups. Also included are the mean scores from this school (the school data are longitudinal).

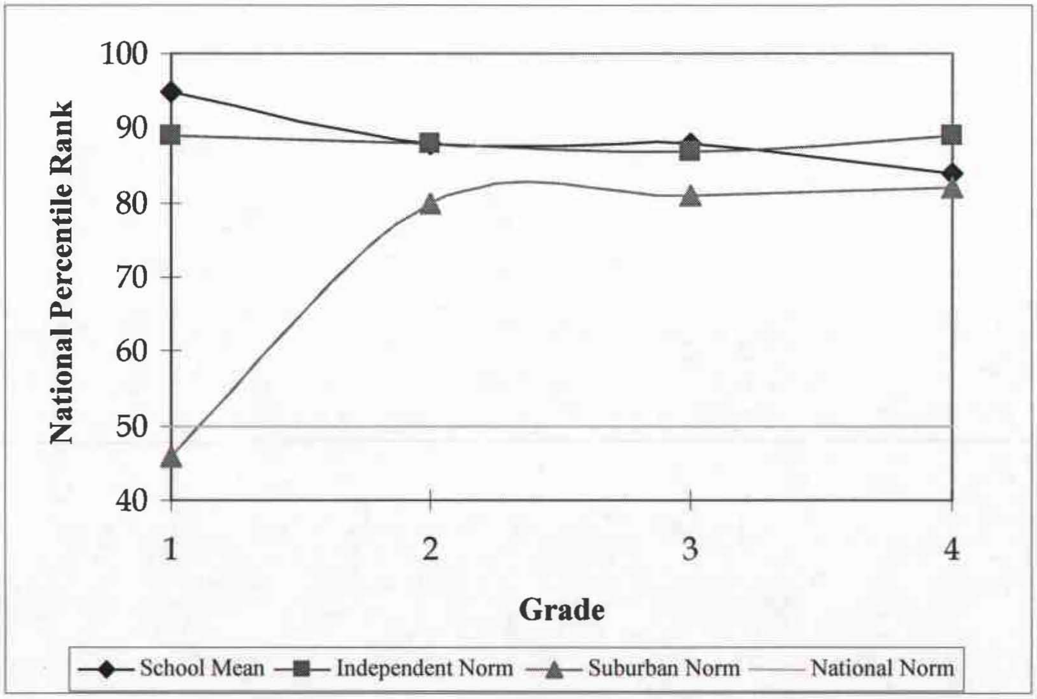

Figure 12. Comparison of alternative norm groups relative to the national norm: CTP III Reading Comprehension.

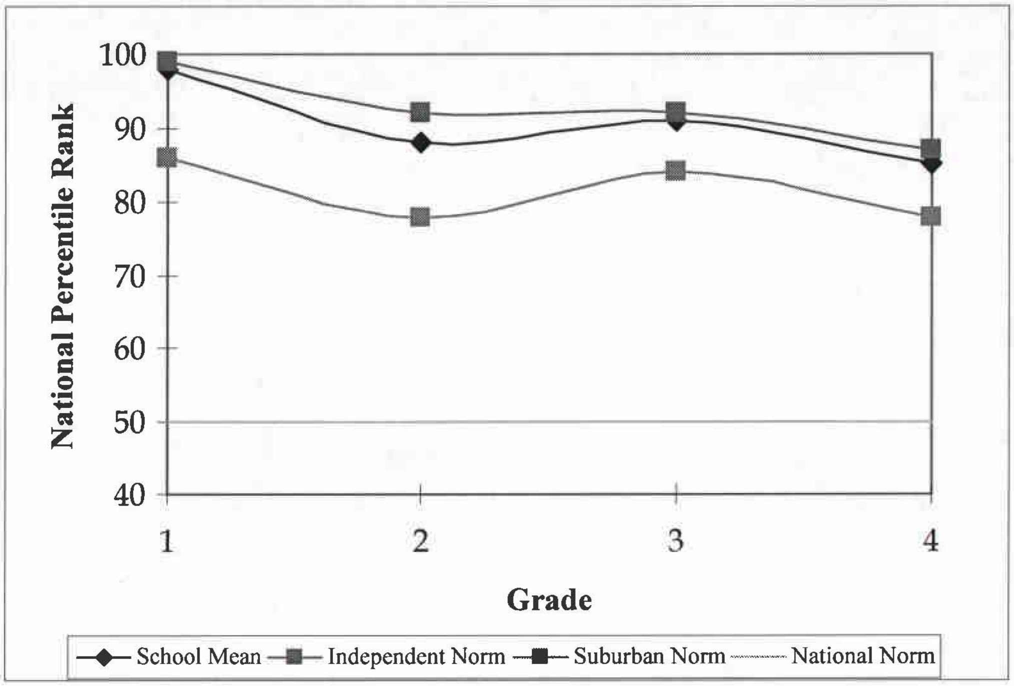

Figure 13. Comparison of alternative norm groups relative to the national norm: CTP III Mathematics.

This instability of the relative position of the different norm groups cannot be written off as simply an artifact of the CTP III. It is also evident in the Private School and High SES norms for the SAT-9. The comparisons of the performance of the three norm groups on Total Reading and Total Mathematics are shown in Figures 14 and 15, respectively.

The analysis of the performance of the norm groups relative to each other is based on cross-sectional, group data using scale scores designed to create a smooth growth curve over time. Yet, from the results, it is obvious that expectations of stability in normative achievement over time are, at best, questionable.

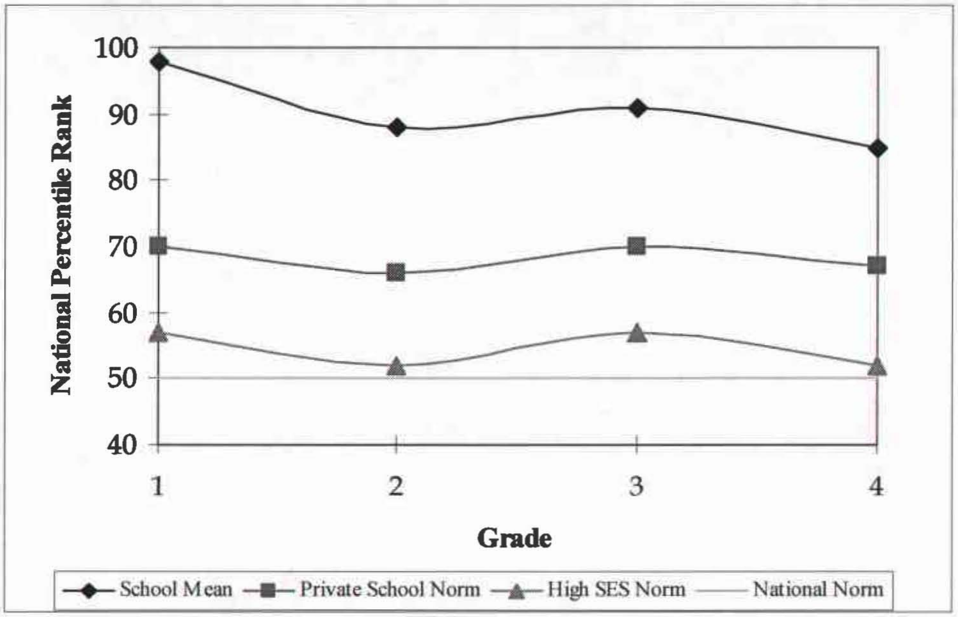

Figure 14. Comparison of alternative norm groups relative to the national norm: SAT-9 Total Reading.

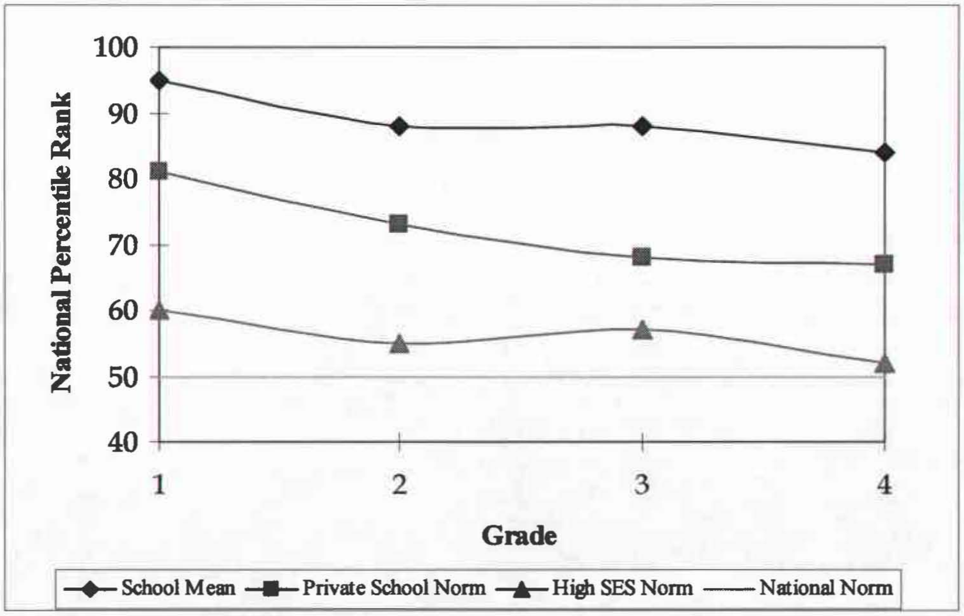

Figure 15. Comparison of alternative norm groups relative to the national norm: SAT-9 Total Mathematics.

CHAPTER 5. CONCLUSIONS

The Validity of Test-Based Inferences about Program Quality

This study is about the validity of the inferences we make about the effectiveness and quality of schools and educational programs on the basis of test scores. To draw valid inferences about the causes of observed changes requires posing and testing hypotheses and the elimination of reasonable competing hypotheses.

At the school whose data were used in this report, mean student test scores showed a persistent pattern of decline over the span of Grade 1 to Grade 4. This pattern was robust, occurring over all cohorts tested at all these grade levels and with two different achievement test batteries. The mean scores for the cohorts included in this study declined from the 95th to the 84th percentile in Reading and from the 98th to the 85th percentile in Mathematics.

Based on the available literature, I identified and explored four reasonable sources for the observed changes in scores all of which have little or no relationship to the effectiveness of the school. These factors are regression toward the mean, the influence of test ceilings, changes in the nature of reading and mathematics in the early elementary grades, and the inherent normative instability in student achievement.

Impact of Regression Toward the Mean

Regression toward the mean is a significant and often neglected factor in understanding changes in test scores. This phenomenon is ubiquitous and too infrequently considered as an explanation for observed changes in scores.

Regression toward the mean has two components: measurement error and imperfect correlations between measures. I used a combined approach to estimate the size of the expected regression effect in a way that I have not seen in the published literature. First, I used established methods to estimate true scores given known test reliabilities. I used these true score estimates to calculate expected scores given imperfect correlations between measures. The published reports I found took one approach or the other.

The expected score results were very comparable to the observed scores. This suggests that, although I had to use proxies for the correlations between the tests, statistical regression is a reasonable explanation for the changes in scores. The change is not, in fact, larger than we should expect. It may even be smaller than what might be expected.

Impact of Test Ceilings

When the scores are at the extremes of the distributions, item gradients are steep, reliability is low, and tests may have marked ceiling effects. These can lead even the best-intentioned observer to false conclusions about the efficacy, or lack thereof, of educational programs.

I provided evidence of the presence of ceilings for reading and mathematics at all grade levels. I also showed that, at Grade 1, nearly 50% of the students were at or near the ceiling for reading, and more than 75% were at or near the ceiling for mathematics. This suggests that while student achievement was indeed above-average, it was also outside the range at which the tests are effective. Thus, the apparent extraordinary level of achievement is likely to have been exaggerated. Put another way, their *true* ranks did not decline from the 95th or 98th percentiles because they were never that high—the percentile ranks were inflated by the test ceilings. The observed declines in percentile ranks therefore say more about the limitations of the test than the achievement of the students.

Impact of Changes in Test Content

The nature of reading and mathematics changes substantially during the early elementary grades. Thus, the content of the tests of these domains change as well.

While it is difficult to quantify the impact of changes in the content of the test, other researchers (e.g., Mills & Jackson, 1990; Storch and Whitehurst, 2002) have demonstrated that the changes in the nature of reading result in, at best, moderate relationships between tests of reading in preschool through Grade 1 and later reading achievement. There are multiple precursors to more mature reading, and the relative importance of each changes over time. Thus, the apparent precocity of the students at Grade 1 should not create expectations that their later achievement will continue to be as distinctive.

In this study, I demonstrated that there are substantial qualitative and quantitative changes in the content of the CTP III and SAT-9. These are likely to contribute to the same instability in the students' scores over time as was demonstrated in the work by Mills and Jackson (1990) and by Storch and Whitehurst (2002).

Impact of Instability of Growth

Finally, longitudinal research on the lack of stability in IQ and on the prediction of achievement in the early grades has demonstrated that individual performance relative to the norm group is inherently unstable, particularly during early childhood. However, there are likely to be prototypic patterns of growth, with the emphasis on the multiplicity of patterns (McCall et al., 1973).

Researchers working in the field of cognitive growth repeatedly point out that the tendency of the group mean is not an adequate substitute for the development of learning curves. To be most useful, such curves would need to be based on longitudinal studies of individuals versus group means (e.g., McCall et al., 1973; Willett, 1988; Wohlwill, 1980). Yet a set of growth curves that could serve as a point of reference to help us interpret the observed data has not been developed. Lacking this, our only fallback is the normative data from ability and achievement tests.

The practicalities of time and the economics of test development constrain the capacity of test developers to go beyond cross-sectional data as they develop the scale scores and norm tables for the achievement test batteries. While some might argue that the linking studies help bridge the levels of the test, these are still estimates from one point in time. The limited longitudinal research available demonstrates that these cross-sectional data cannot reveal the extent of the instability of the performance of individuals or differences in trajectories between groups of students and thus invite false expectations of stability.

This normative instability beyond the study population was demonstrated by a comparison of mean scores for the alternative norm groups. The patterns of scores for the alternative norm groups looks surprisingly similar to those observed at the school under study, given that the smaller number of students at the school could be expected to increase the variability of the scores over time. Thus, the wide spread expectation that students or groups of students will maintain their position relative to the norm group—however tempting to embrace—is based more on impression than on scientific evidence.

Summary

The most facile inference from test data that regresses over time is that the school or educational program is causing the observed changes. If the program is serving below-average students and their scores improve, few are likely to dispute the program's effectiveness. If scores start high and decline, the potential for concern on the part of all stakeholders is great. I have demonstrated the potential of four alternative explanations to account for all or part of the changes observed in the test scores. At this time, we do not have an adequate knowledge base to support causal inferences about the sources of the observed changes. Thus, we cannot accept at face value the theory that the school caused the changes.

I do not suggest that the regression in test scores observed at this school is inevitable. I have first-hand knowledge of other schools (who shall also remain anonymous) that enroll high-achieving students where the test scores do not regress over the elementary grades. In these schools, stability of test scores is a valued outcome and the faculty and school administration work together to achieve this. This is not to suggest that the faculty and staff at these schools are engaged in inappropriate test practices; rather, that for above-average students, regression toward the mean is not unexpected and that to avoid this requires attention to the alignment of curriculum and instruction with the test content. It should also be noted that the students at these schools without regression effects are even more exceptional in their achievement than are the students at the school in this study. Therefore, there is the real possibility that the performance of the students at these schools also varies in its relationship to the norm group, and that this is not apparent because the performance of the students at these schools remains above the test ceiling.

Transportability

It is my aspiration that this work has value beyond the immediate context of the study. Therefore, I have made efforts to ensure the methods and findings are applicable beyond this setting.

Replicable Methods

The approach used in this study was deliberately chosen to be replicable in the context of testing and evaluation offices outside the major test development power houses. As noted in Chapter 1, I use the materials and statistical analyses that, in my opinion, should be part of the "toolkit" of any professional with responsibility for student assessment and program evaluation at or above the school district level.

All reference information in this study comes from the technical reports and norms books routinely published by the commercial test developers. The data analyses were conducted using the information from the standard electronic files of test results from the scoring services (with the exception of the raw scores for the CTP III, which were added to the file after it was received from the scoring service). The statistical methods used are primarily descriptive statistics that can all be calculated with relatively unsophisticated software, such as Microsoft Excel.

Thus, my approach can be easily replicated in testing and evaluation offices across the country. The inquiry and statistical methods can be used by schools and programs that find themselves on either end of the achievement continuum.

Generalizable Findings

Wainer (personal communication, 1999) indicated that once the staff at Educational Testing Service was sensitized to the issue of the regression in test scores by my repeated inquiries, they realized that a number of CTP III user schools had the same pattern of regression in their test scores. (The exact number and names of these schools were not disclosed to me to protect their privacy.)

While the comparability of these other schools to the one in this study is unknown, these findings would generalize to those that with comparable student populations. In addition, the pattern of regression evident in the findings would also generalize to schools and educational programs serving below-average student populations.

Contributions to the Theory of Regression Toward the Mean

This study presents a relatively rare opportunity to empirically analyze the regression effect among a high-achieving population of students. Thus, it provides a test site for the application of the theory as developed by Galton, Campbell and Stanley (1963), Furby (1973), Nesselroade et al. (1980), and others.

As shown in this study, if statistical regression is present, the use of the test reliability as a proxy for the correlation between tests in the population yields apparently reasonable results. Use of the correlation coefficient for the sample—even after correction for attenuation—yields results that are not congruent with "common sense" (i.e., the prediction that the scores for these above-average children would regress all the way to the mean of the general population).

In the exploration of the factors that could reasonably contribute to the explanation of the regression in the scores, I have identified three factors beyond the purely statistical sources of regression that may contribute to the regression effect.

Test ceilings (and floors) may contribute to the regression effect in two ways. First, they decrease the correlation between tests because significant variance in achievement goes undetected in one of the measurements. Second, they return inaccurate normative results. If we measure achievement on two occasions and on both of these occasions the level of achievement is above the ceiling or below the floor of the assessments, no change will be evident, even though it is real. Second, when one result is outside the range measured by the test and the other is within it, this may lead to apparent change where little or no real change may have actually occurred.

Changes in test content—both quantitative and qualitative—essentially change the construct being measured. The relative weights of component aptitudes or developed abilities vary across levels of the tests within an achievement battery or between tests when using indicators from different sources. When this occurs, we can expect scores in the same domain to be correlated but not identical.

Finally, the relative position of individuals may simply change. This could be due to a variety of factors including the effect of education, changing interests and motivations, differences in the timing of the acquisition of skills, the fading of the effects of early acceleration, or other life circumstances.

Limitations of the Current Study

This study has several limitations. The data used had severe truncation and skewness at some points (e.g., Grade 1). The departure from normality in this study may have produced some bias in the estimates reported by an unknown direction; however, most statistics have been shown to be robust in the face of such violations.

The change in tests from the CTP III to the SAT-9 has the potential to be a limitation but, as shown in Chapter 3 and in additional data in the Appendix, the impact of the change from the CTP III to the SAT-9 was no greater than the typical variation between cohorts.

This study included three cohorts of students, which decreased cohort effects and increased the power of the statistics. However, more cohorts of data and data from more schools could help ensure the robustness of the findings across time and educational programs.

I also used cross-sectional normative data from multiple norm groups as proxies for growth curves since no empirically validated curves are available. Empirical data on prototypic learning curves and expected coefficients of stability would have been particularly informative.

Future Directions

We need increased awareness of and attention to factors that are exogenous to educational programs that affect test scores *before* drawing inferences about the effectiveness of curricula, schools, and special programs and services. This need extends to practitioners and in the published educational assessment and evaluation literature.

The likelihood of regression toward the mean and the instability of normative performance, present a dilemma for the evaluation of educational programs serving above- or below-average students: how can we identify program effects with any confidence?

Although this study cannot provide the answer to that question, it does provide evidence of the dangers of relying on purely descriptive studies, suggesting a need to include methods that provide stronger support for evaluative inferences. At a minimum, evidence to assess alternative hypotheses needs to be gathered and evaluated. Methods that help us do this include, but are not limited to, experimental, quasi-experimental, and those that use a process of "detection" that accumulates evidence of likely causal relationships (e.g., when well implemented the CIPP and the success case methods have this potential). Also important in the search for improved evaluative inferences would be a larger knowledge base and benchmark data about the outcomes of educational programs and services serving similar student populations.

A potentially fruitful strategy would be to shift from a normative basis for judging school effectiveness to the measurement of student achievement of standards and/or instructionally significant performance benchmarks. Such an approach would not rely on poorly grounded assumptions of normative stability. In addition, it has the added benefit of transparency in terms of what students know and can do which is *not* an attribute of norm-referenced assessments.

Regardless of whether testing is conducted using norm-referenced or standards-referenced approaches, when achievement is represented in a single developmental scale, we need much more work to increase the integrity of the scales. The creators and marketers of these scales tell us that we can use them to monitor growth across levels of tests. The results of this and other studies show that we take the test developers at their word at our own risk. More importantly, we risk harm to children, teachers, and schools. In the face of the states' rush to comply with the requirements of No Child Left Behind and in an environment where the benefits of "value-added" studies are being promoted, inadequate vertical equating is particularly perilous.

Finally, further empirical work with data from other schools and from other tests is needed to evaluate the reasonableness of using test reliability coefficients as proxies for the correlations between tests, as was done in this study. Also, the availability of prototypic learning curves by cognitive domain, age, and other relevant characteristics would be a great benefit to program administrators, educational program evaluators, the developers of education assessments, and a host of other potential users.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Amrein-Beardsley, A., & Berliner, D. C. (August 4, 2003). *Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine.* Retrieved July 20, 2004, from http://epaa.asu.edu/epaa/v11n25/

Beath, K. J., & Dobson, A. J. (1999). Regression to the mean for nonnormal populations. *Biometrika, 78,* 431–435.

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26,* 155–166.

Bracken, B. A. (1990). Maximizing construct relevant assessment: The optimal preschool testing situation. In B. A. Bracken (Ed.), *The Psychoeducational Assessment of Preschool Children* (3rd ed.). Boston: Allyn and Bacon.

Brown, R. T., & Jackson, L. A. (1992). Ex-Huming an old issue. *Journal of School Psychology, 30,* 215–221.

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practice, 3*(6), 15–16.

Callahan, C. M. (1992). Determining the effectiveness of educational services: Assessment issues: Challenges in gifted education: Developing potential and investing in knowledge for the 21st Century. (ERIC Document Reproduction Service No. ED344416)

Camilli, G., & Bulkley, K. (2001). *Critique of "An evaluation of the Florida A-Plus Accountability and School Choice Program."* Retrieved March 4, 2001, from http://epaa.asu.edu/epaa/v9n7/

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Clemons, T. (2000). A look at the inheritance of height using regression toward the mean. *Human Biology, 72,* 447–454.

Cole, P. G. (1998). Regression to the mean: A ubiquitous concept that explains much about the relationship among data on gifted individuals. *Education Research and Perspectives, 25*(2), 83–98.

Doran, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43–68.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised: Manual for forms L and M.* Circle Pines, MN: American Guidance Service.

Educational Records Bureau. (2001). *ERB testing programs and services: 2001-2002 catalog*. New York: Author.

Educational Testing Service. (1993). *Instructional objectives manual: Relating CTP III to curriculum*. New York: Educational Records Bureau.

Educational Testing Service. (1995). *Comprehensive Testing Program III: Technical report*. New York: Educational Records Bureau.

Educational Testing Service. (1997). *Comprehensive Testing Program III: Spring norms booklet*. New York: Educational Records Bureau.

Educational Testing Service. (2001). *Comprehensive Testing Program III: Fall norms booklet*. New York: Educational Records Bureau.

Educational Testing Service. (2004). *Comprehensive Testing Program 4: Technical report*. New York: Educational Records Bureau.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology, 8*, 172–179.

Furlong, M. J., & Feldman, M. G. (1992). Can ability-achievement regression to the mean account for MDT discretionary decisions. *Psychology in the Schools, 29*, 205–212.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (2nd ed.). Boston: Allyn and Bacon.

Goal 1 Early Childhood Assessments Resource Group. (1998). *Principles and recommendations for early childhood assessments*. L. Shepard, S. L. Kagan, & E. Wurtz (Eds.) Washington, DC: National Educational Goals Panel. Retrieved November 1, 1998 from www.negp.gov/Reports/prinrec.pdf

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test Series, Ninth edition: Test coordinator's handbook.* San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1997a). *Stanford Achievement Test Series, Ninth Edition: Fall norms book.* San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1997b). *Stanford Achievement Test Series, Ninth Edition: Spring norms book.* San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1997c). *Stanford Achievement Test Series, Ninth Edition: Technical data report.* San Antonio, TX: Author.

Hartman, D. P., & George, T. P. (1999). Design, measurement, and analysis in developmental research. In M. H. Bornstein & M. E. Lamb (Eds.), *Developmental psychology: An advanced textbook* (4th ed., pp. 125–195). Mahwah, NJ: Lawrence Erlbaum Associates.

Hills, J. R. (1993). Regression effects in educational measurement. *Educational Measurement: Issues and Practice, 12*(3), 31–34.

Hoover, H. D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice, 3*(6), 8–14.

Hoover, H. D. (1984b). Rejoinder to Burket. *Educational Measurement: Issues and Practice, 3*(6), 16–18.

Huck, S. W., Cormier, W. H., & Bounds, J., William G. (1974). *Reading statistics and research.* New York, NY: Harper & Row.

Ingebo, G. S., & Doherty, V. W. (March 16, 1982). *An empirical study of the effect of regression to the mean.* Portland, OR: Evaluation Department, Portland Public Schools.

Kelly, M. S., & Peverly, S. T. (1992). Identifying bright kindergartners at risk for learning difficulties: predictive validity of a kindergarten screening tool. *Journal of School Psychology, 30*, 245–258.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*, 3–14.

Kruger, J., Savitsky, K., & Gilovich, T. (1999). Superstition and the regression effect. *Skeptical Inquirer* (March/April), 24–29.

La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: a meta-analytic review. *Review of Educational Research, 70*(4), 443–484.

Lindquist, G. T. (1982). Preschool screening as a means of predicting later reading achievement. *Journal of Learning Disabilities, 15*(6), 331–332.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83–102.

McCall, R. B., Appelbaum, M. I., & Hogarty, P. S. (1973). Developmental changes in mental performance. *Monographs of the Society for Research in Child Development, 42*(3, Whole No. 150).

Mee, R. W., & Chua, T. C. (1991). Regression toward the mean and the paired sample *t* test. *The American Statistician, 45*(1), 39-42.

Miller-Whitehead, M. (1999, April). *Tennessee TCAP science scale scores: Implications for continuous improvement and educational reform or is it possible to beat the odds?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec. (ERIC Document Reproduction Service No. ED430030)

Mills, J. R., & Jackson, N. E. (1990). Predictive significance of early giftedness: The case of precocious reading. *Journal of Educational Psychology, 82*, 410–419.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88*, 622-637.

Nitko, A. J. (1983). *Educational tests and measurement: An introduction.* New York: Harcourt Brace Jovanovich.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 Retrieved July 4, 2004 from http://www.ed.gove/policy/elsec/leg/esea02/index.html

Nunnally, J. C. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill.

Paris, S. G., & Lindauer, B. K. (1982). The development of cognitive skills during childhood. In B. B. Wolman & G. Stricker (Eds.), *Handbook of developmental psychology* (pp. 333–349). Englewood Cliffs, NJ: Prentice-Hall.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221–262). New York, NY: American Council on Education and Macmillan.

Popham, W. J. (2000). *Testing! Testing! What every parent should know about school tests* (Second ed.). Boston: Allyn and Bacon.

Rock, D. A., Pollack, J., & Quinn, P. (1995). *Psychometric report for the NELS:88 base year through second follow up.* Retrieved January 28, 2001, from http://nces.ed.gov/pubs95/95382.pdf

Schmidt, S., & Perino, J. (1985). Kindergarten screening results as predictors of academic achievement, potential, and placement in second grade. *Psychology in the Schools, 22*, 146–151.

Shadish, W. R., Cook, T. D., & Leviton, L. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Stanley, J. C. (1990). Leta Hollingworth's contributions to above-level testing of the gifted. *Roeper Review, 12*(3), 166–171.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38*(6), 934–947.

Tallmadge, G. K. (1988, April). *How and how much do biases resulting from regression to the mean affect outcome evaluations?* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED294888)

Tramontana, M. G., Hooper, S. R., & Selzer, S. C. (1988). Research on the preschool prediction of later academic achievement: A review. *Developmental Review, 8*, 89–146.

U. S. Department of Education. (2001). *Elementary & secondary education: Executive summary.* Retrieved July 17, 2004, 2004, from http://www.ed.gov/nclb/overview/intro/execsumm.pdf

Venezky, R. L., Bristow, P. S., & Sabatini, J. P. (1994). *Measuring gain in adult literacy programs. National Center on Adult Literacy: Technical report TR93–12.* (ERIC Document Reproduction Service No. ED 369978)

Wainer, H. (1999). Is the Akebono School failing its best students? A Hawaiian adventure in regression. *Educational Measurement: Issues and Practice, 18*(3), 26–31, 35.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15: 1988–89, pp. 345–422). Washington, DC: American Educational Research Association.

Williamson, G. L., Appelbaum, M. I., & Epanchin, A. (1991). Longitudinal analyses of academic achievement. *Journal of Educational Measurement, 28*, 61-76.

Wohlwill, J. F. (1980). Cognitive development in childhood. In O. G. Brim Jr. & J. Kagan (Eds.), *Constancy and change in human development* (pp. 359–444). Cambridge, MA: Harvard University Press.

Yang, W.-L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*, 33–41.

Yen, W. M. (1983). Tau equivalence and equipercentile equating. *Psychometrika, 48*, 353–370.

Yen, W. M. (1985). Increasing item complexity: A possible cause for scale shrinkage for unidimensional item response theory. *Psychometrika, 50*, 399–410.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299–325.

APPENDIX: SUPPLEMENTARY TABLES

Table A1

Comparison of NCEs Obtained Using the CTP III and SAT-9 for Two Consecutive Years: Reading

Grade	Test	*N*	*Mean*	*sd*	*t*	*p*	*d*
1	CTP III	60	88.2	15.8			
	SAT-9	80	87.5	11.5			
	Difference		0.7	13.5[a]	0.3	.76	0.06
2	CTP III	57	78.0	15.6			
	SAT-9	60	75.0	12.3			
	Difference		3.0	14.0	1.1	.26	0.21
3	CTP III	55	72.5	12.4			
	SAT-9	57	74.6	12.8			
	Difference		-2.1	12.6	-0.9	.38	0.17
4	CTP III	132	71.0	14.8			
	SAT-9	55	64.1	12.2			
	Difference		6.9	14.1	3.0	<.01	0.49

[a] Variances unequal, used Satterthwaite adjustment for unequal variances.

Table A2

Comparison of NCEs Obtained Using the CTP III and SAT-9 for Two Consecutive Years: Mathematics

Grade	Test	*N*	*Mean*	*sd*	*t*	*p*	*d*
1	CTP III	60	96.3	7.8			
	SAT-9	80	87.6	13.5			
	Difference		8.7	11.4[a]	4.8	<.01	0.76
2	CTP III	57	76.2	16.8			
	SAT-9	60	80.7	13.1			
	Difference		-4.6	15.0	-1.6	.10	0.30
3	CTP III	55	70.9	14.7			
	SAT-9	57	79.0	15.6			
	Difference		-8.1	15.1	-2.8	<.01	0.54
4	CTP III	132	68.9	14.0			
	SAT-9	55	67.4	12.4			
	Difference		1.5	13.6	0.7	.49	0.11

[a] *Variances unequal, used Satterthwaite adjustment for unequal variances.*

Table A3

Comparison of NCEs Obtained Using the CTP III for Two Consecutive Years: Reading Comprehension

Grade	Administration	*N*	*Mean*	*sd*	*t*	*p*	*d*
1	Fall 95	60	88.2	15.8			
	Fall 96	80	87.5	11.5			
	Difference		0.7	13.5[a]	0.3	.76	0.05
2	Fall 95	57	78.0	15.6			
	Fall 96	60	75.0	12.3			
	Difference		3.0	14.0	1.1	.26	0.21
3	Fall 95	55	72.5	12.4			
	Fall 96	57	74.6	12.8			
	Difference		2.1	12.6	0.88	.38	0.17
4	Fall 95	132	71.0	14.8			
	Fall 96	55	64.1	12.2			
	Difference		6.9	14.1	3.0	<.01	0.49

[a]Variances unequal, used Satterthwaite adjustment for unequal variances.

Table A4

Comparison of NCEs Obtained Using the CTP III for Two Consecutive Years: Mathematics

Grade	Administration	*N*	*Mean*	*sd*	*t*	*p*	*d*
1	Fall 95	60	96.3	7.9			
	Fall 96	80	87.6	13.5			
	Difference		8.7	11.4[a]	4.8	<.01	0.76
2	Fall 95	57	76.2	16.8			
	Fall 96	60	80.7	13.1			
	Difference		4.5	15.0	1.64	.10	0.30
3	Fall 95	55	70.9	14.7			
	Fall 96	57	79.0	15.6			
	Difference		8.1	15.1	2.83	<.01	0.54
4	Fall 95	132	68.9	14.0			
	Fall 96	55	67.4	12.4			
	Difference		1.5	13.6	0.7	.49	0.11

[a]*Variances unequal, used Satterthwaite adjustment for unequal variances.*

Table A5

Comparison of NCEs Obtained Using the SAT-9 for Two Consecutive Years: Total Reading

Grade	Administration	*N*	*Mean*	*sd*	*t*	*p*	*d*
2	Fall 97	80	75.5	13.1			
	Fall 98	79	77.7	12.0			
	Difference		2.2	12.5	1.1	.27	0.18
3	Fall 97	80	74.1	13.4			
	Fall 98	80	77.6	12.6			
	Difference		3.5	13.0	1.7	.09	0.27
4	Fall 97	132	67.2	12.3			
	Fall 98	144	73.8	13.3			
	Difference		6.6	12.8	4.3	<.01	0.52

Note. Grade 1 data available only for fall 1997.

Table A6

Comparison of NCEs Obtained Using theSAT-9 for Two Consecutive Years: Total Mathematics

Grade	Administration	*N*	*Mean*	*sd*	*t*	*p*	*d*
2	Fall 97	80	80.8	13.0			
	Fall 98	80	80.6	13.2			
	Difference		0.2	13.1	0.11	.91	0.02
3	Fall 97	80	78.3	15.1			
	Fall 98	80	83.5	12.8			
	Difference		5.2	14.0	2.4	.02	0.37
4	Fall 97	132	70.3	13.0			
	Fall 98	144	74.9	14.1			
	Difference		4.6	13.6	2.8	<.01	0.34

Note. Grade 1 data available only for fall 1997.

Table A7

Item gradients around mean raw scores: Reading

	CTP III				SAT-9			
Grade	RS	SS	SS Change	1/3 *sd*	RS	SS	SS Change	1/3 *sd*
	33	314			--	--	--	--
1	32	311	3	5.1	--	--	--	--
	31	309	2		--	--	--	--
	33	318			102	635		
2	32	317	1	4.7	101	632	3	13.8
	31	316	1		100	630	2	
	31	330			75	673		
3	30	328	2	5.1	74	668	5	14.4
	29	327	1		73	664	4	
	--	--	--	--	73	678		
4	--	--	--	--	72	674	4	14.7
	--	--	--	--	71	671	3	

Note. Raw score to scale score conversions are from test norms (Educational Testing Service, 1997; Harcourt Brace Educational Measurement, 1997b). Standard deviations of the scaled scores are from test statistics published by the test developers (Educational Testing Service, 1997; Harcourt Brace Educational Measurement, 1997a, 1997c).

Table A8

Item gradients around mean raw scores: Mathematics

	CTP III				SAT-9			
Grade	RS	SS	SS Change	1/3 *sd*	RS	SS	SS Change	1/3 *sd*
	34	261			--	--	--	--
1	33	258	3	5.5	--	--	--	--
	32	256	2		--	--	--	--
	31	269			67	628		
2	30	267	2	6.5	66	623	5	12.6
	29	265	2		65	618	5	
	59	294			68	658		
3	58	292	2	6.9	67	653	5	13.3
	57	291	1		66	648	5	
	--	--	--	--	66	667		
4	--	--	--	--	65	664	3	12.8
	--	--	--	--	64	660	4	

Note. Raw score to scale score conversions are from test norms (Educational Testing Service, 1997; Harcourt Brace Educational Measurement, 1997b). Standard deviations of the scaled scores are from test statistics published by the test developers (Educational Testing Service, 1997; Harcourt Brace Educational Measurement, 1997a, 1997c).