A VALIDATION STUDY

OF ELEMENTARY SCIENCE ISTEP+ SCORES

Glenn Simonelli

Submitted to the faculty of the School of Education

in partial fulfillment of the requirements

for the degree

Doctor of Education

in the Department of Curriculum and Instruction.

Indiana University

October, 2006

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Education.

_____

Dr. Gary Ingersoll, Ph.D.

_____

Dr. Valarie Akerson, Ph.D.

_____

Dr. Richard, H. Durisen, Ph.D.

September 13, 2006

A VALIDATION STUDY

OF ELEMENTARY SCIENCE ISTEP+ SCORES

Glenn Simonelli

Abstract: The Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) was designed to assess student mastery of key educational goals. The 5th grade ISTEP+ Science Test (5-GIST) is part of the ISTEP+ testing regime. The Indiana Academic Standards were developed to guide instruction in the state, and questions on the ISTEP+ were aligned with these standards. Since its inception, the use of the ISTEP+ exam has been changed to comply with the dictates of both Indiana Public Law 221 and the national No Child Left Behind act. With these modifications, the purpose of the tests has shifted from assessment of individual student academic progress to evaluation of the quality of the educational institution administering the tests. The validity of this use has never been established. The purpose of this study is to assess the validity of the 5-GIST as an instrument for assessing and forming judgments about the quality of science instruction in a particular school. ISTEP+ scores of 2 cohorts of students in a Midwestern school district were converted into Z-scores and tracked from 3rd to 5th grade. A regression line was established to account for the general aptitude and the socio-economic status (SES) of the students. Examining the residuals of the 5-GIST scores revealed that between 57% and 60% of the variance in the scores can be attributed to the general aptitude and SES of the students, leaving between 40% and 43% that can be interpreted as reflecting the effect of the school on student learning.

DEDICATION

Many people have guided and encouraged me over the past four years, but two people in particular stand out. First, my loving wife Michele has unfailingly stood beside me, constantly going out of her way so that I could have the time and space necessary to fulfill all the requirements of this program. She has willingly sacrificed her time and comfort to allow my to pursue me goals, provided valuable feedback and advice, has cheered me when I was discouraged, and never failed to believe in me, even when I began to doubt myself. Her love, compassion and dedication mean more to me than any degree, and any success I may achieve because of it is our success.

Second, Dr. Richard H. Durisen has been a faithful friend and supporter for many years. He has written a grant that has made it possible for me to return to school. He has always looked for ways to promote me to colleagues, for opportunities for me to expand professionally—often going out of his way to provide many of those opportunities himself—and he has willingly given me much of his valuable time to work with and help me on various projects.

For their love, friendship and support over the years, this manuscript is dedicated to both of them.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the guidance and support of Dr. Gary Ingersoll in the design, implementation and interpretation of this study. Through his patient instruction and frequent invaluable feedback this research project was given direction and focus. Dr. Ingersoll has often gone the extra mile in helping me with this project, and this final report reflects his substantial input in both form and substance.

Additionally, the author is grateful for the encouragement received from Valarie Akerson and Richard Durisen over the life of this project. Completion of this work was made much easier because of their enthusiastic support.

This study would not have been possible without the cooperation of Michael Shipmen, Kim Metcalf and Jay Leslie of the Monroe County Community School Corporation. The author is grateful to them for providing access to the data referenced herein.

TABLE OF CONTENTS

PAGE

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS. iv

LIST OF FIGURES. ix

LIST OF TABLES. x

CHAPTER

I STATEMENT OF PROBLEM AND LITERATURE REVIEW 1

Evolution of ISTEP test. 1
Policy Applications of ISTEP+. 3
Indiana Public Law 221. 3
No Child Left Behind 4
Appropriate Use of the 5th Grade ISTEP+ Science Test. 5
Validity Issues. 6
Assumptions 9
The 5th Grade Science ISTEP+ Test 10

II METHODOLOGY. 16

Sample 16
Test Scores 18
Checking for Convergent and Discriminant Validation. 19
Checking Stability of Validity across Cohorts. 21
Evidence for Instructional Impact 21

III RESULTS 23

Convergent and Discriminant Validity 23
Ethnicity as a Moderating Variable. 26
Socioeconomic Status as a Moderating Variable 31
Mobility as a Moderating Variable 34
Additional Demographic Variables. 38
Significance of Covariance 40
Regression Lines 42

CHAPTER PAGE
IV DISCUSSION 48

Convergent and Divergent Validity Matrices. 48
Regression Lines for Ethnicity. 49
Regression Lines for SES. 49
Regression Lines for School Stability 50
T-test of Additional Demographic Variables. 51
Impact of $Science_5$ Regression. 51

V CONCLUSIONS, LIMITATIONS, FOLLOW UP AND RECOMMENDATIONS. 54

Limitations. 56
Follow Up 57
Recommendations 57

REFERENCES 59

LIST OF FIGURES

FIGURE		PAGE
1	Mean ISTEP Z-Score Change for Cohort 1 Students Who Took All 3rd, 5th and 6th Grade ISTEP+ Tests in the Same School (Including 95% Confidence Intervals)	13
2	Mean ISTEP Z-Score Change for Cohort 2 Students Who Took Both 3rd and 5th Grade ISTEP+ Tests in the Same School (with 95% confidence intervals).	14
3	Regression Lines for Ethnicity, Cohort 1.	28
4	Regression Lines for Ethnicity, Cohort 2.	30
5	Regression Lines for SES, Cohort 1.	32
6	Regression Lines for SES, Cohort 2.	33
7	Regression Lines for School Stability, Cohort 1	35
8	Regression Lines for School Stability, Cohort 2	37
9	Mean $Science_5$ Residuals Controlled for $English_3$, $Math_3$ and SES According to School	45
10	Estimated Marginal Means of Standardized Residuals According to School.	46
11	Mean $Science_5$ Z-scores According to School Unadjusted for SES or Aptitude.	53

LIST OF TABLES

TABLE		PAGE
1	Tests Taken by the Two Cohorts Whose Scores Are Examined in this Study.	16
2	Number of Students in Different Descriptive Categories.	18
3	Convergent and Discriminant Validity Matrix, Cohort 1.	20
4	Convergent and Discriminant Validity Matrix, Cohort 2.	21
5	Stability of Correlations over Time	21
6	Convergent-Discriminant Validity Correlations, Cohort 1.	24
7	Convergent-Discriminant Validity Correlations, Cohort 2.	25
8	Squared Correlation Coefficients Reflecting General Achievement. . . .	26
9	Pairwise T-tests for Slopes and Intercepts of Ethnic Groups, Cohort 1. .	29
10	Pairwise T-tests for Slopes and Intercepts of Ethnic Groups, Cohort 2. .	31
11	Pairwise T-tests for Slopes and Intercepts of SES, Cohort 1.	32
12	Pairwise T-tests for Slopes and Intercepts of SES, Cohort 2.	34
13	Pairwise T-tests for Slopes and Intercepts of School Stability, Cohort 1.	36
14	Pairwise T-tests for Slopes and Intercepts of School Stability, Cohort 2.	38
15	Regression Lines of Other Demographic Variables	39
16	Pairwise T-Tests for Slopes and Intercepts of Other Demographic Variables.	40
17	Adjusted R^2 and Significance of Demographic Variables	42
18	Mean Standardized Residuals According to School.	44
19	Tests of Between-Subjects Effects for Residuals	46

CHAPTER I

STATEMENT OF PROBLEM AND LITERATURE REVIEW

All public schools in Indiana are required to use standardized testing to assess academic progress and draw inferences about instructional quality. A premise of the testing program is that superior test scores, in the context of mitigating variables, can be interpreted as evidence of superior instruction. In addition, schools exhibiting superior scores could be scrutinized to identify factors that contribute to their superiority. These factors could then be replicated by other schools seeking to improve the quality of their instruction. However, before standardized test scores can be used to characterize instructional practices as superior or inferior, policy makers must be certain that these scores accurately reflect the instructional quality of a particular school and not other non-school-based variables.

The purpose of this study is to examine a source of evidence for the validity of the 5th Grade ISTEP+ Science Test (5-GIST). The study examines the presumed construct validity of the 5-GIST using procedures reflective of convergent and divergent validity as described by Campbell and Fiske (1959).

Evolution of ISTEP

According to documents on the State of Indiana Department of Education (DOE) web site,[1] the Indiana Statewide Testing for Educational Progress (ISTEP) was first given to students in 1st, 2nd 3rd, 6th, 8th, 9th and 11th grade during the 1987-8 school year. ISTEP was originally constructed as a norm-referenced test[2], but was changed to a criterion-

[1] http://www.doe.state.in.us/istep/2004/pdf/progman2004.pdf
[2] Pass/fail determinations were made according to a student's achievement position relative to the population of students taking the test.

referenced model[3] shortly afterward. Students in each of the grades listed above were assessed in language, mathematics and applied writing. In 1993 the 1st and 11th grade tests were dropped, as was the applied writing portion of the test for the 5 remaining grades assessed. For four years the test was limited to forced-response questions in language and mathematics.

During the 1996-7 school year the testing regime was significantly modified. Applied writing was reinserted into the test, and an applied mathematics section was added. Open-ended responses were once again part of the test, and the test was renamed the ISTEP+. ISTEP+ was to be administered in 3rd, 6th, 8th and 10th grades. During the same year the testing date was changed from the spring to the fall to allow teachers to take advantage of test results in identifying weaknesses in student learning and in planning appropriate corrective measures. Each school was placed in a cohort of schools that were of similar criteria such as size of student population and the percentage of students receiving free lunches. Schools could assess their performance by comparing the percentage of students achieving a passing score in each subject with the results of their cohort schools.

The testing regime remained in this form for 7 years, with the exception that in 1998 the State Board of Education mandated that the 10th grade ISTEP+ test become a gateway test—students had to achieve a passing score on this test to be eligible for graduation. This marked the beginning of the transition of the ISTEP+ test from a "low-stakes" to a "high-stakes" test. However, at that point the sanctions were applied to the students taking the test, not to the schools administering them. This changed with the modifications to the Indiana [legal] Code mandated by Public Law 221 in 1999 that are

[3] Pass/fail determinations were made according to the number of correct responses.

discussed below. In fall, 2003, the ISTEP+ program was modified to include an assessment of student science knowledge in 5th grade, and in fall, 2004, 4th and 7th grade language and math assessments were added to the program.

Policy Applications of ISTEP+

Results of the ISTEP+ test are the primary criteria that the Indiana DOE uses to evaluate the effectiveness of the educational programs offered by Indiana schools. Consequently, school corporations frequently use test results to assess the effectiveness of new curricula (Harris & Gilman, 2003), intervention programs (Jerome & Gilman, 2003) or textbook adoptions (Bolser & Gilman, 2003).

Concerns about the use of ISTEP+ test scores for school or program evaluation date from the earliest days of the testing program. Buechler (1991) conducted phone interviews with over 400 Indiana teachers and focus group interviews with another 65 teachers to determine what they perceived as the main constraints on their effectiveness in the classroom. The ISTEP test was the fourth most frequently cited constraint, mentioned by 16% of the teachers interviewed. The interviews were conducted before any sanctions were contingent upon test results; at the time the interviews were conducted, the ISTEP test was still a low-stakes test.

Indiana Public Law 221

In 1999 the state legislature passed Indiana Public Law 221. This law mandated that schools be placed in different descriptive categories based on the percentage of students in that school achieving a set score on the ISTEP+ tests. Language was inserted into several sections of the Indiana Code mandating that: “the performance of a school's students on ISTEP and other assessments recommended by the education roundtable and

approved by the board are the primary and majority means of assessing a school's improvement. (IC 20-10.2-5-1, Sec. 1. [a]).” A school’s academic performance, and the consequent rewards or sanctions associated with that performance, were determined by the ISTEP+ test scores of their students.

No Child Left Behind

The national No Child Left Behind (NCLB) Act, signed into law in January, 2002, mandates that by 2012 virtually all public school students pass a state administered assessment, and it affixes sanctions to schools failing to achieve this mandate. These sanctions include restructuring, the replacement or reassignment of personnel and, ultimately, school closure. In response to the law, the Indiana Board of Education has abolished its system of cohort schools. It has chosen to use ISTEP+ test results as the means of determining a school’s attainment of Adequate Yearly Progress (AYP). Each school is expected to show improvement in the percentage of students passing the math and language portions of ISTEP+ test until, by 2012, all students pass.

This use of ISTEP+ as the indicator of AYP creates a significant change in the use of the test. As stated earlier, the test was originally designed and used to assess individual student academic progress. The reliability and validity of this use is assessed annually by CTB/McGraw-Hill, the authors of the test. (See, for example, the 2003 ISTEP+ Technical Report, submitted to the Indiana Department of Education by CTB/McGraw-Hill, available from the Indiana Department of Education.) But the sanctions meted out by the dictates of the NCLB Act apply to the school, not the students. Despite its intended use, the de facto use of the ISTEP+ test is now no longer to assess individual student academic achievement, but rather to assess the quality of teaching within a school. To

date, the validity of the use of the ISTEP+ tests to evaluate teaching quality has been a de facto assumption, but it remains to be demonstrated. Considering the potential severity of the sanctions imposed on schools not showing AYP, it is reasonable to demand that the assessment instrument used to apply sanctions be reliable and valid. Furthermore, validity implies more than merely correlating test questions to specific content. Validity also requires that a test be appropriate for its intended use (Rulon,1946). Unfortunately, information about the validity of this use of the testing instrument is not readily available from the state.

Appropriate Use of the 5th Grade ISTEP+ Science Test

The use of the results of the 5th grade ISTEP+ science test (5-GIST) to determine AYP in science presupposes that the test is a reasonable indicator of the quality of science instruction within any given school. That supposition remains to be tested. Despite the delineation of both national and state standards for science instruction, (American Association for the Advancement of Science [AAAS], 1993; National Research Council [NRC], 1996; Indiana Accountability System for Academic Progress, 2002), there are no published reports of studies documenting the accuracy of the test scores in reflecting the degree of effectiveness or competence of a teacher in teaching those standards, and the appropriateness of this use has frequently been called into question. (See, for example, Russell, *et* al, 2004.) Before school corporations can effectively change curricula or pedagogy in an attempt to raise scores on the 5-GIST, more information is needed to ascertain that the test accurately measures science learning. Additionally, the impact of other forms of knowledge such as reading ability, vocabulary, or math computation skills on science test scores should be assessed to

determine what areas of instruction are most effective for schools to focus on in their attempts to improve test scores.

Validity Issues

The degree to which high stakes are contingent upon ISTEP+ test scores relate to their validity. Validity broadly relates to the degree to which a test score accuately reflects what it claims to measure. Rulon (1946) suggested that validity implies "whether the test does the work it is employed to do." However, Messick (1989a) notes that it is important to distinguish between the test and the test score. It is the test score to which validity is referenced. According to Messick (1989a), test validity is a "trichotomy" of three related considerations: content validity, criterion-related validity and construct validity.

Content validity. Content validation offers a judgment of how well a test samples the universe of the subject matter it purports to assess and about which conclusions are drawn (American Psychological Association, 1999). The Grade 5 Science - Guide to Test Interpretation (State of Indiana Department of Education and CTB/McGraw-Hill LLC., 2003) asserts: "The fall 2003 administration of ISTEP+ measured the performance of Indiana's Grade 5 students for the first time against Indiana's Academic Standards in science" (p. 3). Specifically, since the 5-GIST is currently administered at the beginning of grade 5, the grade 4 standards are used as the template for the test's content.

Yet the content validity of the test, although reasonably assumed, is not beyond reproach. The 5-GIST is a paper-and-pencil test, to be completed individually by students without collaboration or discussion. The science section of the Indiana Academic

Standards for 4th grade[4] (Indiana State Board of Education, 2000 – 2001a) lists the following preamble under The Nature of Science and Technology: "Students, working collaboratively, carry out investigations. They observe and make accurate measurements, increase their use of tools and instruments, record data in journals, and communicate results through chart, graph, written, and verbal forms" (p.24). These standards encourage active, hands-on processes. It has never been demonstrated that these goals can be assessed through individually administered, forced response or short answer tests. It is difficult to imagine how the ability to work collaboratively can be assessed by testing students individually. Hence, some of the science skills promoted by the standards appear to be poorly covered by the 5-GIST, if at all, and suggests that content validity is, at best, incomplete.

Although it is reasonable to assume that test questions reflect content mandated by state standards for science education, content validity alone is insufficient in assessing the appropriateness of the use of the 5-GIST in drawing conclusions about the quality of science instruction in a school. According to Cronbach and Meehl, (1959) content validity is most useful when attempting to correlate a behavior with the behavior required by the test-taking process. For example, if a science test required recall of scientific facts from memory by the test taker, then content validation can assure that the test is a reasonable measurement of the test taker's repertoire of scientific information. When using a memory recall test to assess quality of instruction, however, content validation alone can not assure the validity of the test.

[4] The 5-GIST is administered in the fall, roughly 5-7 weeks into the school year. Therefore, it is reasonable to assume that 4th grade science standards are more appropriate for examination than 5th grade. The 5th grade standards Nature of Science and Technology standards, however, contain similar wording.

Criterion-related validity. Validity can be explored predictively, by examining the content of a test, or affectively, by examining test results. Predictive validity includes both content validity and criterion-related validity, and is beyond the scope of this investigation. According to Popham (1978), criterion-related validity is an "attempt to correlate performance on a measure . . . with an independent—that is external—criterion." (p.35) Campbell and Fiske, (1959) refer to this as convergent validity, and it requires confirmation of results by measures independent of each other. To assess this form of validity would require, for example, that student 5-GIST test scores predict concurrently or subsequently an independent assessment of science achievement such as the National Assessment of Educational Progress (NAEP) or the Trends in International Math and Science Study (TIMSS). In the case of the 5-GIST, the attempt is to correlate test scores with quality of teaching. If the scores achieved on the 5-GIST were closely correlated to the results of an independent assessment of science-teaching quality, criterion-related validity could be assumed. To date, no systematic attempt has been made to correlate 5-GIST test scores to other science or science-teaching assessments.

Construct-related validity. Affective validity is often called construct validity, and its evaluation is often approached through an examination of discriminant validity. Maguire, *et* al, (1994) define a construct as "the underlying mechanism that explains not only the behavior on a specific indicator, but other, non-test manifestations as well." Construct validity is an assessment of the appropriateness of a test to its intended use and interpretation, and is generally considered to be a *post facto* test. Messick (1989b) describes this as "an integration of any evidence that bears on the interpretation of meaning of the test scores" (p.17). Campbell and Fiske (1959) advocate the use of

discriminant validation in attempting to show construct validity. They argue that the construct validity of an assessment instrument can be undermined by too high a correlation with other tests from which it are supposed to differ. Popham (1978) lists three steps involved in establishing construct validity:

> (1) A hypothetical construct presumed to account for test performance is identified. (2) One or more hypotheses regarding test performance are derived from the theory underlying the construct. (3) The hypotheses are then tested by empirical methods (p.36).

In the case of the 5-GIST, the hypothetical construct, based on the test's current use, could be that the scores on the test reflect the quality of science instruction within a school. Discriminant validation can be employed to assess the relative contribution of science instruction to 5-GIST scores by identifying and removing influences common to the 5-GIST, English and math test scores. Meaningful variability among the residuals can then be interpreted as contributions unique to specific schools.

Assumptions.

It is assumed that students within a classroom and within a school vary in their academic strengths and weaknesses and that these individual differences affect performance on ISTEP+ tests. Beyond inter-student variability, individual students vary in abilities across content domains. Thus, a student might score higher on one section of the ISTEP+, such as the mathematics section, than a different section, such as English. In that same class, however, another student

might have achieved the converse result. Additionally, there may be general abilities that account for a percentage of overall performance on all tests.

Furthermore, strengths and weaknesses tend to be stable across time. Ding and Davison (2005) found that over time higher and lower ability students learned at approximately the same rate, so that students beginning the testing regime below acceptable levels remained below the level for all 4 years even though they showed a comparable rate of intellectual growth. Similar findings were also reported in a longitudinal study by Rescorla and Rosenthal (2004), which concluded that initial 3rd grade standardized achievement test scores were a strong predictor of future scores. In addition, just as students vary within classrooms and school buildings, classroom climates and teacher quality vary within schools. Likewise schools themselves vary in important ways.

The 5th Grade Science ISTEP+ Test

If the researcher's assumption that ISTEP+ scores are significantly impacted by factors other than teaching quality is born out by the tests suggested above, then it is possible to examine the primary question of this research project: to what extent do the 5-GIST scores reflect the quality of science instruction within a school?

Recall the researcher's assertion that students have differing inherent strengths and abilities, and that these are reflected by the ISTEP+ test scores. That being the case, it is reasonable to assume that many students' abilities in science will be different from their abilities in math and language. According to this assumption, students strong (or weak) in math in 3rd grade are likely to be strong (or weak) in math in 6th grade, but not necessarily likely to be equally strong (or weak) in language in 6th grade or science in 5th

grade. Therefore, comparing individual students' ISTEP+ achievements relative to all the other students in the state of the same grade (i.e., Z-score) across years should yield a closer correlation between 3rd and 6th grade math scores, and 3rd and 6th grade language scores, than between 3rd grade math and 6th grade language, and 3rd grade language and 6th grade math. By extension, the same patterns should be evident when comparing 3rd grade math and 5th grade science, and 3rd grade language and 5th grade science, assuming that the 5-GIST is actually assessing science ability and knowledge. Finding an equally close or closer correlation between 3rd grade math and 5th grade science score as/than 3rd grade math and 6th grade math would suggest that the test is doing a better job assessing math ability than science, or at least that both tests are assessing the same skills, and calls discriminant validation into question; likewise for 3rd grade language and 5th grade science.

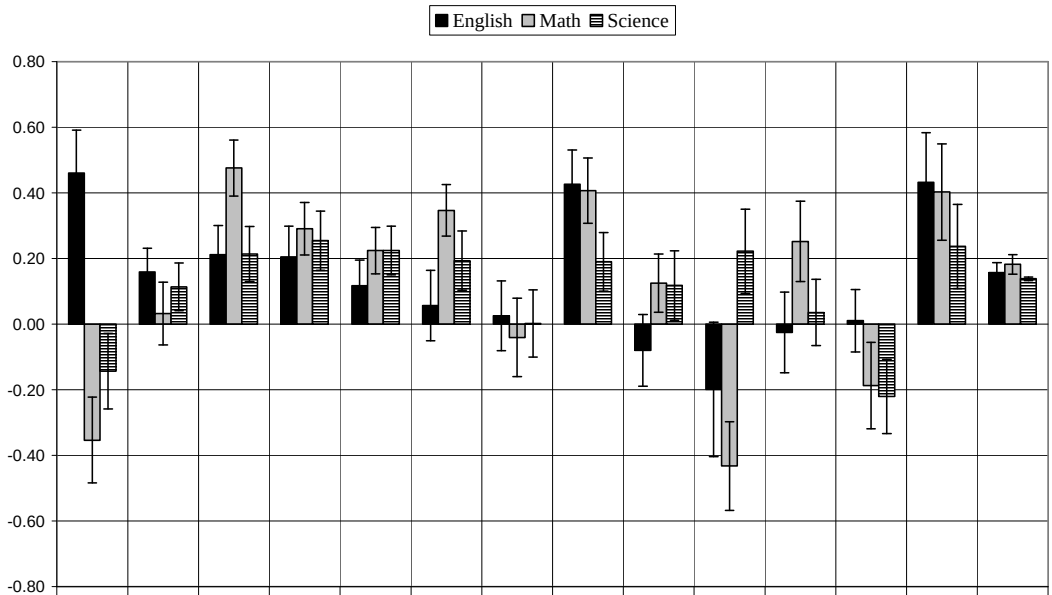
If, however, the earlier statistical tests suggest that quality of teaching has a stronger impact on test scores than other factors, such as innate ability, then it should be possible to identify schools performing exemplary science instruction by identifying schools with large numbers of students whose science scores exceed expectations given their language and math scores. Likewise, schools with large numbers of students whose 5-GIST scores are significantly below their language and math scores can be characterized as offering inadequate science programs.

Some preliminary analysis has already been conducted. Students in Cohort 1 who took all three (3rd, 5th and 6th grade) ISTEP tests in the same school were grouped according to school. The average change in Z-score between the 3rd and 6th grade math and English tests has been calculated for each score by subtracting the 3rd grade math or

English score from the 6th grade score. This reveals whether the scores of the students in a school typically improved or declined over time relative to the state population. Additionally, 3rd grade math and English scores were averaged together and subtracted from the 5th grade science score to offer evidence of the quality of science instruction in that particular school. The results of the investigation can be found on the graphs that follow. The graphs show the changes in Z-scores achieved by students who took all 3rd through 6th (Cohort 1) or 3rd through 5th (Cohort 2) grade ISTEP+ tests in the same school. The purpose of restricting the samples to only these students is to minimize the possibility of having the effects of instruction of other schools brought into host schools by transfer students reflected in the host schools' test results.

Figure 1

Mean ISTEP Z-Score Change for Cohort 1 Students Who Took All 3rd, 5th and 6th Grade ISTEP+ Tests in the Same School (Including 95% Confidence Intervals)[5]:

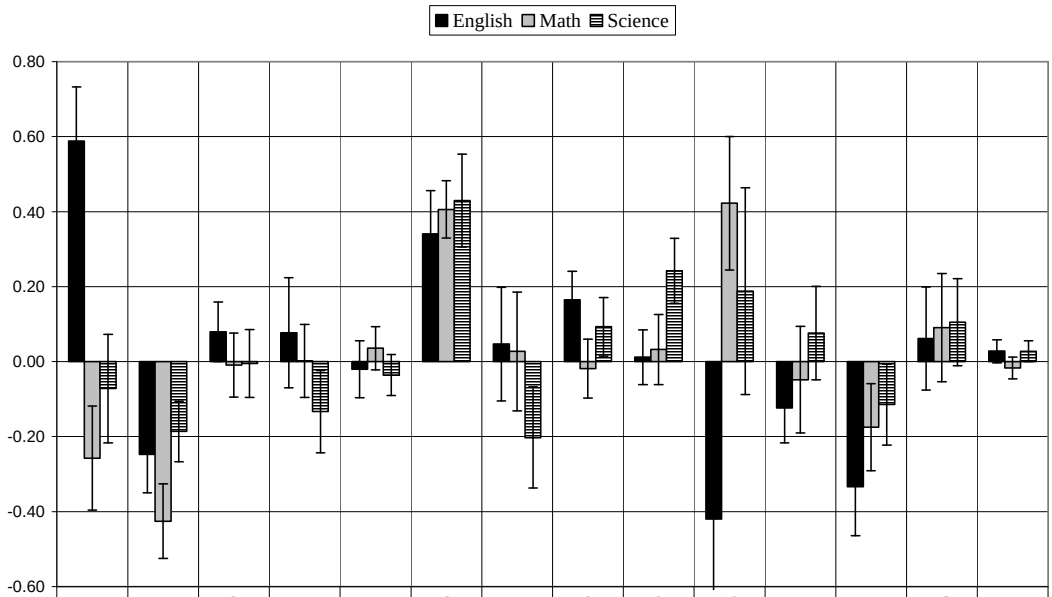


	1 (n=26)	2 (n=47)	3 (n=49)	4 (n=40)	5 (n=87)	6 (n=30)	7 (n=26)	8 (n=53)	9 (n=49)	10 (n=11)	11 (n=18)	12 (n=28)	13 (n=24)	Total
English	0.46	0.16	0.21	0.21	0.12	0.06	0.03	0.43	-0.08	-0.20	-0.03	0.01	0.43	0.16
Math	-0.35	0.03	0.48	0.29	0.22	0.35	-0.04	0.41	0.12	-0.43	0.25	-0.19	0.40	0.18
Science	-0.14	0.11	0.21	0.25	0.22	0.19	0.00	0.19	0.12	0.22	0.04	-0.22	0.24	0.14

[5] The change in science score value was determined by subtracting the average of the $English_3$ and $Math_3$ scores from the $Science_5$ score.

Figure 2

Mean ISTEP Z-Score Change for Cohort 2 Students Who Took Both 3rd and 5th Grade ISTEP+ Tests in the Same School (with 95% confidence intervals)



	1 (n=24)	2 (n=49)	3 (n=54)	4 (n=35)	5 (n=86)	6 (n=33)	7 (n=22)	8 (n=67)	9 (n=55)	10 (n=11)	11 (n=30)	12 (n=30)	13 (n=31)	Total
English	0.59	-0.25	0.08	0.08	-0.02	0.34	0.05	0.17	0.01	-0.42	-0.12	-0.33	0.06	0.03
Math	-0.26	-0.43	-0.01	0.00	0.04	0.41	0.03	-0.02	0.03	0.42	-0.05	-0.17	0.09	-0.02
Science	-0.07	-0.19	0.00	-0.13	-0.04	0.43	-0.20	0.09	0.24	0.19	0.08	-0.11	0.11	0.03

Clearly, there is a difference between schools in how much student test scores improve or decline over time. Many schools show substantial changes. Some schools show improvement of over 0.5 standard deviations (SD) in some areas; other schools show declines approaching the same magnitude. There are obvious differences between the scores of the 5-GIST and those of the English and math tests in many schools. If it can be established that these differences are caused by influences unique to individual schools, then schools offering exemplary science instruction and those exhibiting inadequate instruction can be identified. It is not clear, however, if this difference is the result of the quality of teaching offered by the school or by other factors. Furthermore,

the changes are not always consistent from the first year (Cohort 1) to the second (Cohort 2). Significant differences between cohorts in both impact magnitude and direction are apparent in many areas. This raises the question: Is there unique science achievement variance between schools, or is it accounted for by other factors such a SES, gender, etc.?

CHAPTER II

METHODOLOGY

The purpose of the research is to explore school effects on science teaching and to assess the reasonableness of using scores achieved on the 5-GIST to draw conclusions about the quality of science teaching in individual schools. The study attempts to determine if there is meaningful, unique variance of the science achievement test scores between schools after factors independent of teaching are accounted for.

Sample

The sample population is approximately 1,900 students in a Midwestern suburban/rural school corporation. Students were placed in one of two cohorts. Cohort 1 is composed of students who were in 3rd grade in 2001. Students in Cohort 2 were in 3rd grade in 2002. Table 1 shows the 2 cohorts and the test scores that have been received for all the students in the school corporation:

Table 1

Tests Taken by the Two Cohorts Whose Scores Are Examined in this Study.

	2001	2002	2003	2004
Cohort 1	3rd grade English and Math		5th grade Science	6th grade English and Math
Cohort 2		3rd grade English and Math		5th grade English, Math and Science

There were approximately 1,000 students in Cohort 1 and 900 students in Cohort 2. Four students from Cohort 1 were retained before taking the 5-GIST in 2003. Their scores are discarded.

All students were assigned a random ID code allowing their scores to be tracked from 3rd to 5th or 6th grade, depending on their cohort. In addition to scale scores achieved on the different tests, the following information was recorded: school, reduced/free lunch status, ethnicity, use of an individualized education program (IEP)[6], and limited English proficiency; classroom assignments were not recorded. Additionally a school stability variable was developed. Test scores were examined to determine if students took both 3rd and 5th grade tests in the same school. If both tests were taken in the same school, students were considered to exhibit school stability. Thus, it is possible to examine both individual effects and school effects, but it is not possible to examine classroom effect. For the purposes of this study, variance between classrooms within schools will not be explored and thus that source of variability will be attributed to uncontrolled error.

[6] The Indiana Department of Education refers to students with an IEP as special education students.

Table 2

Number of Students in Different Descriptive Categories.

	Cohort 1 (N_1=1002)		Cohort 2 (N_2=907)	
	Number	Percent	Number	Percent
Gender				
Female	**455**	**45.4%**	**437**	**48.2%**
Male	**547**	**54.6%**	**470**	**51.8%**
Ethnicity				
Asian	**54**	**5.4%**	**51**	**5.6%**
African American	**44**	**4.4%**	**39**	**4.3%**
Hispanic	**23**	**2.3%**	**25**	**2.8%**
Native American	**3**	**0.3%**	**1**	**0.1%**
White	**820**	**81.8%**	**745**	**82.1%**
Multi-Racial	**58**	**5.8%**	**46**	**5.1%**
Socioeconomic Status (SES)				
Free Lunch	**236**	**23.6%**	**203**	**22.4%**
Reduced Lunch	**60**	**6.0%**	**59**	**6.5%**
No Reduced Lunch	**660**	**65.9%**	**641**	**70.7%**
Unreported	**46**	**4.6%**	**4**	**0.4%**
English Proficiency				
Limited Proficiency	**55**	**5.5%**	**55**	**6.1%**
Proficient	**947**	**94.5%**	**851**	**93.8%**
Not Reported			**1**	**0.1%**
Use of IEP				
IEP	**160**	**16.0%**	**132**	**14.6%**
No IEP	**842**	**84.0%**	**775**	**85.4%**

Test Scores

State means and standard deviations for all tests (English, math and science) are available from the state Department of Education. Scale scores provided by the local school corporation were converted to Z-scores—using state means and standard deviations—and thus comparing the achievement of individual students to all the students in the state taking the same test. This allowed the researcher to track individual progress relative to the other students of the state in the same grade. By doing this it was possible to determine if individual student performance improved against the state average from one test to another.

Checking for Convergent and Discriminant Validation

For this study the 3rd grade English and math and the 5th grade science test scores received the most scrutiny, but the 5th and 6th grade English and math scores were used to examine convergent and discriminant validity and to check correlational stability. This examination was attempted by correlating science test scores with those of all other tests. The process is referred to as discriminant validation. It was assumed that the correlation of the 3rd grade with the 5th or 6th grade English scores and the 3rd grade math with the 5th or 6th grade math scores would be greater that the correlation of any of these scores with the 5th grade science score. If the correlation of 3rd grade language, for example, is higher with 5th grade science than with 6th grade language, then that suggests that factors other than the knowledge of the specific subject being tested may be influencing the results.

Table 3

Convergent and Discriminant Validity Matrix

Cohort 1

		3rd grade		5th grade	6th grade
		M	E	S	M
3rd grade test	Math				
	English	<100%			
5th grade test	Science	*<100%*	*<100%*		
6th grade test	Math	**<100%**	<100%	*<100%*	
	English	<100%	**<100%**	*<100%*	<100%

In this matrix, a reliability diagonal does not exist. The correlation of the score on any test with itself is 100%, which is essentially meaningless, so these numbers are eliminated. The validity diagonal is represented by the numbers in bold type. This represents the correlation of the 3rd grade math score with the 6th grade math score and the 3rd grade language score with the 6th grade language score. To show convergent validity, these correlations must be relatively strong. The correlation of the science scores with the other test scores appears in italics. According to Campbell and Fiske, these correlations must be lower than those in the validity diagonal in order to demonstrate convergent and divergent validity. Both convergent and discriminant validity must be demonstrated in order for construct validity to be assumed.

The matrix for Cohort 2 is essentially the same except that the 6th grade test scores are replaced with fifth grade scores:

Table 4

Convergent and Discriminant Validity Matrix

Cohort 2

		3rd grade		5th grade	
		M	E	S	M
3rd grade test	Math				
	English	<100%			
5th grade test	Science	*<100%*	*<100%*		
	Math	**<100%**	<100%	*<100%*	
	English	<100%	**<100%**	*<100%*	<100%

Checking Stability of Validity Across Cohorts

R-square values of the test correlations were computed, and the following table was constructed to check the stability of the English/science and math/science correlations across groups:

Table 5

Stability of Correlations over Time.

	Cohort 1	Cohort 2
English$_3$/Science	**R^2, (compare with Cohort 2 English$_3$/Science R^2)**	**R^2, (compare with Cohort 1 English$_3$/Science R^2)**
Math$_3$/Science	**R^2, (compare with Cohort 2 Math$_3$/Science R^2)**	**R^2, compare with Cohort 1 Math$_3$/Science R^2)**

Evidence for Instructional Impact

Using linear regression analysis it is possible to statistically estimate the degree to which different criteria account for science test scores. Given some common variance between English, math and science, does there remain significant unique variance of science scores among schools? The goal was to remove all the contributors unique to the individual student—essentially demographic data and achievement—leaving residuals that will be interpreted to represent school impact. The first step was to determine the contribution of English and math variance to science variance. Next the contribution of demographic data was determined. Once an idealized regression line had been established the residuals for each school were examined. For this analysis only the scores of those students who took all tests in the same school were included. This was to prevent the possibility of the residuals reflecting the instruction delivered by a different school due to students transferring into the school during the testing regime. Since all other common influences have been removed, the residuals can be interpreted to reflect the impact of the science instruction afforded by the school.

CHAPTER III

RESULTS

The data analyses of the present study are organized to address each of the hypotheses and a set of ancillary issues related to documenting the convergent and discriminant validity of the Grade 5 ISTEP+ science achievement measure. Ultimately, if that measure is to be used as an indicator of the quality of science instruction within a school, the ISTEP+ science achievement measure must simultaneously possess meaningful variance that is non-redundant to other indicators, while at the same time cross-validating other indicators. To review, all student test data were standardized as Z-scores based on respective state means and standard deviations prior to analysis.

Convergent and Discriminant Validity.

Cohort 1. Table 6 presents the convergent and discriminant validity matrix for Cohort 1. Shown are Pearson bivariate correlations with pairwise deletions, for all pairs of tests taken by the same student. That is, if a student took the 3rd and 5th grade ISTEP+ within the school district but not the 6th grade, then correlations between 3rd and 5th grade tests only were included in the calculations. All correlations were statistically significant at the 0.01 level.

Table 6

Convergent-Discriminant Validity Correlations

Cohort 1

		Grade 3		Grade 5	Grade 6	
		English	Math	Science	English	Math
Grade 3 tests	English	-				
	Math	.671	-			
Grade 5 test	Science	*.690*	*.682*	-		
Grade 6 tests	English	**.735**	. 625	*.767*	-	
	Math	. 664	**.744**	*.774*	.740	-

Shared Covariance with General Achievement. Treating the $English_{3\text{-}6}$ scores and the $Math_{3\text{-}6}$ scores as test/retest yields reliability ratings of .735 and .744 respectively (bold type). These are higher than the correlations (italics) between $English_3$ and $Science_5$ (.690) and $Math_3$ and $Science_5$ (.682), but were marginally lower than the correlations between $Science_5$ and the $Math_6$ and $English_6$ scores (.774 and .767, respectively). This difference, however, should be interpreted with caution since the $Math_3$ – $Science_5$ correlation is a validity coefficient while $Math_3$ – $Math_6$ is a reliability (stability) coefficient.

Cohort 2. Table 7 presents the convergent and discriminant validity matrix for Cohort 2. The pattern of correlations for Cohort 2 is consonant with Cohort 1. In this case the highest correlation is between 5th grade science and math. Reliability ratings were in a comparable range with Cohort 1. A test of goodness of fit of the two patterns yields $X^2_{(df=9)} < 1.0, ns$.

Table 7

Convergent-Discriminant Validity Correlations:

Cohort 2

		Grade 3		Grade 5		
		English	Math	Science	English	Math
Grade 3 tests	English	-				
	Math	.697	-			
Grade 5 tests	Science	*.735*	*.679*	-		
	English	**.700**	.621	*.767*	-	
	Math	.684	**.755**	*.807*	.735	-

Shared Covariance with General Achievement. A valid indicator of science achievement must be demonstrably independent of general achievement. Table 8 presents the squared bivariate correlations of $English_3$ and $Math_3$ achievement with $Science_5$ achievement and the combined squared multiple correlation of the combined contribution of $English_3$ and $Math_3$ to $Science_5$ science achievement. The squared correlation (the coefficient of determination) reflects the percent of shared variance between and among the independent and dependent variables.

Table 8

Squared Correlation Coefficients Reflecting General Achievement

	Cohort 1	Cohort 2
$English_3/Science_5$	.48	.54
$Math_3/Science_5$	.47	.46
$English_3$+ $Math_3/Science_5$	.56	.60

For Cohort 1, 48% of the variance in $Science_5$ achievement is attributable to variance in $English_3$ achievement; similarly, 47% can be attributed to variance in $Math_3$ scores. The combined multiple correlations of $English_3$ and $Math_3$ accounts for 56% of variance in $Science_5$ achievement. For Cohort 2, 54% in grade 5 science achievement is attributable to variance in $English_3$ achievement; similarly, 46% is attributed to variance in $Math_3$ scores. The combined multiple correlations of $English_3$ and $Math_3$ accounts for 60% of variance in $Science_5$ achievement. While the patterns are parallel across the two cohorts, it would be unwise to combine the two cohorts. To do so might mask different patterns across schools across time.

Note, the $English_3/Science_5$ covariance is greater that the $Math_3/Science_5$ covariance in both cohorts. The covariance with the $Math_3$ scores was consistent across cohorts, but covariance with $English_3$ across cohorts showed somewhat less consistency. However, Fisher's *r* to Z transformation yielded an effect size of .52 for Cohort 1 and .60 for Cohort 2. Both reflect large effect sizes.

Ethnicity as a Moderating Variable

An assessment of the potential differential role of ethnicity was initially conducted using linear regression lines with $English_3$ predictive of $Science_5$ scores.

Figure 3 displays regression lines for ethnicity within in Cohort 1. Both Hispanic (n=6) and Indian subgroups (n=3) were excluded because there are too few cases. Multiethnicity was also excluded because of the possibility of different interpretations of the term by students. The slope of the regression line is given in the form of $Science_{5i.group} = \beta_{i.group} English_{3i.group} + a_{group}$, where $Science_{5i.group}$ is the predicted student score on the $Science_5$ within the specific group, $\beta_{i.group} English_{3i.group}$ is the slope of $English_3$ achievement on $Science_5$ achievement for the specific group, and a_{group} is the intercept for the group. Relative impact of slope and intercept for groups were compared.[7] Table 9 presents pairwise comparisons of slopes and intercepts for ethnicity.

[7] Statistical comparison of slopes is accomplished by $t_{df=n1+n2-2} = \frac{\beta_i - \beta_j}{Se_{\beta i - \beta j}} = \frac{\beta_i - \beta_j}{\sqrt{Se_{\beta i}^2 + Se_{\beta j}^2}}$.

Statistical comparison of intercepts is accomplished by $t_{df=n1+n2-2} = \frac{a_i - a_j}{Se_{ai-aj}} = \frac{a_i - a_j}{\sqrt{Se_{ai}^2 + Se_{aj}^2}}$.

Figure 3

Regression Lines for Ethnicity

Cohort 1

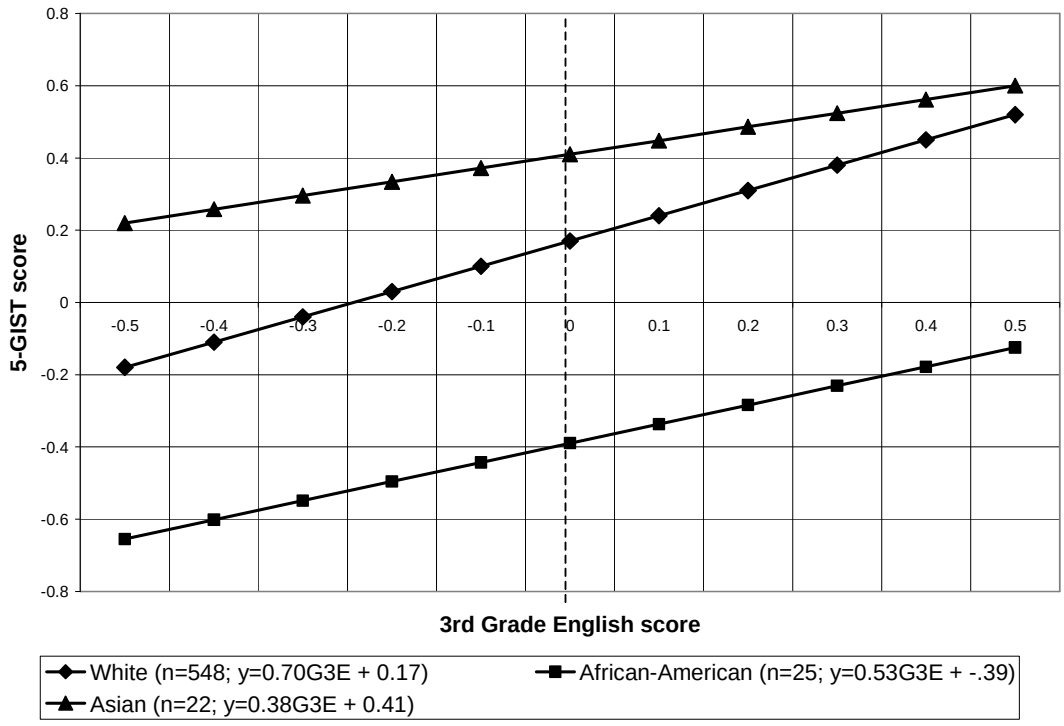


Table 9

Pairwise T-tests for Slopes and Intercepts of Ethnic Groups

Cohort 1[8]

		Comparing Slopes			Comparing Intercepts	
	$\beta_{i.group}$	Asian	Afr-Amer	a_{group}	Asian	Afr-Amer
White	0.70	2.62	1.22	0.17	-1.83	3.65
Asian	0.38	-	-0.83	0.41	-	4.07
African American	0.53	0.83	-	-0.39	-4.07	-

All t-tests assume infinite degrees of freedom. Interpretation of ethnic comparisons is somewhat problematic since sample sizes for Asian and African-American students are small and standard errors inflated. The difference in slopes between the White and Asian students is significant at the 0.01 level, i.e. the slope for white students was greater. The difference in intercept between the African-American student population and both the white and Asian student populations is statistically significant at the 0.001 level[9]. This pattern can be seen also with Cohort 2. Figure 4 displays regression lines for ethnicity within Cohort 2, and Table 10 the respective pairwise comparisons.

[8] Two-tailed t-test. Outcome >2.58 indicates significance of p<.01. Given the large sample size, p<.01 is used to avoid inflating Type 1 error.

[9] Since the scores are Z-scores, the analyses are effectively "grand mean centered".

Figure 4

Regression Lines for Ethnicity

Cohort 2

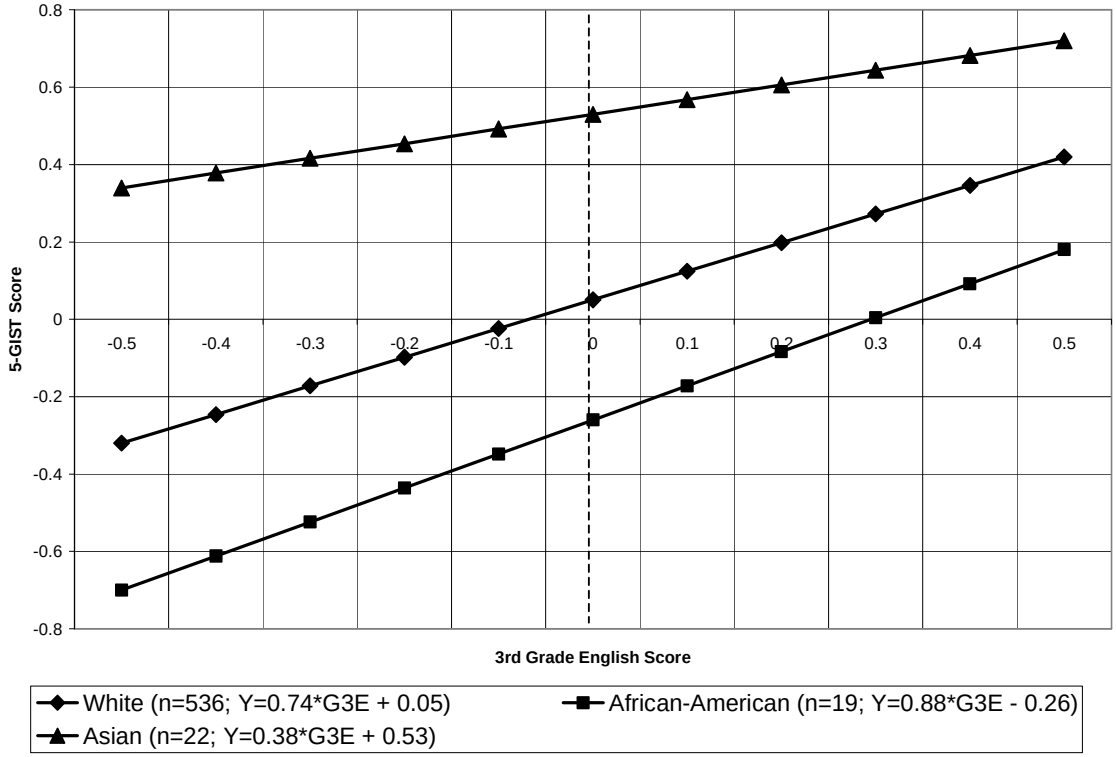


Table 10

Pairwise T-tests for Slopes and Intercepts of Ethnic Groups

Cohort 2

		Comparing Slopes			Comparing Intercepts	
	$\beta_{i.group}$	Asian	Afr-Amer	a_{group}	Asian	Afr-Amer
White	0.74	2.58	-0.78	0.05	-3.00	2.43
Asian	0.38	-	-2.25	0.53	-	3.95
African American	0.88	2.25	-	-0.26	-3.95	-

As with Cohort 1, the slope for Asian students is significantly different from the other slopes. This may be a result of a few unusual cases. Conversely, the difference in intercepts between the three ethnic groups appears to be stable ($p<.01$).

Socioeconomic Status as a Moderating Variable

Linear regression lines with $English_3$ predictive of $Science_5$ scores were established to assess the potential differential role of socio-economic status (SES) as reflected by reduced/free lunch status. SES regression also reveals differences between groups. Figure 5 displays regression lines for SES within in Cohort 1 and Table 11 the respective pairwise comparisons.

Figure 5

Regression Lines for SES

Cohort 1

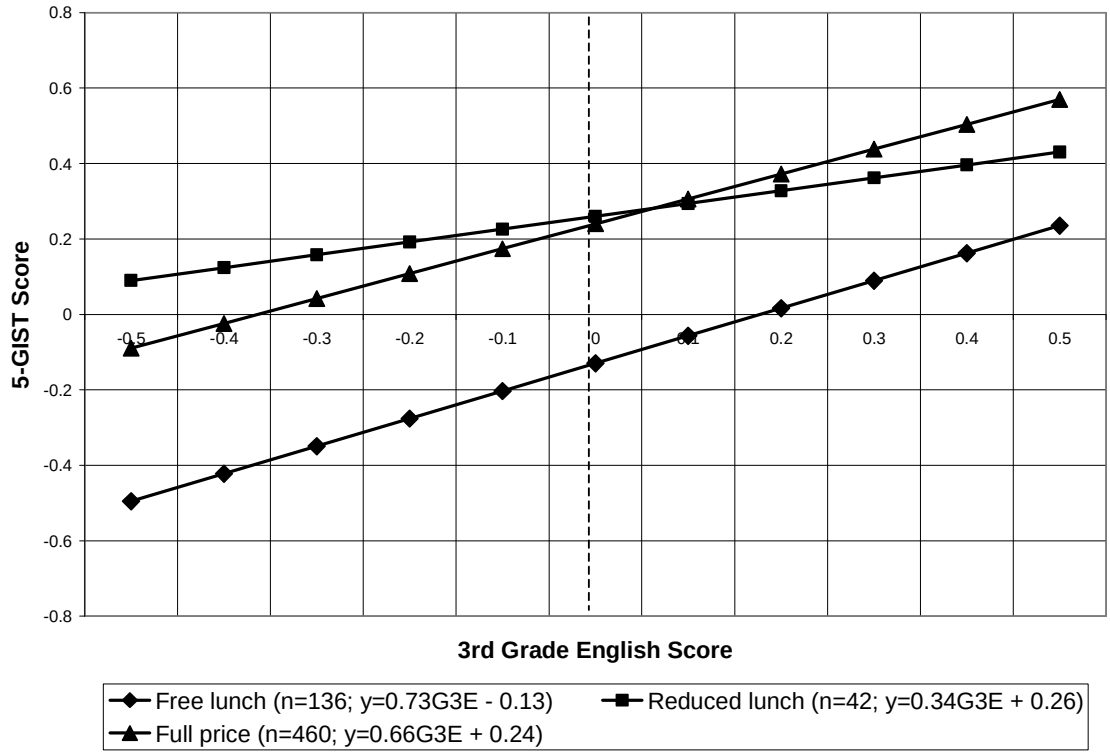


Table 11

Pairwise T-tests for Slopes and Intercepts of SES

Cohort 1

		Comparing Slopes			Comparing Intercepts	
	$\beta_{i.group}$	Reduced	Full	a_{group}	Reduced	Full
Free lunch	0.73	3.20	0.87	-0.13	-3.98	-4.65
Reduced price	0.34	-	-3.06	0.26	-	1.04
Full price	0.66	3.06	-	0.24	-1.04	-

The difference in slope between the reduced price lunches and the free and full-priced lunches is significant at the 0.01 level. The differences in intercepts between the free-lunch students and both the reduced- and full-price students are significant at the 0.001 level.

Figure 6 displays regression lines for SES within in Cohort 2 and Table 12 the respective pairwise comparisons.

Figure 6

Regression Lines for SES

Cohort 2

Table 12

Pairwise T-tests for Slopes and Intercepts of SES

Cohort 2

		Comparing Slopes			Comparing Intercepts	
	$\beta_{i.group}$	Reduced	Full	a_{group}	Reduced	Full
Free lunch	0.77	0.21	1.46	-0.17	-1.52	-4.19
Reduced price	0.74	-	0.63	0.02	-	-1.09
Full price	0.66	-0.63	-	0.14	1.09	-

The regression lines indicate that free lunch students in both cohorts do not progress as well as most other students based on $English_3$ Z-scores, remaining approximately 0.3-0.4 SD below the full-price-lunch members of their cohort. However, for Cohort 2 a pairwise t-test shows that the slopes are not significantly different. The difference in intercepts between the free and full-price regression lines is significant beyond the 0.001 level.

Mobility as a Moderating Variable

Precise student mobility data were not available for analysis, but an approximation of the impact of student mobility was assessed by comparing the schools at which at a given student took the 3rd grade and 5th grade tests. Students who took both tests at the same school were classified as exhibiting school stability, while students taking the tests at different schools within the local school district were classified as mobile. Students who did not have scores for either the 3rd or the 5th grade test were not

included in the regression. This would include any students transferring into or out of the district between the time that the 3rd and 5th grade tests were administered. Hence, the "mobile" designation refers to only intra-district mobility.

Figure 7 displays regression lines for school stability in Cohort 1 and Table 13 the respective pairwise comparisons.

Figure 7

Regression Lines for School Stability

Cohort 1

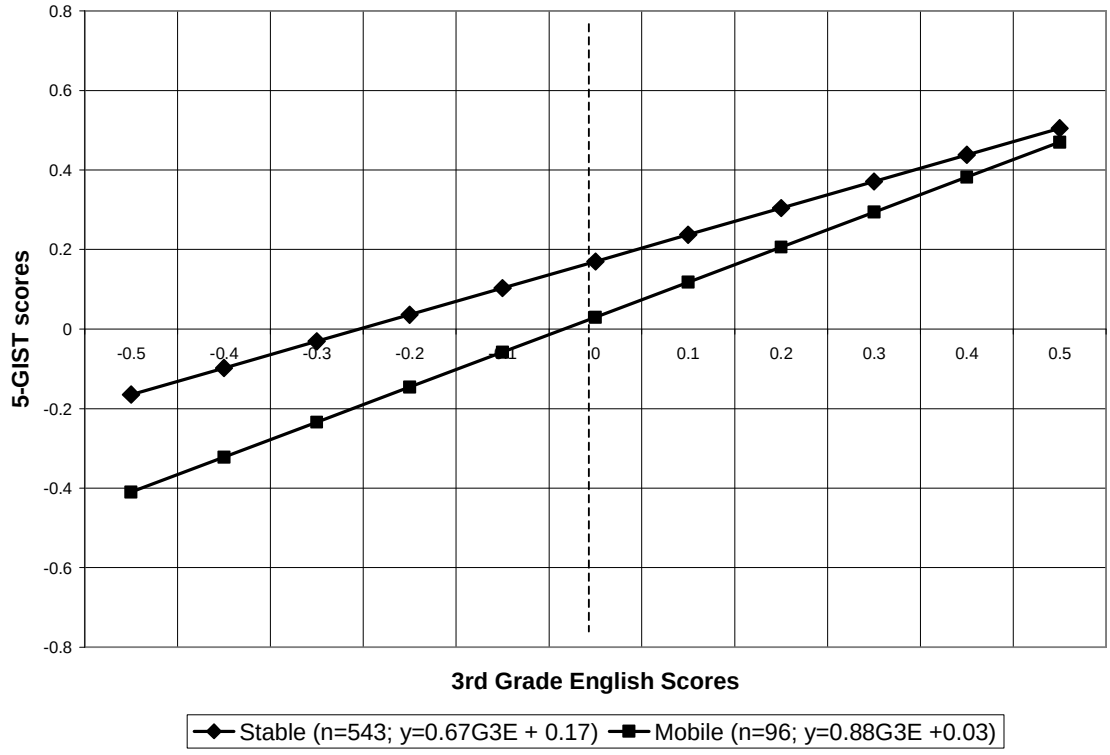


Table 13

Pairwise T-tests for Slopes and Intercepts of School Stability

Cohort 1

		Comparing Slopes		Comparing Intercepts
	$\beta_{i.group}$	Mobile	a_{group}	Mobile
Stable	0.67	-2.46	0.17	1.54
Mobile	0.88	-	0.03	-

The t-test shows that the difference in slope between groups is significant at the 0.05 level.

Figure 8 displays regression lines for school stability in Cohort 2 and Table 14 the respective pairwise comparisons.

Figure 8

Regression Lines for School Stability

Cohort 2

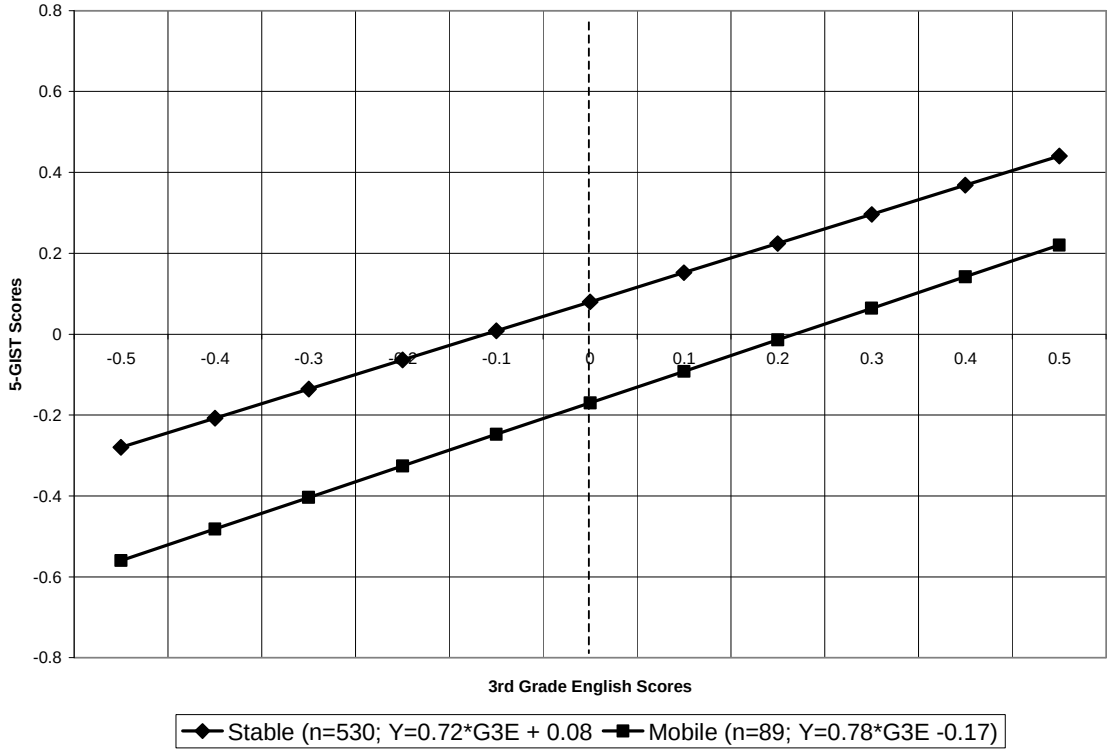


Table 14

Pairwise T-tests for Slopes and Intercepts of School Stability

Cohort 2

		Comparing Slopes		Comparing Intercepts
	$\beta_{i.group}$	Mobile	a_{group}	Mobile
Stable	0.72	-0.85	0.08	3.48
Mobile	0.78	-	-0.17	-

For Cohort 2 there is no significant difference in slope between the two groups, but the difference in intercept is significant at the 0.001 level.

Additional Demographic Variables.

Table 15 below lists the slopes of the regression lines for the other demographic variables.

Table 15

Regression Lines of Other Demographic Variables

Cohort 1			Cohort 2	
Demographic Variable	Number	Slope of Regression	Number	Slope of Regression
Gender:	M (n=357) F (n=282)	0.75*G3E + 0.20 0.67*G3E + 0.07	M (n=329) F (n=290)	0.80*G3E + 0.07 0.67*G3E + 0.02
Limited English:	Yes (n=22) No (n=617)	0.39*G3E + 0.27 0.72*G3E + 0.14	Yes (n=24) No (n=594)	0.76*G3E + 0.15 0.73*G3E + 0.04
IEP:	Yes (n=104) No (n=535)	0.84*G3E + 0.09 0.65*G3E + 0.19	Yes (n=86) No (n=533)	0.69*G3E – 0.15 0.70*G3E + 0.08

Table 16

Pairwise T-Tests for Slopes and Intercepts of Other Demographic Variables

	Cohort 1				Cohort 2			
Variable	$\beta_{i.group}$	Slopes	a_{group}	Intercepts	$\beta_{i.group}$	Slopes	a_{group}	Intercepts
Gender								
Male:	0.75	1.38	0.20	2.19	0.80	2.42	0.07	0.90
Female:	0.67	-	0.07	-	0.67	-	0.02	-
Lim. English								
Yes:	0.39	-1.06	0.27	0.40	0.76	0.27	0.15	.093
No:	0.72	-	0.14	-	0.73	-	0.04	-
Sp. Education								
Yes:	0.84	2.03	0.09	-0.97	0.69	-0.08	-0.15	-1.63
No:	0.65	-	0.19	-	0.70	-	0.08	

None of the differences in slopes or intercepts is significant at the .01 level.

Significance of Covariance.

Despite different regression lines for some of the demographic variables, a test of significance of the variables revealed that most did not have a statistically significant impact on test scores when combined with the effects of the $English_3$ and $Math_3$ scores on the $Science_5$ scores. Table 17 shows the impact of demographic variables when combined with 3rd grade test scores.

Cohort 1. Checking for covariance between $Science_5$ scores (dependent variable) and $English_3$ scores (independent variable) reveals an adjusted R^2 of .476. This is

statistically significant beyond the .001 level. Adding $Math_3$ as an independent variable increases R^2 to .560, accounting for 56% of the variance in $Science_5$ scores. The large percentage of the variance in $Science_5$ scores due to 3rd grade scores tends to mask the influence of most of the demographic variables. The only other variable shown to be significant was SES.

Cohort 2. Both SES and school stability appear to influence test results. No other demographic variables were significant below the p=0.05 level. The improvement in R^2 value achieved by adding school stability to the $English_3$, $Science_5$ and SES regression is 0.002, which is so slight it is effectively insignificant. Therefore, to generate a consistent, meaningful regression formula across cohorts, only $English_3$ and $Math_3$ scores and SES will be applied as control variables.

Table 17

Adjusted R^2 and Significance of Demographic Variables

	Cohort 1		Cohort 2	
Dependent variables	Significance	Adjusted R^2	Significance	Adjusted R^2
English	>.001	.476	>.001	.560
English Math	>.001 >.001	.560	>.001 >.001	.597
English Math Ethnicity	>.001 >.001 .151	.561	>.001 >.001 .304	.597
English Math SES	>.001 >.001 >.001	.571	>.001 >.001 .002	.603
English Math School Stability	>.001 >.001 .359	.560	>.001 >.001 .003	.602

Regression Lines

Cohort 1. Creating a regression line for the three independent variables (3^{rd} grade $English_3$ and $Math_3$ scores plus SES) and then examining the residuals essentially removes the outside influences and allows examination of individual school effects and a comparison of the relative quality of those effects. For Cohort 1 the regression formula is $Science_5 = 0.417*English_3 + 0.374*Math_3 + 0.139*SES - 0.181$, where $Science_5$ equals

the predicted Z-score on the $Science_5$, $English_3$ equals the Z-score on the $English_3$ test, $Math_3$ equals the Z-score on the $Math_3$ test and *SES* equals socioeconomic status as inferred from free/reduced school lunch status.

Cohort 2. For Cohort 2 the formula for the regression line is $Science_5 = 0.492*English_3 + 0.313*Math_3 + 0.107*SES - 0.225$, with all variables the same as for Cohort 1. Combining these two formulae results in the following formula for the regression line for both cohorts: $Science_5 = 0.453*English_3 + 0.341*Math3 + 0.124*SES - 0.206$. This is the formula used for the rest of the regression analysis.

Table 18 shows the mean residuals for Cohorts 1 and 2 of the $Science_5$ scores according to the school that administered the $Science_5$ test.

Table 18

Mean Standardized Residuals According to School

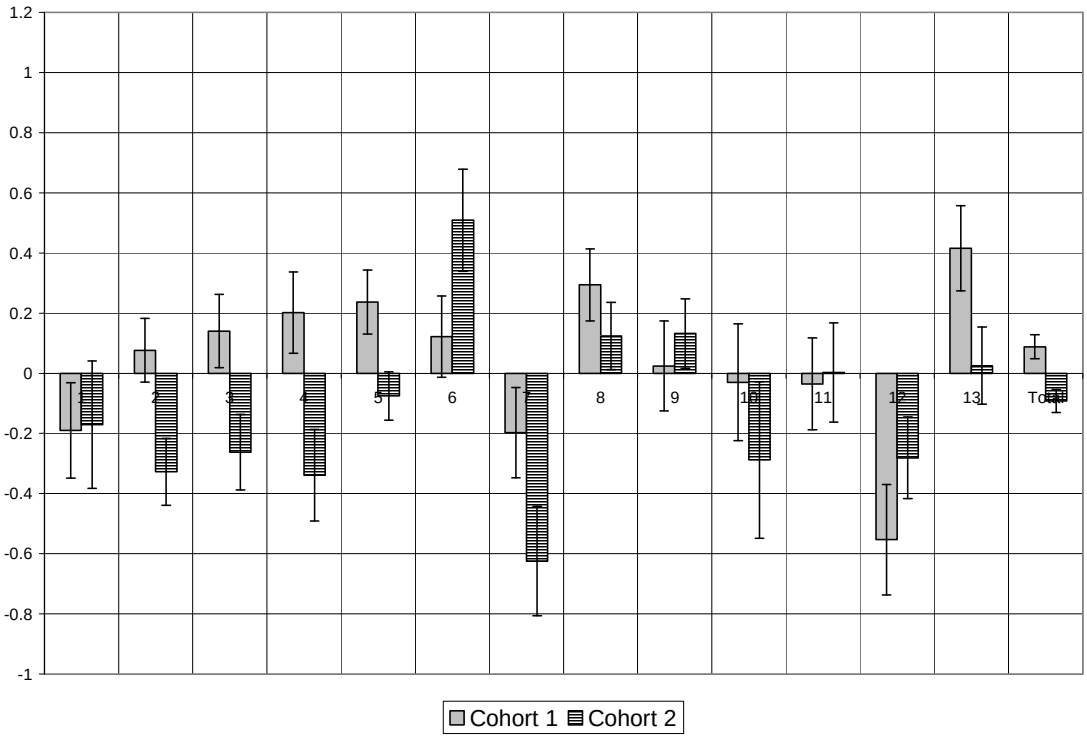
School Number	Cohort	Mean	Cohort	Mean
1	1	-0.189	2	-0.171
2		0.077		-0.328
3		0.141		-0.262
4		0.202		-0.338
5		0.237		-0.075
6		0.123		0.510
7		-0.197		-0.624
8		0.294		0.124
9		0.025		0.132
10		-0.030		-0.288
11		-0.035		0.003
12		-0.553		-0.281
13		0.416		0.026
Total		0.089		-0.092

The standard error of the estimate (SEE) for both cohorts is 0.647. Although there are differences between cohorts within a school, in all cases they are less than the SEE. The residuals, which are interpreted as the "school effect," show marked differences from the raw $Science_5$ Z-scores. In all cases, school effect refers to the effect of the school that administered the $Science_5$. Since school stability was not shown to have a significant impact on the residuals, the effect of intra-district transfers between 3^{rd} and 5^{th} grade is discounted.

Figure 9 is a graph of the residuals listed in Table 18. In this chart the scores are adjusted to account for the scores on the $English_3$ and $Math_3$ tests and SES.

Figure 9

Mean $Science_5$ Residuals Controlled for $English_3$, $Math_3$ and SES According to School



Most schools appear fairly consistent across cohorts. Where there are some differences, they are statistically small and usually less than 0.5 SD. Schools 6, 8 and 13 appear to exert a positive school effect, relative to state means, while schools 1, 7 and 12 appear to exert a negative effect. The other schools appear to have an effect close to the state norms.

Figure 10 and Table 19 compare the marginal means of the residuals for cohort and school effects.

Figure 10

Estimated Marginal Means of Standardized Residuals According to School

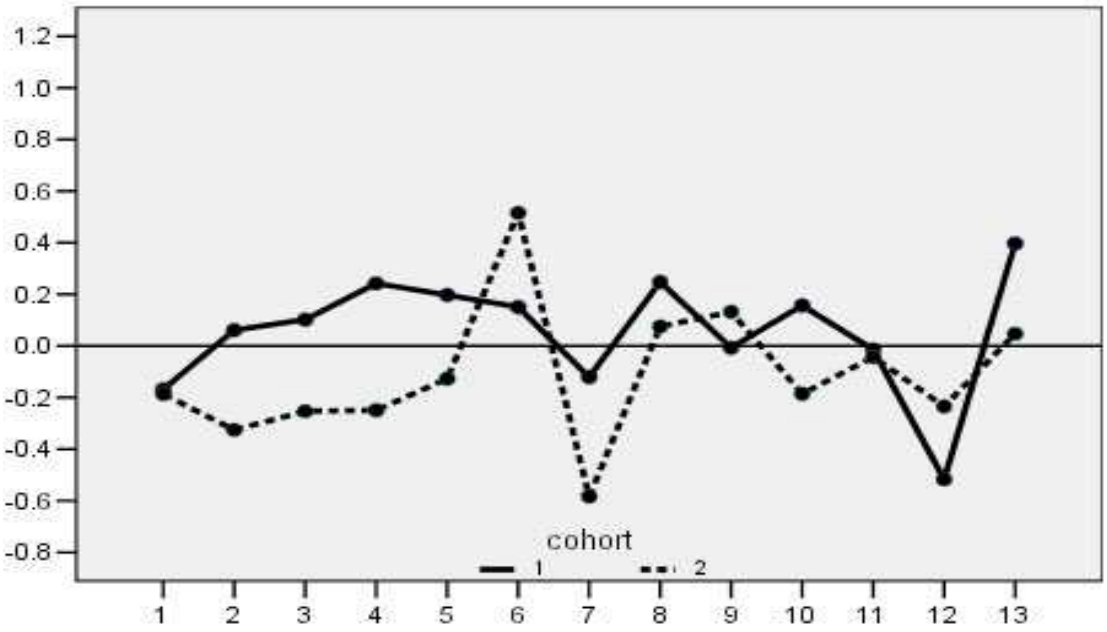


Table 19

Tests of Between-Subjects Effects for Residuals

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	71.269(a)	25	2.851	2.970	.000
Intercept	.712	1	.712	.742	.389
cohort	6.966	1	6.966	7.257	.007
school	40.566	12	3.381	3.522	.000
cohort * school	21.260	12	1.772	1.846	.037
Error	1180.731	1230	.960		
Total	1252.000	1256			
Corrected Total	1252.000	1255			

a. R Squared = .057 (Adjusted R Squared = .038)

Because of the large number of subjects, a significance level greater than .01 is considered not significant. Table 19 shows a significant (<.001) effect of schools on $Science_5$ scores. It also shows a significant cohort effect (.007). However, the difference in school effect between cohorts (see Table 18) is approximately 0.18 SD, which, although statistically significant, is of little practical significance as it represents a relatively slight difference in terms of actual achievement. Although the school effect is statistically significant (.000), interaction between school and cohort is not (.037).

CHAPTER IV

DISCUSSION

The results of the 5-GIST are used to assess the quality of science instruction offered by schools. The validity of such a use has not been shown in the research literature. The purpose of this study is to determine if there is meaningful variance between schools in the scores of the 5-GIST that can not be accounted for by general aptitude and SES. Any variance independent of these factors may reasonably be attributed to school effects. Many of the results of the analysis yield insights into the validity of using the 5-GIST scores to draw conclusions about the quality of science instruction afforded by the administering school. Convergent and Divergent Validity Matrices

Divergent validity is reflected in Tables 6 and 7. The 3rd – 6th grade (for Cohort 1) and 3rd – 5th grade (for Cohort 2) are, effectively, reliability coefficients and range from .700 to .755 over two or three years, depending upon the cohort. Validity coefficients between $Science_5$ and $English_6$ and $Math_6$ for Cohort 1 and the $Science_5$ and $English_5$ and $Math_5$ for Cohort 2 range from .621 to .807 or squared correlations from .386 to .651. Interestingly, for Cohort 1 the correlation between $English_6$ and $Math_6$ scores (.740) is not as high as the correlations of both these tests with the $Science_5$ (.767 and .774, respectively), which suggests a cognitive maturity effect, i.e. age. This pattern is similar for Cohort 2, but perhaps less unexpected since all three tests were administered during the same year. The high degree of covariance between the three tests

suggests that all three are influenced by the same phenomenon, which will be referred to as "general aptitude" in this study.[10]

In any case, the validity coefficients suggest that science achievement, reflected in performance on the ISTEP+ measures accounts for variance independent from general achievement reflected in English achievement, Math achievement, or the combination of the two. Thus, there remain other influences on the science test scores that are independent of general ability.

Regression Lines for Ethnicity

A comparison of regression lines across ethnic groups reveals that, even controlling for general achievement, differences in overall achievement (reflected in differences in intercept) remain. $Science_5$ performance of African-American students consistently lagged behind those of white and Asian students who achieved similar scores on the 3rd grade English test. Most cases show a growth discrepancy of over 0.5 SD. However, that observation must be tempered due to the low number of African-American (25) and Asian (22) students. Differences in the slope for Asian students from other students may be more a function of the "flatness" of that slope relative to the others.

[10] The label "general aptitude" may be unfair, as this implies that certain groups of students may be cognitively below other groups. Yet, the Z-scores of the 3rd grade tests have the greatest predictive facility for later test scores of any variable. The argument that consistent school effect from before 3rd grade through 5th or 6th grade is the major influence on scores is difficult to sustain when looking at the influence of SES, as measured by free-lunch status, on 3rd grade Z-scores. Students paying the full price for their lunches achieve significantly higher scores than those students receiving free lunches *in every school*. This is the case with both English and math scores—in every school full-price-lunch students outperform free-lunch students. Whether or not this relates directly to general aptitude is open to question—"test-taking ability" may be a more accurate term—but the argument that the K-3 school effect of every school is having a greater impact on one group of students over the other is difficult to accept. A more likely explanation is that this discrepancy results from factors independent of school effect.

Again, the low number of Asian students permits a few unusual results to unduly influence the slope of the Asian regression line[11].

Regression Lines for SES

A comparison of regression lines across socioeconomic groups reveals that, even controlling for general achievement, differences in overall achievement (reflected in differences in intercept) remain. $Science_5$ performance of poorest students (those on free lunch) consistently lagged behind other students. The unusual slope of the reduced-price regression line for Cohort 1 (Figure 5) is significant but difficult to interpret. A review of the test scores for these students suggests that the relatively low number of reduced-price students has allowed the poor performance on the science test by a few students scoring near or above the 1.0 level on the 3rd grade English test to unduly influence the slope of the line. (Removing the 4 students with the highest scores on the $English_3$ test from the regression analysis changes the slope of the regression line to .62, which is almost identical to the slope for full-priced-lunch students.) Despite the anomaly with the slope of the reduced-price regression line for Cohort 1, it is clear from the difference in intercepts between the free- and full-price-lunch regression lines for both Cohorts that SES does have a major impact on $Science_5$ achievement. (See Tables 8 and 9.)

Regression Lines for School Stability

For Cohort 1, the regression lines for school stability differ in slope and intercept and appear to converge toward the upper end of the scale (around 0.5 SD) indicating that the impact of geographic mobility may be more detrimental at the lower end of the

[11] A likely explanation is that several non-native Asian students have 3rd grade English scores depressed because of their lack of English proficiency. By the time these students have entered 5th grade their English skills have improved and have less of an attenuating effect on their test scores.

achievement scale. This compares to regression lines for ethnicity and socioeconomic status that seem to imply a consistent advantage or disadvantage.

However, Cohort 2 revealed no significant difference in slopes between the two groups, although the difference in intercepts was significant at the 0.001 level. In both cases school stability appears to have a positive impact on test scores, although the distribution of the impact across ability levels varies between cohorts.

T-tests of Additional Demographic Variables

A surprising result of the t-tests for IEP status is that the difference in intercepts between students with an IEP and non-IEP students is not statistically significant. This may be because general aptitude is already accounted for in the $English_3$ and $Math_3$ scores. The distribution of students with an IEP is truncated with a median score on, e. g., the $English_3$ test of-0.73. Thus, meaningful comparisons are problematic.

Impact on $Science_5$ Regression

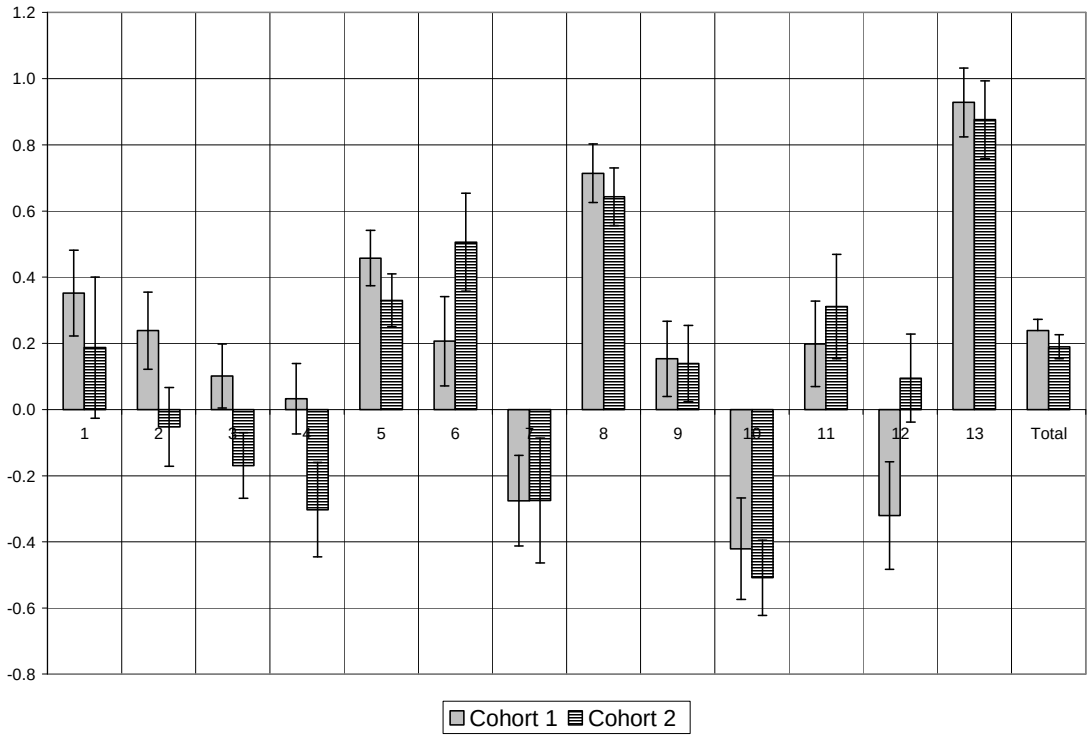
Of all the demographic variables, only SES was shown to have a consistent, statistically significant impact on $Science_5$ scores. Both ethnicity and school stability influences were accounted for—or at least subsumed—within SES. Adding SES to general aptitude accounted for a combined 57% of the variance in $Science_5$ achievement. That still leaves 43% of $Science_5$ achievement unaccounted for. The next issue to which the achievement data were addressed was whether this independent 43% of $Science_5$ achievement could be meaningfully disaggregated to be interpreted as the “school effect”. Quite clearly, significant differences in overall of $Science_5$ achievement were observed among schools, even when the influence of previous general achievement and socioeconomic status were statistically controlled.

Interpretation of Residuals

Figure 11 shows the mean $Science_5$ scores for students of Cohorts 1 and 2 according to school. This chart shows how the students in a particular school performed relative to the state mean. As shown in the graph, Schools 1, 5, 6, 8 and 13 all have $Science_5$ scores above expected performance, while Schools 7 and 10 were consistently below state averages. For some schools, Figure 9 (based on mean achievement controlled for general achievement and socioeconomic status of students) paints a picture of the school's effect on $Science_5$ achievement markedly different from that pictured in Figure 11. The gains registered by schools 5, 8 and 13 are more modest after accounting for the general aptitude and SES of the students, while School 6 still shows a positive school effect. School 10, which showed $Science_5$ scores below state means, nonetheless appears to show a school effect closer to the "expected" mean. Schools 7 and 12 however, appear to show a negative school effect on $Science_5$ scores.

Figure 11

Mean Science$_5$ Z-scores According to School Unadjusted for SES or Aptitude



Although this does not translate directly into teaching quality—other factors such as overall environment or quality of resources may also have an influence—it is probably reasonable to assume that teaching quality, defined as the ability of teachers to influence test scores[12], is a significant component of school effect. It should be noted, the particular classroom assignment of the students is not recorded with the test scores, individual classroom effects could not be determined.

[12] This is the de facto definition of the term. It is simply articulating assumptions inherent in the reliance on the exclusive use of standardized tests to draw conclusions about school quality for NCLB purposes. The author does not necessarily endorse this definition.

CHAPTER V

CONCLUSIONS, LIMITATIONS, FOLLOW UP AND RECOMMENDATIONS

The study sought to answer two questions about the 5-GIST:

1) Is there evidence of the construct validity of the test?

2) Can the test be used to make reasonable inferences about the quality of science instruction within schools?

The analysis appears to answer both questions in the affirmative, albeit with qualifications. Independent residual variance in 5-GIST test scores was uncovered after accounting for factors independent of teaching. Furthermore, this finding was consistent across cohorts. Some variability between cohorts is to be expected due to changes in teaching staffs from one year to the next, yet, as evinced by the similarity of the two regression formulae for the cohorts, repeatability appears robust. Assuming positive content validation, it would be reasonable to conclude that this unique variance is a reflection of the science content knowledge of the test-taker.

The variance in the between-school mean residuals supports the contention that the 5-GIST can be used to identify differences in the quality of science instruction offered by the schools. In the school corporation investigated in this study, two schools (#6 & #13) showed a positive effect on science scores of one of two cohorts in excess of 0.4 SD. Conversely, two schools (#7 & #12) showed scores of a cohort lagging more than 0.4 SD. Long-term tracking of school effect, particularly with schools making concerted efforts to improve science instruction, would be desirable to determine to what extent school effect can be improved.

However, it must be emphasized that this analysis is only a first step in the process ultimately leading to the use of 5-GIST scores in fairly and accurately assessing the impact of schools on student learning. There remain some reservations about the use of the test as the sole indicator of quality. First, substantial differences remain between what is measured in a paper and pencil test of knowledge and important objectives related to process skills among students. It has yet to be shown that the ability to perform active scientific processes such as inferring, investigating, or controlling and manipulating variables can be accurately assessed through a testing program that relies so heavily on reading and writing ability, (as evinced by the high correlation between $English_3$ and $Science_5$ test scores). Better controls are needed for assessing school variance; test scores provide only a limited picture of the impact of instruction on students, and that picture is often distorted by influences outside of schools' control.

Some notable influences on 5-GIST scores independent of instruction were revealed by the analysis. The single greatest influence was the general aptitude of the students as reflected by the scores achieved on the 3rd grade English and math tests. Ethnicity, SES and school stability were also shown to impact scores and were reflected in statistically significant differences in regression lines and/or intercepts. However, when considering the impact of these variables, it was discovered that accounting for SES alone, in combination with general aptitude, was sufficient to account for the effect of all three variables. Thus, if test results are disaggregated according to general aptitude and SES, further disaggregation by ethnicity and mobility may not be necessary.

IEP status did not prove to have a significant impact on test scores. Students with IEPs are usually given test accommodations, and it is possible that these accommodations

are sufficient to allow them to record achievement similar to their regular education peers of similar general ability and SES. However, it is important to remember that these results apply only to the 5-GIST and can not be taken as indicators for other subject areas. It would be interesting to see if IEP status has a significant impact on subsequent English and math test scores.

Limitations

The classification of school stability/mobility was determined by comparing the school at which the 3rd and 5th grade tests were taken. It is likely that many mobile students—e.g., those that moved between 1st and 3rd grade, or moved after taking the 3rd grade tests but then moved back to the same school before taking the 5th grade tests—were not classified as mobile. Hence, the impact of mobility on test scores may have been diluted in the analysis. The inability to identify a statistically significant student mobility effect on the 5-GIST should not be interpreted as proof that student mobility does not impact test scores, but rather, as a possible limitations of available data.

Although school effect does appear to be identifiable in this study, teacher effects are not. Data about the individual class assignments of the subjects, and hence, the teacher responsible for the science education of the subject, are not recorded and therefore not available for analysis. Variations in the effectiveness of science education within a school must of necessity be treated as error.

The school corporation participating in this study is largely suburban and rural, homogeneous and generally of higher SES. Replicating this study in other school corporations, particularly urban corporations or those with more diverse populations, would broaden generalizability.

The validation established by this study should not be construed as content validation. Although the study appears to identify a school effect on test scores, this does not confirm nor deny that the tests are an accurate reflection of the degree to which the test-takers have mastered the science concepts and process skills as mandated by the state science education standards. A separate content-validation study would be required to verify this.

Follow Up

Several follow up studies would be desirable. A similar study examining 5th grade English and math scores could be undertaken to test consistency of results across subjects. In particular, it would be interesting to see if the discovery that IEP status did not have a statistically significant impact on 5-GIST scores is found also with English and math tests. Repeating this study with a third cohort would help refine the regression formula and strengthen conclusions about the statistical significance of the difference between cohorts and between schools. Beginning in the spring of 2006, students in Indiana take a second science assessment in 7th grade. Comparisons of 5th and 7th grade science tests scores will be possible in the near future.

Using regression to identify exemplary and inadequate schools goes beyond the intentions of this study. However, a reasonable use of these results would be to investigate schools showing consistently positive school effects to see if uniquely effective educational strategies can be identified in those schools and reproduced in other schools.

Recommendations

Although unique between-school variance in scores was identified by this study, this should not be construed as an endorsement of the use of the 5-GIST as the sole instrument for the labeling of schools. While approximately 40-43% of the variance was attributed to school effect, the major influences of the test scores—accounting for between 57% and 60% of the variance—were general aptitude and SES. While such a large percentage of variance attributable to factors independent of instruction may not be an issue if the sole use of the test is to assess the science content knowledge of the test-taker—the use for which it is designed—it can obfuscate objective evaluation of the quality of instruction provided by an institution. If the results of the study are consistent across subject areas, then many schools may be unjustly sanctioned unless these outside influences are taken into consideration in the analysis. As was shown in this study, it is possible for schools to have a positive effect upon student science learning and still achieve results below state means.

The stated goal of the school reform movement is to improve the quality of instruction. The universal administration of a test, only partially sensitive to teaching influence, as the sole indicator of instructional quality affords only a limited view with which to pass judgment. It is hoped that the results of this study will initiate a reevaluation of the current use of the 5-GIST.

REFERENCES

American Association for the Advancement of Science (AAAS) (1993). Project 2061: Benchmarks for science literacy. New York: Oxford University Press.

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

Bolser, S. and Gilman, D. A. (2003). Saxon Math, Southeast Fountain Elementary School: Effective or ineffective? (ERIC Document Reproduction Service No. ED 474 537.)

Buechler, M. (1991). Constraints on teachers' classroom effectiveness: The teachers' perspective, policy bulletin. Educational Policy Center, Indiana University, Bloomington, IN. (ERIC Document Reproduction Service No. ED 361 302.)

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological bulletin, 56 (2), 81-105.

Cronbach, L. J., and Meehl, P. E. (1959). Construct validity in psychological tests. In Minnesota studies in the philosophy of science 1 (3rd ed.). Minneapolis: University of Minnesota Press.

Davis, H. S. (1998). Effects of absence and Cognitive Skills Index on various achievement indicators: A study of ISTEP scores, discrepancies, and school-based math and English tests of 1997-1998 seventh grade students at Sarah Scott Middle School, Terre Haute, Indiana. (ERIC Document Reproduction Service No. ED 423 302.)

Ding, C. S., and Davison, M. L. (2005). A longitudinal study of math achievement gains for initially low achieving students. Contemporary educational psychology, 30, 81-95.

Harris, A. N. and Gilman, D. A. (2003). Implementing the Shurley Method at Reelsville Elementary School to raise achievement scores: Effective or ineffective? (ERIC Document Reproduction Service No. ED 475 825.)

Indiana State Board of Education (2000 – 2001a). Indiana's academic standards: Grade 4, [On-line]. Available: http://www.doe.state.in.us/standards/Docs-2004/English/PDF/K-8/Grade04.pdf .

Jerome, K. and Gilman, D. A. (2003). Writing improvement programs: Does this type of intervention really work? (ERIC Document Reproduction Service No. ED 474 935.)

Maguire, T., Hattie, J and Haig, B. (1994). Construct validity and achievement assessment. The Alberta journal of educational research, 40(2), 109-126.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. Educational researcher, 18 (2), 5-11.

Messick, S. (1989b). Validity. In R.L. Linn, (Ed.), Educational measurement, 3rd edition. New York: American Council on Education/Macmillan Publishing Company.

National Research Council (1996). National science education standards. Washington, D.C.: National Academy Press.

Popham, W. J. (19787). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Rescorla, L. and Rosenthal, A. S. (2004). Growth in standardized ability and achievement test scores from 3rd to 10th grade. Journal of educational psychology, 96(1), 85-96.

Rulon, P. J. (1946). On the validity of educational tests. Harvard educational review, 16, 290-296.

Russell, M., Higgins, J. and Raczek, A. (2004). Accountability, California style: Counting or accounting? Teachers college record, 106,(11), 2102-2127.

State of Indiana Department of Education and CTB/McGraw-Hill LLC. (2003). The grade 5 science: Guide to test interpretation. Monterey, CA: The McGraw-Hill Companies, Inc.