LEADERSHIP AND POLICY STUDIES

ACHIEVEMENT TRADEOFFS AND NO CHILD LEFT BEHIND

MATTHEW G. SPRINGER

Dissertation under the direction of Professor James W. Guthrie

Despite speculation that the No Child Left Behind Act of 2001's (NCLB) finely tuned attention to improving academic opportunities for traditionally low-performing students and student subgroups compromises educational opportunities of high-performing students, there is limited empirical evidence that NCLB actually inhibits the progress of high-performing students. Consequently, ideological predispositions have dominated public interest in distributional effects under NCLB. A Student X Subject general linear model with school and Year X Grade fixed effects is estimated to isolate whether a school, based on prior year's performance, has targeted resources to (a) students in a failing subgroup, (b) students in a failing subject, and/or (c) students failing math on a failing subgroup in Idaho. There is strong evidence that NCLB's threat of sanctions increased incentives for schools and school districts to elevate learning opportunities for traditionally low-performing students and student subgroups, but that the increased performance by traditionally low-performing students and student subgroups did not occur at the expense of traditionally high-performing students. It appears that Idaho's response to NCLB is one of improved efficiency and not

achievement tradeoffs, in that traditional public schools in the state did more with the same level and distribution of resources as in years past.

Approved_____Date_____

ACHIEVEMENT TRADEOFFS AND NO CHILD LEFT BEHIND

By

Matthew G. Springer

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Leadership and Policy Studies

December, 2006

Nashville, Tennessee

Approved

Professor James W. Guthrie

Professor R. Dale Ballou

Professor Michael J. Podgursky

Professor Kenneth K. Wong

To my wife, Joy, and my son, Sam.

ACKNOWLEDGEMENTS

Financial assistance from the Smith Richardson Foundation and Vanderbilt University's Peabody Center for Education Policy is gratefully acknowledged.

I am grateful for the guidance and helpful suggestions of my dissertation committee members, James W. Guthrie, R. Dale Ballou, Michael J. Podgursky, and Kenneth K. Wong. I am also indebted to my brother, Jeffrey A. Springer, for his time and attention to detail in the preparation of this research. Additionally, I would like to express my appreciation to Eric A. Houck for his insights and comments.

Finally, I would like to thank my family, and especially my wife and my son, for their encouragement, patience, and understanding. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

This dissertation investigates the strategic resource-allocation decision making of traditional public schools and public charter schools seeking to close the achievement gap among low- and high-performing students in response to the No Child Left Behind Act of 2001 (NCLB), as well as factors hypothesized to explain this decision making.

NCLB is the reauthorization of the nation's omnibus Elementary and Secondary Education Act of 1965 (ESEA). NCLB is a sweeping amalgam of legislative elements and prior reform ingredients recast into a blueprint for standards-based reform. The central purpose of NCLB is that all traditional public school students, and defined student subgroups thereof, reach academic "proficiency" by the 2013-2014 academic year. NCLB monitors progress toward meeting academic "proficiency" through Adequate Yearly Progress (AYP) calculations, a series of minimum competency performance targets that must be met by schools and school districts to avoid sanctions of increasing severity. In theory, NCLB's threat of sanctions increases incentives for schools and school districts to elevate learning opportunities for traditionally low-performing students and student subgroups.

However, some assert that NCLB's finely tuned attention to improving academic opportunities for traditionally low-performing students and student subgroups compromises educational needs and opportunities of high-performing, academically accelerated students (Colangelo, Assouline, & Gross, 2004a, 2004b; Davidson &

Davidson, 2005; Gallagher, 2004; Renzulli, 2005; Ruf, 2005; Sausner, 2005). In a two-part report sponsored by the John Templeton Foundation, Colangelo et al. (2004a) argued that NCLB ignores high-performing students to the detriment of their expected learning trajectories:

> Schools pay lip-service to the proposition that students should learn at their own pace; in reality, for countless highly able children the pace of their progress through school is determined by the rate of progress of their classmates. . . . [T]he No Child Left Behind legislation, which aims to bring all children up to proficiency, is the national focus on education. This is an admirable goal and worthy of our efforts. However, NCLB ignored those students who are well above proficiency, and these students are also worthy of our best effort. (pp. 1-2)

Despite speculation that high-achieving students have been "deceived" and "denied" as a result of NCLB (Sausner, 2005, p. 1), there is limited empirical evidence that NCLB's minimum-competency standards actually inhibit the progress of high-performing students. Furthermore, there is scant scientific understanding of factors that may explain strategic resource-allocation decision making in response to minimum-competency accountability systems. Consequently, ideological predispositions have dominated burgeoning public and scholarly interest in distributional effects under NCLB's accountability system.[1]

As Congress approaches reauthorization of NCLB in 2007, it is vital to quantify the frequency and magnitude of achievement tradeoffs occurring under NCLB and the processes by which resource-allocation decisions are undertaken. Identification will help policy makers determine whether gains of marginal-performing students are indeed

[1]For instance, advocates contend that accountability policies will make teachers and schools more effective, thus ensuring high-quality instructional curricula for all students. Opponents, on the other hand, argue that accountability policies will lead to a "leveling" effect from a "one-size-fits-all" curriculum. For a more thorough analysis of potential effects of high-stakes testing delineated by students, teachers, administrators, and policymakers, see Stecher (2002) and Koretz, McCaffrey, and Hamilton (2001).

occurring at the expense of high-performing students and, consequently, whether present accountability policies should be modified.[2]

Research Questions

This dissertation analyzes both longitudinal, student-level test score and school-level accountability data from Idaho. Student-level test score data were furnished by the Northwest Evaluation Association (NWEA). School-level accountability data were collected from the Idaho State Department of Education. These data were then used to answer the following research questions:

1. Have Idaho schools responded to NCLB by raising the achievement of failing students relative to the achievement of nonfailing students?

2. Have Idaho schools responded to NCLB by raising the achievement of failing subgroups relative to the achievement of nonfailing subgroups?

3. Have Idaho schools responded to NCLB by raising the achievement of failing students at the expense of high-performing students?

4. Do other programmatic features of NCLB, such as severity of sanctions, overlapping student subgroups, and high-stakes testing in terminal grades, better explain strategic resource-allocation decision making?

5. Does increased market sensitivity in public charter schools impact resource-allocation practices?

---

[2]The United States Department of Education is exploring use of growth model calculations to determine Adequate Yearly Progress. Earlier this year 20 states submitted proposals to participate in a pilot growth model program. Although eight of these proposals made it to a second round of review (i.e., Alaska, Arkansas, Arizona, Delaware, Florida, North Carolina, Oregon, and Tennessee), North Carolina and Tennessee were the only states selected to participate.

Defining Strategic Resource-Allocation Decisions

Investigating strategic resource-allocation decision making by schools seeking to close the achievement gap requires a measure that captures systematic shifts in intraschool resource distribution. Recognizing that no formal accounting system tracks allocation of resources at the student or classroom level,[3] distributional inequities in student achievement among low- and high-performing student achievement are used to infer a reprioritization of intraschool resources.

An NCLB-induced resource-allocation decision is detected if a greater than expected increase in the achievement of traditionally low-performing students occurs in tandem with a less than expected increase in the achievement of traditionally high-performing students. The term *low-/high-achievement tradeoff* is used throughout this dissertation to imply a strategic resource-allocation decision made by a school in response to Idaho's high-stakes accountability program.

Reliance on distributional inequities to infer strategic resource-allocation decision making in response to high-stakes accountability assumes implicitly that there is a resource constraint on schools. A resource constraint implies, in effect, that elevating the performance of traditionally low-performing students necessitates that schools give up something, somewhere else. If Idaho public schools indeed operate consistently within this zero sum view, then examining the distribution of student achievement to infer strategic resource-allocation decision making is fitting.

[3]For a more complete discussion of data deficiencies in contemporary educational research, practice, and policy see Guthrie (2006). Guthrie argues that the next generation of education reform necessitates design and implementation of a comprehensive national data system linking inputs, throughputs, and outcomes at the student, school, and classroom levels.

However, if schools operate in the absence of resource constraints, then it is difficult to infer tradeoffs. Theoretically, schools unconstrained by resources are capable of elevating the outcomes of low-performing students without diverting attention and resources away from other students. Conversely, schools simply may become more efficient, in that they respond to systemic incentives by doing more with the same level and distribution of resources as in years prior. In light of these potential confounding scenarios, alternative explanations will remain to be explored if this dissertation finds no evidence of achievement tradeoffs.

Outline

This dissertation is divided into six chapters. Chapter II provides a summary review of relevant literature. Although considerable interest and controversy surround achievement tradeoffs in high-stakes accountability programs, surprisingly little empirical research has addressed the issue of achievement tradeoffs using student-level achievement data. To date, most scholarly research has examined the association between (a) accountability programs and mean achievement growth; (b) accountability programs and school, classroom, or teacher behavior and practices indirectly linked to student-level achievement data; or (c) accountability programs and system gaming (e.g., teachers altering test scores and/or assisting students with test questions). Chapter II concludes by discussing why further research is warranted.

Chapter III describes Idaho's Assessment and Accountability Program (ISAAP). ISAAP's genesis was a set of content and achievement standards established in 1994, and subsequently revisited in 2000 and 2001, by the Idaho legislature, State Board of

Education, and Citizen's Commission on Assessment and Accountability. ISAAP complied with federal guidelines and regulations associated with NCLB as of 2003. Chapter III also characterizes resource-allocation decision making by schools in response to ISAAP by borrowing insight from economic and psychological theory and by building upon education domain-specific theory and empiricism found in pre-NCLB research.

Chapter IV describes data sources, data development, and lastly the basic estimation strategies employed in measuring the presence and magnitude of achievement tradeoffs and in examining several mediating factors that may help explain strategic resource-allocation decision making. Indicators of interest include whether (a) a school has failed to meet state-prescribed proficiency, (b) a student is part of a student subgroup that failed to meet state-prescribed proficiency, and (c) a student has failed to meet state-prescribed proficiency in the subject for which a school also failed. Identified mediating factors that may explain further why one student category is rewarded at the expense of another include whether (a) a student is part of more than one student subgroup held accountable to the state's minimum competency proficiency targets (e.g., economically disadvantaged and black), (b) a student is in a terminal grade, and (c) differences in market sensitivity across traditional public schools and public charter schools.

Chapter V reports results from a series of general linear models used to estimate whether and, if so, how schools are responding to Idaho's minimum competency accountability program. The dependent variable across all reported models is a fall-to-spring student gain score in mathematics as measured by the Idaho State Assessment Test (ISAT). Results consistently indicate positive and statistically significant gains for traditionally low-performing students. These results persist after (a) adjusting regressors

for meaningful variation between grades in average student performance gains and (b) standardizing the dependent variable in an effort to gauge the potential influence of mean reverting measurement error on estimates. There is strong evidence that NCLB's threat of sanctions increases incentives for schools and school districts to elevate learning opportunities for traditionally low-performing students and student subgroups. Moreover, the increased performance by traditionally low-performing students and student subgroups do not occur at the expense of traditionally high-performing students. Indeed, results indicate that NCLB's threat of sanction has increased the efficiency in which traditional public schools operate in Idaho.

CHAPTER II

REVIEW OF RELEVANT LITERATURE

Although considerable interest and controversy surround the issue of achievement tradeoffs in high-stakes accountability programs, surprisingly little empirical research has addressed the issue using student-level achievement data. This deficiency is due in large part to data limitations and to the fact that minimum competency accountability programs were not as potent a feature of state public education systems until federal enactment of NCLB in 2002.[4]

To date most scholarly research has examined the association between (a) accountability programs and mean achievement growth; (b) accountability programs and school, classroom, or teacher behavior and practices indirectly linked to student-level achievement data; or (c) accountability programs and system gaming (e.g., teachers altering test scores and/or assisting students with test questions). As such, and as the following summary review of literature concludes, advent of NCLB's accountability structure and the unique policy conditions present in states not yet rigorously examined

[4]Generally speaking, accountability programs have existed in many forms for a long time (Kirst, 1990; Resnick, 1982; Stecher, Hamilton, & Naftel, 2005). However, it was not until codification of NCLB's reform agenda and its rigid incentive system and ambitious performance targets that the public education landscape was truly characterized by high-stakes accountability. Even though accountability programs were implemented in approximately 80% of states prior to NCLB, only 14 states used student performance measures to assign discrete grades or ratings to all schools and/or school districts prior to NCLB (Reback, 2006). Furthermore, only 17 states ever fully complied with the Elementary and Secondary Education Act of 1994 directives (Wanker & Christie, 2005). Yet, by the close of NCLB's first year of implementation, all 50 states were in compliance with NCLB policies.

warrant further research on the relationship between achievement tradeoffs and high-stakes accountability programs.

Accountability Programs and Mean Achievement

Prior studies have modeled whether statewide accountability systems altered predicted math and reading achievement, as measured by the National Assessment of Educational Progress (NAEP) or by a combination of outcome measures, including NAEP, ACT and SAT (two college placement tests), and Advanced Placement (AP) scores.

Carnoy and Loeb (2002) constructed an index of the intensity of the accountability mechanism in all 50 states in order to analyze whether the strength of accountability mechanisms was systematically related to mean mathematics gains, as measured by NAEP. Using data from 1996 to 2000, and after adjusting analyses for changing inclusion rates of special education and limited English proficiency students, they found a positive and significant relationship between NAEP math scores and strength of accountability mechanisms across all race/ethnicity categories at the eighth-grade level and for African-American fourth graders. In substantive terms, Carnoy and Loeb concluded that a two-step increase on their 0-to-5 accountability scale was associated with an approximate one half a standard deviation higher gain in the percentage of students that achieved a proficiency rating of at least basic.

Hanushek and Raymond (2005) modeled whether introduction of a statewide accountability system altered predicted math and reading gains using a sample of

approximately 40 states.[5] A state's accountability system was considered "introduced" if the state published outcome information aggregated at the school level and provided a means by which this information could be interpreted. Drawing upon two cohort observations from NAEP to estimate math and reading growth, Hanushek and Raymond concluded that pre-NCLB accountability systems led to improved student performance in the aggregate. However, their work concluded that pre-NCLB accountability instruments were not effective in closing achievement disparities by race.

In response to significant growth on statewide assessments and NAEP mathematics assessment from 1990 to 1997, Grissmer and Flanigan (1998) conducted an in-depth case study of North Carolina and Texas education systems to better understand test score gains. Their study reported that significant gains in student achievement could not be attributed to increased school spending, reduced student-teacher ratios, or increased levels of teacher training. Instead, Grissmer and Flanigan attributed achievement gains to state-level accountability programs closely linked to academic standards.

Conversely, in a set of controversial studies,[6] Amrein and Berliner (2002a, 2002b) found no effect, or in some cases a negative effect, of accountability structures on student achievement gains. In their first study, Amrein and Berliner (2002a) identified 18 states whose public school accountability policies were characterized by the "most severe consequences." They then modeled the impact of these programs on several outcome measures, including NAEP, ACT and SAT, and AP scores. Amrein and Berliner

[5]Reported sample size ranged from 38 to 42 states and 68 to 138 observations.

[6]For more information on the controversy surrounding the Amrein and Berliner papers see Amrein-Beardsley and Berliner (2003), J. P. Greene and Forster (2003), Raymond and Hanushek (2003), Thompson (2003), and Winter (2002).

concluded that these states' accountability mechanisms exerted no positive effect on student outcomes. Moreover, in a follow-up study that included an additional 17 states, Amrein and Berliner again reported no positive effect of public school accountability policies on student outcomes.

Accountability Programs and School, Classroom, or Teacher Behavior and Practices

A second strand of prior research on educational accountability involved use of survey and observational data to study the impact of accountability structures on school, principal, and teacher behavior and practices. In a case study of four Chicago elementary schools, two of which were high-performing and two low-performing, Diamond and Spillane (2004, p. 1148) examined differential responses by teachers and administrators in terms of (a) responses to incentive structures, (b) interpretation and use of test score data, and (c) setting of instructional priorities. They concluded that low-performing schools on probation focused on raising the performance of students within student subgroups considered failing, whereas high-performing elementary schools appeared to allocate time and energy equally across all students and grades.

In 2002, Ladd and Zelli reported results from surveys administered in 1997 and 1999 to a random stratified sample of approximately 70 North Carolina elementary school principals. These surveys examined principals' behavioral responses to their respective state's accountability program. Ladd and Zelli reported that principals responded by redirecting resources to math and reading, integrating math and reading into other courses and into extracurricular activities, increasing time spent with teachers to improve instruction, and shifting attention from high- to low-performing students in an

effort to avoid probationary status (p. 54). Furthermore, principals reported that their respective program's incentive structure made it even more difficult to fill teacher and administrative vacancies in traditionally "hard-to-staff" schools.

Koretz, Barron, Mitchell, and Stecher (1996) conducted a large-scale evaluation of the Kentucky Instructional Results Information System (KIRIS), finding that teachers operating under the state's mandated accountability program elevated their expectations of high-performing students, but not necessarily of low-performing or special need students. Their results were based on surveys administered to random representative samples of fourth-grade teachers, eighth-grade mathematics teachers, and fourth- and eighth-grade principals across Kentucky, as well as interviews with 115 principals and 216 teachers in the state. A subsequent study by Koretz and Barron (1998, p. 1) reported that student gains in response to Kentucky's accountability program could be attributed to "item-specific coaching tailored to reused items" on high-stakes assessments, and that these gains were sufficiently sizable in ensuing years to discount somewhat the argument that test familiarization, and not the accountability program's incentive structure, had spurred the vast and sustained growth in student achievement witnessed over the period under examination.

Accountability Programs and System-Gaming

Presently, a rapidly growing strand of research is investigating whether schools respond to accountability programs by "gaming" the system. Studies have documented schools gaming accountability systems in a spectrum of manners, including (a) focusing excessively on a single test and altering test scores and/or assisting students with test

questions (Goodnough, 1999; Koretz et al., 1996; Jacob & Levitt, 2003), (b) strategically reclassifying students as special education and limited English proficiency (Cullen & Reback, 2002; Deere & Strayer, 2001; Figlio & Getzler, 2002; Jacob, 2005), (c) using discipline procedures to ensure that low-performing students will be absent on test day (Figlio, 2005), (d) manipulating grade retention policies (Haney, 2000; Jacob, 2005), (e) misreporting administrative data (Peabody & Markley, 2003), (f) acquiescing in parents' demands for test exemptions and waivers (Neufeld, 2000), and (g) planning nutrition-enriched lunch menus prior to test day (Figlio & Winicki, 2005).

Jacob and Levitt (2003), for example, developed a statistical algorithm to detect unexpected fluctuations in students' test scores and unusual patterns of answers for students within a classroom. The validity of their instrument was confirmed through a series of "retesting" experiments. When applied to Iowa Test of Basic Skills (ITBS) results from the Chicago Public School System from 1993 through 2000, Jacob and Levitt found that the likelihood of cheating in a classroom whose prior performance was one standard deviation below the mean increased by roughly 29% in response to accountability policies.

Using detailed student-level panel data and controlling for student-level fixed effects, Figlio and Getlzer (2002) examined whether initiation of the Florida Comprehensive Assessment Test (FCAT) influenced public schools' decision making on special education assignments. They found that low-performing students and students from low socioeconomic backgrounds were significantly and substantively more likely to be reclassified into exempted disability categories and that this practice was significantly more likely to occur in high-poverty schools than in their more affluent counterparts.

This evidence suggested that Florida public schools might have responded to the state's accountability program by reclassifying certain students in order to reduce the contribution of these students to the aggregate measures of test performance upon which schools were judged.

Accountability Programs and Achievement Tradeoffs

With the exception of one study from NCLB's first year of operation, previous research on the effect of accountability programs on student outcomes have focused exclusively on pre-NCLB accountability programs.

In reaction to extraordinary student progress reported by the state of Texas over the period between 1993 and 1999, Deere and Strayer (2001), for instance, analyzed achievement data for 3rd- through 8th- and 10th-grade students nested in approximately 4,290 traditional public school campuses.[7] They found that students' passing rates on high-stakes accountability tests in math, reading, and writing increased substantially relative to those on low-stakes tests in social studies and science. Furthermore, when conditioned on initial score, the average test score increment in math or reading (t+2 – t+1) decreased from low- to high-performing student subgroups. That is, students clustered in the lowest sextile had the greatest average score increment (M = 4.12; r =

[7]Mean campus Texas Accountability and Assessment System (TAAS) score pass rates "skyrocketed" from 1993 to 1998 in math, reading, and writing. Deere and Strayer (2001) noted that, "Along with a 13% increase in the average math score and a 7% increase in the average reading score, passing rates rose sharply over the period. Between 1994 and 1999, the fraction of all students who passed the math test increased from 58% to 85%, and the fraction who passed the reading test increased from 74% to 86%. Over this period the fraction of schools that received the 'exemplary' designation (the highest accountability rating) rose from 1.3% to 11.7% as passing rates increased. Likewise, the 'recognized' rating (second highest rating) was given to 36.8% of school in 1999, up from 13.1% in 1994" (p. 2). Deere and Strayer's dataset included approximately 1.7 million observations per year, or an approximate total of 8.5 million student-year observations for the period encompassing the 1993 through 1999 school years.

4.67), and students in the highest sextile had the smallest average score increment (*M* = 1.48; *r* = 1.27). Even after taking into consideration regression to the mean, Deere and Strayer found that students at or below the passing level exhibited the greatest improvement on tests.

Holmes (2003) analyzed achievement data for third- through eighth-graders in North Carolina to examine the distributional effects of accountability programs within traditional public schools. Holmes advanced Deere and Strayer's (2001) work by constructing an indicator variable that measured teachers' short-run incentive to target instruction to particular students. Holmes (2003) concluded that North Carolina's incentive program led to distributional inequities, in that students with greater potential to influence a school's accountability rating from a given increase in test scores were targeted with more resources than those students whose expected test score growth was less relevant under the system's incentive structure. On average, achievement gains for "targeted students" were 1.47 to 3.79 points greater in reading and 1.25 to 2.62 points greater in math during the 1999-2000 school year.

Relying on student-level panel data generated from the Texas Assessment of Academic Skills (TAAS), Reback (2006) explored more subtle intricacies of potential achievement tradeoffs occurring under Texas' Accountability Plan. Reback calculated the effect exerted by an improvement in the scores of certain students or student subgroups on the probability a school would improve its accountability rating. Using marginal effects as indicator variables in equations predicting actual test score gains, Reback found that a one standard deviation increase in his incentive measure was associated with approximately a .007 standard deviation increase in a student's place in the statewide

achievement distribution. His study concluded that marginally performing students ultimately exhibited the most improvement, while low- and high-performing students gained less than peers in schools where incentives to elevate the performance of marginally performing students were weaker.

Booher-Jennings' (2005) case study of an urban elementary school located in a low-socioeconomic neighborhood in Texas explored "how" and "why" teachers responded to the state's accountability program during NCLB's inaugural year. Booher-Jennings found teachers in the school practiced a form of "educational triage" in which resources were focused on students close to the passing score to the detriment of peers farther from the proficiency threshold. Students close to the threshold were pulled out for small group tutoring sessions with the lead teacher for approximately 90 minutes each day, while the rest of the class was given seat work under the supervision of local college students hired by the school. Booher-Jennings described the observed targeting of individualized instruction as a dynamic process in which pass-rate thresholds were re-defined by teachers and administrators according to achievement test results reported by the state to the school.

Finally, Chakrabarti (2006) examined whether public schools facing voucher threats under Florida's Opportunity Scholarship Program (OSP) behaved strategically. OSP provided families taxpayer-funded tuition payments for their children to attend another academic institution, public or private, if their present public school was identified as "failing."[8] Chakrabarti reported that failing public schools responded by focusing additional attention on low-performing students, but that increased performance

[8]In January 2006, the Florida Supreme Court in a 5-2 opinion declared the Opportunity Scholarship Program unconstitutional because voucher recipients could use taxpayer dollars for private education.

by low-performing students did not come at the expense of high-performing peers, as evidenced by a decline in high-performing peer achievement of consistently less than 1% per annum over the period under study. Chakrabarti also detected strategic behavior whereby failing schools seeking to avoid sanction focused additional attention on the FCAT writing test as the easiest of the high-stakes assessments for students to pass.

Why More Research is Needed

Additional research on the impact of high-stakes accountability programs on achievement tradeoffs is warranted for several reasons. First, evidence on accountability programs resulting in tradeoffs that negatively affect high-achieving students is decidedly mixed. Studies of pre-NCLB accountability programs in Texas and North Carolina suggested the presence of achievement tradeoffs working to the detriment of traditionally high-performing students, while evidence from Florida indicated that elevated achievement of low-performing students did not come at the expense of high-performing peers' estimated gains.

Second, prior research suggests that idiosyncratic features of state accountability policies might differentially impact teacher and school behavior (Swanson & Stevenson, 2002). Consequently, it is likely that findings from one state may not replicate necessarily in another.[9] For instance, NCLB permits considerable latitude across states in the level of

[9]Even though Booher-Jennings conducted her research during the 2002-03 school year, the Texas Accountability Program differed substantively from NCLB accountability regulations and mandates. Booher-Jennings (2005, pp. 260-261) noted that Texas' treatment of special education students, the minimum number of students needed when reporting subgroup results, and inclusion of mobile students in AYP calculations did not conform to NCLB mandates.

academic achievement that constitutes "proficiency."[10] In a Thomas B. Fordham Foundation report that evaluated accountability programs in 30 states, Cross, Rebarber, and Torres (2004) reported significant variation in the standards states require for a student to be labeled "proficient." Kingsbury (2003), too, found significant variance in their evaluation of proficiency standards in 14 states, noting in particular that differences among state standards reflected "dramatically different visions of what it means [for a student] to be proficient" (p. 13). Recognizing that NCLB provides states considerable autonomy in the design of their respective accountability programs, one might also expect intrastate variation in achievement tradeoffs due to variation in state prescribed proficiency targets[11] and minimum student subgroup sizes (or *n* sizes),[12] as well as to relative penetration of marketplace competition from charter and private schools.

Third, distributional effects in the context of NCLB's long-term agenda may be different from those elicited in Florida, North Carolina, and Texas. NCLB requires that all public school students, and defined student subgroups thereof, reach academic "proficiency" by the 2013-2014 school year. This target is significantly more ambitious

[10]For instance, NCLB guidelines simply state that, "Each State shall demonstrate that the State has adopted challenging academic content standards and challenging student academic achievement standards that will be used by the State" (P.L. 107-110, Section 1111(b)(1)(A)). NCLB also permits flexibility in the minimum size of subgroups, selection of factors used to calculate AYP, determination of targets for incremental improvement, and definition of full academic year.

[11]D. R. Greene, Trimble, and Lewis (2003) offered an interesting analysis of the effect of multiple standard-setting procedures employed by the Kentucky Department of Education (2001). Furthermore, according to an Education Commission of the States (2004) report, the percentage of schools failing to make AYP in the 2002-03 school year varied from a low of 8% in Minnesota to a high of 87% in Florida. The report indicated that these disparities are due in large part to variation in standards and proficiency levels across states.

[12]For a thorough discussion on the role of minimum *n* sizes and state accountability system design see deputy secretary Raymond Simon's (2006) testimony before Committee on Education and the Workforce entitled, *No Child Left Behind: Disaggregating Student Achievement by Subgroups to Ensure All Students Are Learning*.

than that set by Texas's Accountability Program and North Carolina's ABC Accountability Plan. Additionally, consequences for schools that fail to meet AYP are much more severe under NCLB than those first imposed by precursor programs in Texas, North Carolina, and Florida. For these reasons, distributional effects resulting from NCLB may be greater than those detected by previous studies, necessitating requantification of potential tradeoffs within the context of NCLB.

Fourth, there is considerable variation in the types of strategies for improving failing schools identified by districts. These strategies likely shape, in turn, the nature of school, principal, and teacher responses. For example, in 2005 the Center for Education Policy identified 18 different strategies in use by districts to improve failing schools, including: (a) use of student achievement data to inform instruction and other operational decisions, (b) reallocation of resources to support school improvement, (c) hiring of additional teachers to reduce class size, and/or (d) use of best practice research to inform decisions about improvement strategies (Center for Education Policy, 2005, p. 15). Some of these strategies may be considerably less vulnerable to the type of strategic resource-allocation decision making that works against particular groups of students.

Finally, present debate over achievement tradeoffs and NCLB tends to be informed by hyperbole and unfounded assertion. With congressional proceedings for NCLB's reauthorization scheduled for 2007, the time is opportune to provide policy makers with sound empirical evidence as they seek to determine whether present accountability rules should be modified to reward gains across a wider spectrum of achievement, and whether gains of traditionally low-achieving students are occurring at the expense of other students.

CHAPTER III

IDAHO'S STATEWIDE ASSESSMENT AND ACCOUNTABILITY PROGRAM

How schools respond to high-stakes minimum-competency accountability programs resides at the heart of the NCLB debate. Do schools behave strategically to incentives built into NCLB? Do decision makers allocate resources disproportionately to students on the margin of passing? Do these strategic resource allocation decisions occur at the expense of students at either tail of the achievement distribution? Do certain features of accountability better explain achievement tradeoffs than others?

Despite the intensity and gravity of the NCLB debate as Congress approaches reauthorization proceedings in 2007, the resource allocation decision making of K-12 public schools in response to strong incentives imbedded in high-stakes accountability programs continues to lack a well-defined theoretical foundation. Furthermore, the idiosyncratic nature of NCLB implementation across states adds an additional layer of complexity to efforts to define and quantify the relationship between incentive structures and achievement tradeoffs. As means to bring some clarity to this dynamic, this chapter examines the mechanics of Idaho's Statewide Assessment and Accountability Program before briefly characterizing resource-allocation decision making by Idaho traditional public schools and public charter schools in response to ISAAP.

Idaho's Statewide Assessment and Accountability Plan

Designed to meet federal guidelines and regulations associated with NCLB, ISAAP was approved by the United States Department of Education (USDE) on June 10, 2003. ISAAP's genesis was a set of content and achievement standards established in 1994, and subsequently revisited in 2000 and 2001, by the Idaho legislature, State Board of Education, and Citizen's Commission on Assessment and Accountability.

ISAAP holds all public schools[13] in Idaho accountable on the following three dimensions: (a) proficiency scores in math, (b) proficiency scores in reading, and (c) minimum participation rates in testing. Under the system, schools must meet or exceed math and reading performance thresholds and minimum participation rates in and across 10 student subgroups to avoid sanction.[14] ISAAP's minimum *n* for accountability purposes is 34 students.

Table 1 displays ISAAP's annual proficiency goals and the intermediate incremental percentage increase in the aggregate required of schools to reach 100% proficiency by the 2012-13 school year. During the 2002-03 school year, for instance, schools needed a minimum of 66% and 51% of students in each student subgroup to score proficient or better in reading and math, respectively, in order to make Adequate Yearly Progress. Annual proficiency goals for schools are calculated using a uniform averaging procedure across grade levels in a school. Schools are not held accountable for

[13]Idaho public schools are defined as those elementary and secondary schools established and maintained at public expense.

[14]The 10 subgroups include: (a) all students, (b) African American, (c) Asian, (d) American Indian/Alaskan Native, (e) Hispanic, (r) White, (g) Hawaiian/Other Pacific Islander, (h) students with disabilities, (i) limited English proficient, and (j) economically disadvantaged.

Table 1

Annual Proficiency Goals and Intermediate Incremental Increase Required to Reach 100% Proficiency by the 2012-13 School Year

Goals	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13
Reading											
Annual	--	3	3	3	3	3	3	4	4	4	4
Intermediate	66	66	72	72	78	78	84	84	92	92	100
Math											
Annual	--	4	5	5	5	5	5	5	5	5	5
Intermediate	51	51	60	60	70	70	80	80	90	90	100

student and student subgroup performance in individual grades. Furthermore, a minimum of 95% of students in each subgroup must be tested to avoid sanction.

During the 2003-04 school year, Idaho entered into a flexibility agreement with the USDE whereby any school missing the 95% participation target would determine an adjusted participation rate by averaging its 2003-04 school year rate with that of the prior year. If the 2-year average was 95% or greater, the participation rate target for 2003-04 was met. For the 2004-05 school year, and all years thereafter, a school seeking to average participation rate across years must include the three most recent years of data in their calculation (Idaho State Board of Education, 2006).

ISAAP also holds all public schools accountable on a fourth dimension, generically labeled the “third indicator” by NCLB. For the 2002-03 and 2003-04 school years, Idaho’s third indicator was tied to proficiency in language usage for all students aggregated at the school level, but not by subgroup. Beginning in the 2004-05 school year, however, the language usage indicator was replaced by one of the following three

options chosen at the discretion of the school district: (a) increase percentage of students scoring advanced; (b) decrease percentage of students scoring below basic; or (c) demonstrate growth in student performance, as measured by a predetermined computerized remediation program (Idaho State Board of Education, 2006). The first two options apply both to math and to reading tests of the Idaho Standards Achievement Test (ISAT), whereas the third option is specific to the remediation program selected by the district.

ISAAP measures student content knowledge and skill using an internet-enabled testing system called the Idaho Standards Achievement Test, which was developed by the Northwest Evaluation Association (NWEA), a national testing firm based out of Lake Oswego, Oregon. ISAT evaluates students in reading, math, and language arts and is scored on a single cross-grade and equal-interval scale ranging from 150 to 300 using the Rasch Unit (RIT) methodology (Cronin, Kingsbury, McCall, & Bowe, 2005). ISAT is administered to all public school students in grades 2 through 10 in the fall and spring.[15] Although scores are equated across grades, they are not equivalent across subjects.

Spring ISAT results are compared to grade-specific benchmark scores to gauge whether a student, student subgroup, and school met ISAAP's minimum proficiency standards. Both the student score required to be considered proficient and the percentage of students in a school needed to avoid ISAAP sanctions are delineated by content, grade, and level in Table 2.

Figure 1 provides a brief description of ISAAP's four prescribed levels of proficiency (i.e., below basic, basic, proficient, and advanced). A student must score at

[15]For information about the reliability, validity, and alignment of NWEA assessments see Northwest Evaluation Association's (2004) report *Reliability and Validity Estimates: NWEA Achievement Level Test and Measures of Academic Progress* and Kingsbury (2003).

Table 2

Approved Idaho Standards Achievement Test (ISAT) Proficiency Scores

	Grade								
ISAT	2	3	4	5	6	7	8	9	10
Reading									
Basic	174	185	192	198	203	207	210	213	216
Proficient	182	193	200	206	211	215	218	221	224
Advanced	193	204	211	217	222	226	229	232	235
Language Arts									
Basic	176	186	193	200	204	207	211	213	214
Proficient	184	194	201	208	212	215	219	221	222
Advanced	197	207	214	221	225	228	232	234	235
Math									
Basic	174	185	194	202	208	214	222	229	231
Proficient	185	196	205	213	219	225	233	240	242
Advanced	201	212	221	229	235	241	249	256	258

Note. Districts are permitted to impose more stringent proficiency standards. However, all districts are currently following these accountability standards.

the “proficient” level or better to be considered passing. The math statewide percentage of proficient and advanced scoring students in grades 3 through 8 and 10 has ranged from a low of 53% for eighth-graders in 2003 to a high of 90.3% for fourth-graders in 2005. When pooling across grades and years approximately 75% of students score proficient or better on the spring math ISAT test.

ISAAP motivates schools to meet minimum proficiency standards by imposing sanctions of increasing severity upon schools that fail to meet minimum proficiency targets for two consecutive years. If failing to make AYP for two consecutive years, a school will receive technical assistance from its district and must offer parents the option

ADVANCED: Exceeds Standards

The student demonstrates thorough knowledge and mastery of skills that allows him/her to function independently above their current educational level.

a) The student demonstrates a comprehensive understanding of all relevant information relevant to the topic at level.

b) The student demonstrates comprehension and understanding of knowledge and skills above his/her grade level.

c) The student can perform skills or processes independently without any significant errors.

PROFICIENT: Meets Standards

The student demonstrates mastery of knowledge and skills that allow him/her to function independently on all major concepts and skills related to their educational level.

a) The student demonstrates a comprehensive understanding of all information relevant to the topic at level.

b) The student can perform skills or processes independently without any significant errors.

BASIC: Below Standards

The student demonstrates basic knowledge and skills usage but cannot operate independently on concepts and skills related to his/her educational level. Requires remediation and assistance to complete tasks without significant errors.

a) The student has an incomplete knowledge of the topic and/or misconceptions about some information.

b) The student requires assistance and coaching to complete tasks without errors.

BELOW BASIC: Critically Below Standards

The student demonstrates significant lack of skills and knowledge and is unable to complete basic skills or knowledge sets without significant remediation.

a) The student has critical deficiencies of relevant knowledge of topic and/or misconceptions about some information.

b) The student cannot complete any skill set without significant assistance and coaching.

Figure 1. Description of Idaho proficiency standards.

to transfer their children to another public school, the transportation expense of which is born by the district. If choice is not an available option, a school must offer supplemental services, such as tutoring, after school classes, and summer classes, to eligible students in

reading and math.[16] Technical assistance provisioned by the district may consist of school improvement planning and implementation, data analysis, identification and implementation of scientifically based instructional strategies, professional development, and budget analysis. Furthermore, the school must complete a 2-year school improvement plan for district review within 90 days of its identification as failing.

If the same school fails to meet AYP the following year, the school must implement the intervention improvement plan developed in concert with its district the year prior, in addition to continuing to offer choice and supplemental services to eligible students. If failing to meet AYP for four consecutive years, the school is subject to corrective action. Under corrective action, the responsible district must enroll the school in a technical assistance program and/or implement at least one of the following seven courses of action: (a) provide to school staff appropriate, scientifically based professional development to elevate the achievement of low-performing students; (b) institute a new curriculum grounded in scientifically based research and provide appropriate professional development to support curriculum implementation; (c) extend the length of the school year or school day in an effort to improve instruction and increase student learning; (d) replace school staff deemed responsible for the school failing to make AYP; (e) significantly decrease management authority at the school; (f) restructure the internal organization of the school; or (g) appoint one or more external experts to advise the

[16]Regardless of Title I status, all schools are subject to sanction if they do not make AYP for two consecutive years. The sanctions to which non-Title I schools are subject differ from Title I schools only when a district permits a non-Title I school to offer computerized remediation programs, remedial online classes, after-school academic programs, or any other district-sponsored remedial or tutoring services in place of supplemental services preapproved by the state.

school on its improvement plan and on the specific issues underlying the school's continued failure to make AYP.

In Year 5, a failing school must plan for restructuring, including providing teachers and parents opportunity to comment on and participate in development of a restructuring plan. A school's restructuring incorporates at least one of the following five elements: (a) replacement of all or most of school staff; (b) contracting with an entity with a track record of effectiveness, such as a private management company, to assist in the continued operation of the school as a public concern; (c) turning operation of the school over to the state education agency; (d) reopening the school as a public charter school; or (e) implementation of any other major restructuring of school governance consistent with state-specified principles of restructuring. Figure 2 summarizes the schedule of sanctions under ISAAP for schools failing to make AYP.

Years 1 & 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8
Alert Status	Improve-ment 1	Improve-ment 2	Improve-ment 3	Improve-ment 4	Improve-ment 5	
School on Alert	Technical Assistance School Choice Supplemental Services Create Improvement Plan	Technical Assistance School Choice Supplemental Services Implement Improvement Plan	Technical Assistance School Choice Supplemental Services Corrective Action Planning	Technical Assistance School Choice Supplemental Services Implement Corrective Plan Restructuring Planning	Technical Assistance School Choice Supplemental Services Implement Restructuring Planning	School Starts Over

Figure 2. Sanction schedule for schools failing to make Adequate Yearly Progress.

ISAAP also includes a “Safe Harbor” provision for schools and districts that are making progress in elevating student achievement but have not yet met prescribed proficiency targets. The provision is designed to prevent overidentification of failing schools. If a district or school misses a subgroup proficiency target in reading, math, or both subjects, it still can avoid sanction if: (a) the subgroup reduced the percentage of students below proficient by 10% when compared to the previous year; (b) the school as a whole, and each subgroup with 34 or more students, met the language arts goal; and (c) the school attained a 95% test participation rate (Idaho State Board of Education, 2006).

Under ISAAP, schools may also be rewarded for elevating students’ academic achievement. The Idaho State Board of Education recognizes schools for academic excellence in two distinct manners: as “Distinguished Schools” and/or with Additional Yearly Growth (AYG) Awards. Schools gain eligibility for “distinguished” classification by being among the top 5% of schools exceeding AYP intermediate targets and by significantly reducing achievement gaps between student subgroups. For schools that have met AYP, eligibility for AYG awards entails demonstrating improved proficiency among student subgroups or by elevating school-level proficiency in the aggregate by greater than 10% in a given year.

Of 705 Idaho schools that submitted reports for AYP determination during the 2002-03 school year, 394 schools made all 41 goals, 212 schools missed at least one goal and did not make AYP, and 99 schools were not eligible for designation due to the Safe Harbor provision, minimum reporting requirements, or school type (Idaho State Board of Education, 2006). Of the 212 schools missing at least one goal, 50 schools failed reading but not math, 21 schools failed math but not reading, and 31 schools failed both subjects.

81 of these 212 schools failed at least one of the racial student subgroups in either math or reading, 38 of which failed in both subjects. Of the 81 schools failing at least one of the primary racial student subgroups (i.e., White, Hispanic, Native American, Black, and Asian), almost every school failed exclusively in the White and/or Hispanic subgroups. Schools failing in the Hispanic subgroup tended to fail more often in reading, while schools failing in the White subgroup, conversely, failed substantively more frequently in math. Lastly, 128 schools failed in the Free and Reduced Price Lunch student subgroup in either reading and/or math, 49 schools of which failed in reading alone, but only 15 schools of which failed in just math.

During the 2003-04 school year, and of 731 schools reporting AYP calculations, 513 schools made all 41 goals, 111 schools missed at least one goal, and 117 schools were not eligible for designation. Of the 111 schools missing at least one goal, only 7 schools failed reading but not math, only 3 schools failed math but not reading, and just 9 schools failed both subjects; 69 of these 111 schools, however, failed at least one of the racial student subgroups in either math or reading, 40 of which failed in both subjects. Of the 69 schools failing at least one of the primary racial student subgroups (i.e., White, Hispanic, Native American, Black, and Asian), almost every school failed exclusively in the Hispanic subgroup. Schools failing in the Hispanic subgroup failed more often in Reading. Lastly, 68 schools failed in the Free and Reduced Price Lunch student subgroup in either reading and/or math, 35 schools of which failed in reading alone, but only 9 schools of which failed in just math.

Following high-stakes testing in the spring of the 2004-05 school year, 345 schools of 752 reporting made all 41 goals, 259 schools failed to make AYP, and 149

schools were exempted. Of the 259 schools missing at least one goal, 25 schools failed reading but not math, 18 schools failed math but not reading, and 29 schools failed both subjects; 121 of these schools failed at least one of the racial student subgroups in either math or reading, 69 of which failed in both subjects. Of the 121 schools failing at least one of the primary racial student subgroups (i.e., White, Hispanic, Native American, Black, and Asian), almost every school failed exclusively in the Hispanic subgroup, although the number of schools failing AYP in the White student subgroup did rise slightly from the 2003-04 school year. The difference in the number of schools failing in the Hispanic subgroup in reading versus math grew substantially in the 2004-2005 school year. Lastly, 159 schools failed in the Free and Reduced Price Lunch student subgroup in either reading and/or math, 75 schools of which failed in reading alone, but only 22 schools of which failed in just math.

Characterizing ISAAP-Induced Resource-Allocation Decisions

By building upon theoretical and empirical discussions found in pre-NCLB research,[17] and by borrowing insight from economic and psychological theory,[18] this section briefly characterizes resource-allocation decision making by schools in response to ISAAP. When an accountability system such as ISAAP is implemented, schools are

[17]Specifically, this framework builds upon work by Deere and Strayer (2001), Holmes (2003), Reback (2006), and Chakrabarti (2006), and draws from an international study by Burgess, Propper, Slater, and Wilson (2005) that examines the impact of the United Kingdom's accountability program on the distribution of student gains following implementation in 1988.

[18]Although economic and psychological perspectives on schooling may appear an unlikely match, MacPhail-Wilcox (1988, p. 233) insightfully argues: "They are disciplines with similar goals, kindred concepts, and related propositions. Economists investigate ways that resources are allocated to satisfy human wants, and motivation scholars identify the needs undergirding wants and the processes that activate and sustain behavior. Both seek to explain how these phenomena affect satisfaction, behavior, and productivity."

motivated to focus resources on high-stakes activities and students on the threshold of passing. As the likelihood grows that the total percentage of any one subgroup is below state-defined proficiency standards, so too does the worth of focusing resources on these students to the school. Concomitantly, there is a limit to this notion, particularly in that schools may shift resources away from habitually low-performing students. Nonetheless, if a school believes it can make AYP, and penalty aversion is a significant factor, then schools will change behavior, particularly when resource constrained.

I predict a shift of resources away from students at either tail of the achievement distribution and toward marginal students, under the hypothesis that devoting resources well above or well below the passing threshold will have very small marginal effects on the likelihood of a school making AYP. Under this premise, the frequency and magnitude of the low-/high-achievement tradeoff is a function both of whether a school failed to meet ISAAP defined proficiency standard in the prior year and of how much effort might need be expended by that school to reach proficiency, as defined and measured by how far the marginal student in each of the school's one or more failing student subgroups is from reaching the ISAAP-defined proficiency standard in a particular year.

Schools not subject to AYP determination under NCLB's Safe Harbor provision are likely to respond strategically in a manner similar to that of failing schools not seeking or ineligible for Safe Harbor exemption. Eligibility for Safe Harbor designation necessitates that schools decrease the percentage of students in the nonproficient student subgroup by 10% from the preceding school year on both reading and mathematics indicators. As a consequence, one might argue that a serially underperforming school with little prospect of reaching proficiency in the near term has equal, if not greater,

incentive to improve the performance of marginal students, or risks losing Safe Harbor eligibility, failing AYP, and becoming subject to NCLB sanctions. While Idaho does not report specifically which schools avoided failing status through Safe Harbor exemption, I am able to capture these schools in the greater pool of failing schools by calculating the yearly performance of each school's applicable student subgroups in relation to the ISAAP-defined proficiency standard. Nevertheless, the overall number of schools exempted under Safe Harbor as a percentage of total failing schools appears small. In 2004, for instance, Idaho reported that the number of schools identified for school improvement reduced by approximately 3.5%, or 4 about schools, following the state's granting of Safe Harbor exemptions.

In either scenario, whether that of the school close to passing AYP or that of the serially underperforming school seeking to decrease its percentage of failing students per annum, this characterization of ISAAP-induced resource-allocation decision making assumes that schools: (a) respond strategically to high-stakes accountability programs, (b) are both well-informed and well-intentioned in their resource decision making, and (c) face resource constraints. Expectancy theory, psychology's classical behaviorist model of motivation, offers an extensive empirical research base on why and how individuals and organizations respond to incentives. Expectancy theory states that individuals react to external stimuli such as sanctions, and that individual responses are a product of personality, skills, knowledge, experience, ability, and so forth (Vroom, 1964).[19] Past

[19]More specifically, expectancy theory is composed of three elements: expectancy probability, instrumentality probability, and valence. Expectancy probability (E→P) is the belief that one's effort (E) will result in attainment of desired performance (P) goals. Instrumentality probability (P→R) is the belief that if one does meet performance (P) goals, he or she will receive a greater reward (R). Valence (V(R)) is the value (V) an individual places on the reward (R). Ultimately, Vroom (1964) and others suggests that valence x (expectancy (instrumentality)) = motivation.

research suggests incentives are highly effective in motivating behavior, even when employees are strongly intrinsic (Locke & Latham, 1990; Mohrman & Lawler, 1996), while a growing body of educational accountability research indicates that schools act strategically in response to high-stakes accountability programs (see, e.g., Chakrabarti, 2006; Figlio, 2005; Jacob & Levitt, 2003; Reback, 2006).

The informational immediacy and transparency of ISAAP's reporting system provides support for the argument that schools are capable of making knowledge-based resource-allocation decisions. All ISAT results are reported to teachers, schools, and districts within 72 hours of test administration in the fall and spring (Northwest Evaluation Association, 2006). Moreover, NWEA's testing system automatically analyzes data prior to release, providing schools, administrators, and teachers with percentile rank, achievement score, projected proficiency on state tests, and year-over-year test score growth for each student following administration of the fall and spring tests (Northwest Evaluation Association, 2006). These data and information enable a school to gauge how each student and student subgroup performs in relation to ISAAP's minimum competency standards and to make informed resource-allocation decisions in response to ISAAP proficiency standards within the first few weeks of the school year.

This research may not detect a tradeoff, though, if a school becomes more efficient by substituting resources across outcomes. Education is a complex and multidimensional enterprise (Baker, 1992; Holmstrom & Milgrom, 1991), and ISAAP does not encompass all relevant activities of the state's education system.[20] For instance, ISAAP holds schools accountable in math, reading, and language arts, but not science,

[20]Prendergast (1999, p. 21) noted that, "The use of explicit contracts could cause agents to focus too much on those aspects of the job included in the contract to the detriment of those that are excluded."

social studies, and so forth. As such, schools may focus additional resources on high-stakes tests and subjects (i.e., math, reading, and language arts) to the detriment of low-stakes activities (Center for Education Policy, 2005; Chakrabarti, 2006; Deere & Strayer, 2001).[21]

Schools also may become more efficient by tracking students by ability. As first noted in Deere and Strayer (2001), a principle assumption in the study of achievement tradeoffs is that curriculum will focus on the least rigorous portion of content tested and required to meet proficiency thresholds because the marginal gain in the passing rate from teaching more advanced material will be small. If schools are not restricted to offering the same instruction to all students in the same classroom, then tradeoffs may not be evident when measured by the distribution of student achievement. Of course, this solution assumes schools can easily track students or differentiate the curriculum, and do so without incurring additional cost.

It is also possible that districts and/or private foundations provide resource support to schools designated as needing improvement. Schools could use these additional resources to hire more teachers or extend the school day. If schools receiving additional district-level or private foundational resources use these resources to avoid ISAAP sanction, distributional effects may be dampened.

[21]Future iterations of this research will be able to examine if Idaho public schools reduce instruction in low-stakes subjects when data for the 2005-06 school year become available. In 2005, Idaho entered into a flexibility agreement with the US Department of Education that permits districts to alter accountability program features required under NCLB. Specifically, the language arts portion of the ISAT was a high-stakes exam in Idaho for 3 years (i.e., 2002-03, 2003-04, 2004-05 school years). Starting in the 2005-06 school year, however, the language arts test was replaced by state-defined student growth measures. To test for a subject tradeoff one can compare achievement before and after Idaho entered into flexibility agreement.

CHAPTER IV

DATA SOURCES, DATA DEVELOPMENT, AND ESTIMATION STRATEGIES

NWEA's Growth Research Database

The primary data for this study are drawn from NWEA's Growth Research Database (GRD), a data system that collects longitudinal student-level achievement results from approximately 2,200 school districts in 45 states. The GRD assessment system, as described by Cronin et al. (2005), is a rich database for educational policy research considering:

> All scores for the NWEA assessment in a subject area reference a single, cross-grade, equal-interval scale developed using Item Response Theory methodology. . . . The RIT scale is designed to measure student growth and performance across time as well as to take advantage of strong measurement theory and experimental design, and have been demonstrated to be extremely stable over twenty years of development and use (Kingsbury, 2003). This stability holds for each subject area measurement scale (reading, mathematics and language usage) and across grades levels from 3 to 8 within subjects (Northwest Evaluation Association, 2002). (p. 16)

Starting with the 2002-03 school year, NWEA's GRD contains test score information for over 90% of Idaho students in mathematics, reading, and language arts. GRD assigns each student a unique identifier as long as the student is enrolled in Idaho's public school system. This tracking mechanism offers researchers access to multiple observations on each individual student in the sample and opportunity to construct a panel dataset. GRD also contains demographic and other relevant information on students and schools, including race/ethnicity, gender, free and reduced price lunch status, grade,

date of birth, school type, and school size. School finance data, however, are not available.

Data Development

Development of the data included selecting eligible schools and students. GRD contains demographic information and ISAT scores for students in traditional public schools, nontraditional public schools (e.g., students enrolled in charter or virtual/on-line schools), and private schools (e.g., Catholic, other religious, and nonreligious). Eligible cases were restricted to students enrolled in traditional public schools or public charter schools because private schools are not held accountable under ISAAP.

There are 51 schools in Idaho for which the total tested student population is less than ISAAP's minimum prescribed *n* of 34. Under ISAAP, AYP for these schools is calculated using 3 years of achievement data to obtain a more consistent and reliable determination due to their small sample size. These unusually small schools were eliminated, resulting in exclusion of a total of 727, 730, and 708 student observations for the 2002-03, 2003-04 and 2004-05 school years, respectively.

Reconfigured schools were also deleted from the sample. ISAAP does not hold reconfigured schools accountable until 3 full years of achievement data is collected. Across all 3 years, there were a total of 10 reconfigurations, or eight school consolidations and two school deconsolidations. Eliminating these schools reduced the total number of student observations by 1,921, or approximately 1% percent of all traditional public school and charter school students in grades three through eight.

Eligible student observations were restricted to students enrolled in grades three through eight because grades above and below this band are not subject to sanction under ISAAP. The dependent variable relies on a fall-to-spring student gain score; therefore, students without both fall and spring RIT scores in a given school year were excluded from analysis. This restriction was also placed on data in light of the fact that Section 112.03 of Idaho's Administrative Procedures Act requires that only students who are continuously enrolled in the same public school from the end of the first 8 weeks, or 56 calendar days, of the school year through the spring test administration period be included in AYP calculations. While coding procedures employed are not based on student attendance patterns as defined in Section 112.03, this restriction is believed to produce a closer approximation to actual AYP calculations. In total, these restrictions eliminated less than 17,500 student observations, or an average of 5.5% of student observations per year. Moreover, results are robust to estimation of tradeoffs when including non-full-year students in school-level proficiency calculations.

Because estimates might be influenced by atypical values, outliers were flagged according to the following algorithm. For each school, the first (Q1) and third (Q3) quartile values were identified for a given variable, and any value that was 1.5 IQR below Q1 or 1.5 above Q3 was flagged as suspicious value. Values that were 3 IQR below Q1 or 3 IQR above Q3 were flagged as extreme outliers. Results are robust to estimation of tradeoffs when removing no outliers, when removing only extreme values, or when removing both suspicious and extreme values.

To investigate data reliability following cleaning and development procedures, resultant student achievement descriptive statistics were compared to reported values

published in a biannual series of statewide results brochures maintained by the Idaho State Board of Education (ISBE). These comparisons are reported in Table 3.

Table 3

Comparison of RIT Scores Calculated Post-Data-Development to RIT Scores Reported by Idaho State Department of Education

	Fall 2002		Spring 2003		Fall 2003		Spring 2004		Fall 2004		Spring 2005	
Grade	Calcu-lated	Report-ed	Calcu-lated	Report-ed	Calcu-lated	Report-ed	Calcu-lated	Report-ed	Calcu-lated	Report-ed	Calcu-lated	Report-ed
3	189.9	189.8	202.1	202.1	192.8	192.7	204.9	205.0	194.1	194.1	204.7	204.8
4	200.4	200.3	211.6	211.7	202.6	202.5	214.6	214.6	204.1	204.0	216.2	216.3
5	208.8	208.5	216.6	216.6	209.2	209.1	219.8	219.8	212.6	212.4	221.3	221.4
6	215.1	214.9	223.7	223.7	217.1	216.9	226.3	226.3	219.2	219.1	225.8	225.9
7	222.6	222.3	229.0	228.9	224.1	223.7	230.7	230.7	225.9	225.8	232.3	232.4
8	229.2	228.7	234.0	233.9	230.1	229.7	237.2	237.2	231.6	231.3	237.8	237.9

Between 83% and 86% of third- through eighth-grade students are White. Between 11% and 15% are Hispanic or Latino, and the remaining 3% are Black/African American, Asian/Pacific Islander, or American Indian/Native Alaskan. Between 34% and 40% of students were identified as economically disadvantaged during the spring semester as defined by free and reduced price lunch status. The fall-to-spring gain score in math has a mean of 8.91 points with a standard deviation of 6.92 points when pooled across all grades and years. These sample means are weighted by the number of students at each campus who take the ISAT.

Measuring Low-/High-Achievement Tradeoffs

Three previous research studies examined systematic shifts in intraschool resource-distribution patterns in response to high-stakes accountability programs using a student gain score as the dependent variable.[22] Deere and Strayer (2001) and Holmes (2003) constructed a student gain measure using student test score changes from time t+1 to t+2 on the left-hand side of the equation where both time periods represent spring test administrations. Isolating student gain on the left-hand side of the equation is an attractive approach in that the difference between the two measurement errors is picked up by the disturbance term.

Reback (2006) advanced Deere and Strayer (2001) and Holmes's (2003) specification by developing a student gain score that compared a student's performance to typical gains at that particular point in the achievement distribution. He converted each student's test score to a Z-score based on the performance of a student in the same grade with an identical score from the previous year. Reback (2006) argued that this specification was superior to that of Deere and Strayer (2001) and Holmes (2003) because results were robust to the influence of mean reversion. In comparing a student to peers with the same initial score, Reback controlled for the amount of mean reversion expected from this group of students. Once a student's test-score gain was measured in relation to peers in the same group, the amount of mean reversion that group might have exhibited in comparison to other groups was no longer an issue.

[22]Other studies examining subgroup-level data have estimated shifts in the percentage of students scoring in different performance levels within and between failing schools and nonfailing schools (Chakrabarti, 2006), or the second difference in average pass rates pre- and post-NCLB (Ballou, Liu, & Rolle, 2005). An international study by Burgess et al. (2005) examined the difference between a student's test score at age 16 and at age 14 where age 14 test score was used as a regressor.

While these studies all identified viable options for constructing a dependent variable capable of capturing a low-/high-achievement tradeoff, this dissertation employs an alternative approach. ISAT is administered twice per year, allowing for construction of a fall-to-spring gain score for each individual student in the 2002-03, 2003-04, and 2004-05 school years. A fall-to-spring gain score is advantageous because spring-to-spring gain scores have the confounding influence of the summer months, meaning that any gain (or potential loss) in a student's ISAT score due to what the school provided cannot be disentangled easily from how much gain (or loss) occurred as a result of summer activities.[23]

This dissertation also generates a standardized fall-to-spring test-score gain for each student based on a comparison of a student's nominal gain and the average gain in achievement for all students in Idaho by applicable year and grade. Similar to the approach employed in Texas charter school research by Hanushek, Kain, Rivkin, and Branch (2005) and Booker, Gilpatric, Gronberg, and Jansen (2006), the initial achievement distribution in Idaho is defined by fall ISAT score and divided into 20 equal intervals for each unique combination of year and grade. The mean and standard deviation test-score gain for all students starting in a particular interval for each unique combination of year and grade are then computed. The standard fall-to-spring test-score gain for each student is calculated as the difference between that student's nominal gain and the mean gain of all students in the interval over the standard deviation of all student gains in the interval. Gains in each interval are therefore distributed with a mean of zero and standard deviation of one. Ultimately, this standardization of gains allows me to test

[23]Alexander, Entwisle, and Olson (2001) used hierarchical growth models to demonstrate the seasonality of children's learning over the school year and summer months.

whether the model specifications using the unstandardized fall-to-spring test-score gains are robust to bias resulting from mean reversion.

Measuring a School's Short-Run Incentive to Target Resources

The incentive structure embedded in NCLB-mandated state accountability programs has led some to claim that students that could most influence a school's AYP rating will receive a greater proportion of educational resources, especially if the total percentage of any one subgroup is below state-defined proficiency standards. Without taking into consideration a school's short-run incentive to target instruction, research studies seeking to quantify the magnitude and frequency of the low-/high-achievement tradeoff will underestimate strategic resource-allocation decisions made in response to high-stakes accountability.[24] Unfortunately, there are no definitive constructs that indicate a school's short-run incentive to target particular students.

Holmes (2003) was the first researcher to construct an indicator for measuring a teacher's incentive to target instruction in response to high-stakes accountability systems. Holmes' strategy consisted of two components. First, he calculated how each student's predicted test score fell from the state accountability system's defined performance target. Second, he calculated the importance to a school of getting a particular student over the performance target. Then to identify a student's relative importance, Holmes

[24]Since ISAAP was implemented statewide in 2002 there are no true counterfactuals against which to compare schools' strategic resource-allocation decision making. Investigating the impact of ISAAP's incentive structure on resource-allocation practices requires examining natural variation found within Idaho's K-12 school system. However, as noted by Reback (2006), because all schools are exposed to the accountability program, results may actually underestimate distributional effects as schools seek to make permanent changes to raise pass rates, regardless of short-run incentives. This effect may be intensified by NCLB's goal toward 100% proficiency by the 2012-13 school year.

determined the underlying distribution of student quality in a school and then situated each student on this continuum based upon their predicted spring achievement in relations to North Carolina's accountability system and defined performance thresholds.

Holmes, however, imposed a fixed bandwidth value that restricted the estimated degree of resource targeting at the school level. Despite estimating that some schools must target students 20 units below the performance threshold, Holmes restricted the maximum distance a "targeted" student could fall from the performance threshold to 5 units below the performance standard. This restriction biases resultant estimates of whether schools target resources to marginal students.

Burgess et al. (2005) estimated the impact of the proportion of marginal students in a school on the achievement gains of traditionally low-performing students in the United Kingdom. For each student observation, they interacted the proportion of marginal students in a school with an indicator variable that took the value of 1 if the student's ability was considered low. The coefficient on the conditioning effect reflected the impact of the United Kingdom's accountability policy. Burgess et al.'s strategy was limited, however, in that their low-performing student dummy was based on a non-high-stakes test taken by students at age 14, while their dependent variable of interest were results from a high-stakes test taken at age 16.

Reback (2006) developed a more complex technique for estimating a school's incentive to improve the expected performance of certain students. He estimated the marginal effect of a hypothetical improvement in the expected performance of a particular student on the probability that a school would obtain a certain accountability rating. Reback's indicator permitted him to test whether students earned higher than

expected test score gains when schools were subject to stronger short-run incentives to focus effort and resources on marginally performing students.

This dissertation constructs a series of indicator variables to measure whether a particular student is likely to be targeted. The first indicator variable constructed, $F.ISAT.M_{ij(t-1)}$, is a dummy variable denoting whether a student failed the fall administration of the ISAT math test. Failure is defined as any score below the state-defined proficiency standard for that student's particular grade and year combination. Approximately 50% of all fall test takers did not meet grade specific performance threshold.

$F.ISAT.M.ADJ_{ij(t-1)}$, an adjusted version of the first indicator variable, is a dummy variable denoting whether a student that failed fall administration of the ISAT math test was expected to pass the test's spring administration by making normal gains over the school year. Using the average fall-to-spring gain for each grade, this adjustment helps reclassify as passing those failing students in the fall who were unlikely to be targeted with resources given expectations of meeting the state-defined proficiency standard in the spring. Approximately 73% of all test takers were expected to pass the spring ISAT after taking into consideration average expected gains.

The second indicator variable constructed, $ST.GAP.M_{ij(t-1)}$, uses the distance of a student's score from the state-defined passing threshold to indicate the probability that a school overlooks that student. $ST.GAP.M_{ij(t-1)}$ is defined as the gap between the passing threshold and a student's test score. For example, if the state-defined proficiency standard is 50 and a particular student in the fourth grade receives a 55 on the test, the student is

assigned a value of 5 for $ST.GAP.M_{ij(t-1)}$. The average $ST.GAP.M_{ij(t-1)}$ value across all years and grades was 3.36.

$ST.GAP.M.ADJ_{ij(t-1)}$, an adjusted version of the second indicator variable, takes into consideration a student's expected fall-to-spring test-score gain in a particular grade. $ST.GAP.M.ADJ_{ij(t-1)}$ denotes the gap between the passing threshold and a student's test score after adjusting for whether that student is expected to pass the spring administration of the ISAT math test by making normal gains over the year. For example, assuming then that fourth graders were expected to gain 3 points on average between the fall and spring, a student with an $ST.GAP.M_{ij(t-1)}$ value of 5 would be assigned a value of 2 for $ST.GAP.M.ADJ_{ij(t-1)}$. The average $ST.GAP.M.ADJ_{ij(t-1)}$ value is .40.

$ST.GAP.M_{ij(t-1)}$ and $ST.GAP.M.ADJ_{ij(t-1)}$, however, impose a strong assumption; that being, increased gains of a student who starts 20 points below the cut-off equals the amount by which the gain will diminish when a student starts 20 points above the cut-off. By imposing such a strong assumption on the relationship of $ST.GAP.M$ and $ST.GAP.ADJ.M$ to student gains, the estimates are likely to be biased. Therefore, $ST.GAP.PASS.M_{ij(t-1)}$ and $ST.GAP.FAIL.M_{ij(t-1)}$ are created to relax this assumption. $ST.GAP.PASS.M_{ij(t-1)}$ is defined as the gap between the passing threshold and a student's test score that passes the fall ISAT. $ST.GAP.FAIL.M_{ij(t-1)}$ is defined as the gap between the passing threshold and a student's test score that failed the fall ISAT. These variables are also modeled after adjusting for average gain by grade and year.

This dissertation relies on a single indicator, $F.PROF.M_{jt}$, to capture a school's incentive to target instruction. This indicator, and the third constructed, is a dummy

variable denoting whether a school met ISAAP's minimum proficiency standard. $F.PROF.M_{jt}$ replicates ISAAP's proficiency standard calculations under ISAAP by determining whether each traditional public school, and all defined subgroups therein, met the state-prescribed proficiency threshold in each of the 3 years represented in the panel following administration of the ISAT math test in the spring. This approach helps account for the potential presence of achievement tradeoffs both in Safe Harbor schools that must act strategically in the short run to increase the percentage of students scoring at or above proficiency by 10% year-over-year and in schools where the size of certain failing subgroups fluctuates around minimum *n* size of 34, thereby making that subgroup's continued exemption from AYP designation an uncertainty and preserving the school's incentive to elevate the performance of students at or near the proficiency cut-off. Four different versions of the third indicator were constructed using minimum *n* thresholds of 34, 32, 28, and zero students. Each construction yielded similar results in cross-wise comparisons.

A final variable is used to denote whether a student in the subgroup for which a school failed to meet proficiency. The variables are identified as $F.SUB.HISP.M$ and $F.SUB.WHITE.M$. Variables are limited to White and Hispanic subgroups because approximately 96% of all Idaho students are either White or Hispanic and AYP determination for non-White or non-Hispanic subgroups effects less than 3% of traditional Idaho public schools.

Other Factors Influencing the Low-/High-Achievement Tradeoff

Mediating factors may explain further why one student category is rewarded at the expense of another in the resource-allocation decision-making process. In particular, this dissertation examines if the outcome of a low-/high-achievement tradeoff varies depending upon: (a) the severity of sanction faced by a school, (b) whether a student is in a terminal grade, (c) whether a student is in more than one failing subgroup, and (d) differences in market sensitivity among traditional public and public charter schools.

Severity of Sanction

NCLB monitors schools' progress toward meeting academic "proficiency" through AYP calculations and motivates them to meet minimum proficiency standards by imposing sanctions of increasing severity for failure to meet AYP. Schools may avoid being labeled as failing AYP, however, by (a) reducing the percentage of students below proficient by 10% when compared to the previous year, (b) meeting the language arts proficiency goal, and (c) maintaining a 95% test participation rate (Idaho State Board of Education, 2006). Additionally, subgroups with less than 34 students are not held accountable.

Given that the severity of these sanctions varies across schools, this dissertation creates two dummy variables to denote the school improvement status of a "failing" school, defined herein as a school that failed to meet the prescribed proficiency target in at least one of the 10 subgroups designated under ISAAP following administration of the ISAT math test in the fall and/or spring of the prior year. $SCH.IMP.Y_{jt}$ equals 1 if a school failed to meet state-defined proficiency standard in at least 1 year, and 0 if a

school met proficiency in all years; $SCH.IMP.YY_{jt}$ equals 1 if a school failed to meet proficiency in two consecutive schools years and 0 if a school met proficiency in all years. It is hypothesized that a school's investment in marginal students will increase with each additional and successive failure to meet proficiency.

Terminal Versus Nonterminal Grade

As noted in Chapter II, and discussed in previous research (Deere & Strayer, 2001; Reback, 2006), the estimation strategy assumes that a school's response to ISAAP is a "one year optimization problem." That is, in order to avoid sanction, schools are only concerned with allocation of resources to elevate the performance of marginal students in a single year. In reality, schools likely invest in certain students over a multi-year period.

To examine if schools indeed respond strategically to the period for which a school is held accountable for a student's academic performance, this dissertation compares student achievement gains in terminal and nonterminal grades. A binary indicator variable, $TRMGRD_{ij(t+1)}$, takes the value of 1 if student *i* in school *j* is in a terminal grade in the spring, or 0 if the student is not. If schools indeed employ a "one year optimization" strategy in response to ISAAP, then one would expect to detect no difference in the performance of marginal students in terminal grades relative to that of marginal students in nonterminal grades. Conversely, if schools are more likely to invest more heavily in students for which they are accountable for extended periods of time, then one might expect to detect instances of marginal students in terminal grades underperforming in relation to marginal peers in nonterminal grades. Given that Idaho holds all subgroups accountable for the failing performance of any subgroup in the prior

year, the latter scenario seems less likely, and one would expect any evidence of this effect to be random and isolated.[25]

Students in Overlapping Categories

ISAAP holds schools accountable for student performance in 10 subgroup categories. However, subgroup classifications are not mutually exclusive. Schools with economically disadvantaged, limited English proficient, and/or disabled students are held accountable for the performance of these students across at least three dimensions–the "all students" subgroup, a "race/ethnicity" subgroup, and at least one of the three aforementioned subgroup classifications. If a school faces the threat of sanctions due to the performance of White and economically disadvantaged students in math, for instance, it is hypothesized then that low-income White students will be targeted more than affluent White students or non-White low-income students.

Market Sensitivity of Public Charter Schools

This dissertation also compares strategic resource decision-making practices within and between traditional public schools and public charter schools. Charter schools are held accountable under NCLB, and hence subject to the same sanctions as traditional public schools. Thus, there are incentives for charter schools to engage in the same strategic tradeoffs as regular public schools. However, charter schools, by virtue of

[25]Reback (2006) conducted a similar analysis, concluding: "The effect of the incentive for a school to improve the math performance of particular students is more than twice as large among students in the terminal grades of their schools. In addition, relatively high achieving students in terminal grades are much more significantly hurt by having many classmates who could influence the school's rating that year. Similarly, the lowest achieving students in terminal grades are helped far less than usual when they have many classmates who could influence the school's rating that year" (p. 22).

funding and enrollment policy, are more market sensitive than their traditional public school counterparts, a condition that may attenuate allocation of resources to particular groups of students. A binary indicator variable, $CHARTER_{ijt}$, takes the value of 1 if a school is a public charter school or a value of 0 if a school is a traditional public school.

Estimation Strategy

The basic general linear model is specified as follows. Let $\Delta Y_{ijt} = Y_{ijt} - Y_{ij(t-1)}$ be the math test-score gain for student *i* in school *j* from fall to spring administration of ISAT, where *t* denotes spring administration of the test. Then

$$\Delta Y_{ijt} = Y_{ijt} - Y_{ij(t-1)} = \alpha_0 + \alpha_1 F.PROF.M_{j(t-2)} + \alpha_2 HISPANIC_{ijt} + \alpha_3 WHITE_{ijt} + \alpha_4 F.PROF.M_{j(t-2)} \times F.SUB.HISP.M_{ijt} + \alpha_5 F.PROF.M_{j(t-2)} \times F.SUB.WHITE.M_{ijt} + \alpha_6 F.PROF.M_{j(t-2)} \times F.ISAT.M_{ij(t-1)} + \alpha_7 F.PROF.M_{j(t-2)} \times F.SUB.HISP.M_{ijt} \times F.ISAT.M_{ij(t-1)} + \alpha_8 F.PROF.M_{j(t-2)} \times F.SUB.WHITE.M_{ijt} \times F.ISAT.M_{ij(t-1)} + \alpha_9 F.ISAT.M_{ij(t-1)} + \alpha_{10} F.ISAT.M_{ij(t-1)} \times HISPANIC_{ijt} + \alpha_{11} F.ISAT.M_{ij(t-1)} \times WHITE_{ijt} + \chi_{jt} + \mu_j + \eta_g \times \gamma_t + e_{ijt}.$$

The value on α_1 represents the difference between the average fall-to-spring test-score gain in math for passing non-Hispanic and passing non-White students enrolled in schools that failed to meet ISAAP's proficiency standard in math in at least one student subgroup the previous spring and the average fall-to-spring test-score gain in math for passing non-Hispanic and passing non-White students enrolled in schools that did meet the standard. The sign on α_1 indicates the direction in which the average test-score gain of passing non-Hispanic and passing non-White students enrolled in failing schools differed from that of racially similar passing students enrolled in passing schools. A negative sign on α_1 coefficient that is statistically different from zero denotes a tradeoff

whereby passing non-Hispanic and passing non-White students enrolled in failing schools gained less than passing non-Hispanic and passing non-White students enrolled in passing schools.

The value on α_2 represents the difference between the average fall-to-spring test-score gain in math for passing Hispanic students enrolled in schools that met ISAAP's proficiency standard in math the previous spring and the average fall-to-spring test-score gain in math for passing non-Hispanic and non-White students also enrolled in schools that met the standard. The sign on α_2 indicates the direction in which the gains of passing Hispanic students enrolled in passing schools differed from those of passing non-Hispanic and non-White students also enrolled in passing schools. A negative sign on the α_2 coefficient that is statistically different from zero denotes that passing non-Hispanic and non-White students in passing schools gained more on average than similarly situated passing Hispanic students.

The value on α_3 represents the difference between the average fall-to-spring test-score gain in math for passing White students enrolled in schools that met ISAAP's proficiency standard in math the previous spring and the average fall-to-spring test-score gain in math for passing non-Hispanic and non-White students also enrolled in schools that met the standard. The sign on α_3 indicates the direction in which the gains of passing White students enrolled in passing schools differed from those of passing non-Hispanic and non-White students also enrolled in passing schools. A negative sign on the α_3 coefficient that is statistically different from zero denotes that passing non-Hispanic and non-White students gained more on average than similarly situated passing White students.

The value on α_4, identified by $F.PROF.M_{j(t-2)} \times F.SUB.HISP.M_{ijt}$, represents whether the average test-score gain in math of passing Hispanic students enrolled in schools that failed to meet ISAAP's proficiency standard in math due to poor performance by Hispanic students the previous spring differed from that of passing non-Hispanic and non-White students also enrolled in failing schools, save for those schools with a failing Hispanic and/or White subgroup. A positive and statistically significant value on α_4 indicates greater gains on average by passing Hispanic students in schools with a failing Hispanic subgroup than by passing non-Hispanic and non-White students enrolled in schools with non-Hispanic and non-White failing subgroups.

The value on α_5, identified by $F.PROF.M_{j(t-2)} \times F.SUB.WHITE.M_{ijt}$, represents whether the average test-score gain in math of passing White students enrolled in schools that failed to meet ISAAP's proficiency standard in math due to poor performance by White students the previous spring differed from that of passing non-Hispanic and non-White students also enrolled in failing schools, save for those schools with a failing Hispanic and/or White subgroup. A positive and statistically significant value on α_5 indicates greater gains on average by passing White students in schools with a failing White subgroup than by passing non-Hispanic and non-White students enrolled in schools with non-Hispanic and non-White failing subgroups.

The value on α_6, identified by $F.PROF.M_{j(t-2)} \times F.ISAT.M_{ij(t-1)}$, represents whether the average test-score gain in math of failing non-Hispanic and non-White students enrolled in failing schools, save for failure by Hispanic and/or White subgroup, differed from that of passing non-Hispanic and non-White students also enrolled in failing schools, save for failure by Hispanic and/or White subgroup. A positive and

statistically significant value on α_6 indicates greater gains on average in math by failing non-Hispanic and non-White students enrolled in failing schools, save for failure by Hispanic and/or White subgroup, than by similarly-situated and passing non-Hispanic and non-White students.

However, the values on these latter three interactions do not explicitly explain the issue of interest, namely if schools responded to ISAAP's incentive structure by targeting resources to failing students in failing subgroups. To provide a more refined specification of resource targeting, this model includes a Student X Subgroup conditioning effect. This three-way interaction permits examination of whether Idaho schools responded to ISAAP by targeting resources to failing students in failing subgroups at the expense of similarly situated peers who met proficiency.

The value on α_7, identified by $F.PROF.M_{j(t-2)} \times F.SUB.HISP.M_{ijt} \times F.ISAT.M_{ij(t-1)}$, is the Failing Student X Hispanic Subgroup interaction. α_7 represents whether the average test-score gain in math of failing Hispanics students enrolled in schools that failed to meet ISAAP proficiency standards due to failing math performance by Hispanic subgroup differed from the average test-score gain in math of (a) passing Hispanic students also enrolled in schools that failed due to failing math performance by the Hispanic subgroup and/or (b) failing non-White and non-Hispanic in failing schools, save for failure by Hispanic and/or White subgroups. With respect to the former, a positive sign and statistically significant value on α_7 indicates greater gains on average in math by failing Hispanic students than by passing Hispanic students in schools that failed due to poor math performance by the Hispanic subgroup, suggesting that these schools

targeted resources to failing students in a failing subgroup at the expense of their passing peers.

The value on α_8, identified by $F.PROF.M_{j(t-2)} \times F.SUBWHITE.M_{ijt} \times F.ISAT.M_{ij(t-1)}$, is the Failing Student X White Subgroup interaction. α_8 represents whether the average test-score gain in math of failing White students enrolled in schools that failed to meet ISAAP proficiency standards due to failing math performance by White subgroup differed from the average test-score gain in math of (a) passing White students also enrolled in schools that failed due to failing math performance by the White subgroup and/or (b) failing non-White and non-Hispanic in failing schools, save for failure by Hispanic and/or White subgroups. With respect to the former, a positive sign and statistically significant value on α_8 indicates greater gains on average in math by failing White students than by passing White students in schools that failed due to poor math performance by the White subgroup, suggesting that these schools targeted resources to failing students in a failing subgroup at the expense of their passing peers.

Nonfailing schools may also have responded strategically to ISAAP. It is plausible nonfailing schools desire to avoid negative publicity associated with being labeled failing in the future and therefore allocate additional resources to failing students. The value on α_9, identified by $F.ISAT.M_{ij(t-1)}$, represents whether the average test-score gain in math of failing non-White and non-Hispanic students in passing schools differed from that of similarly situated passing non-White and non-Hispanic peers. A positive and statistically significant value on α_9 suggests nonfailing schools may have responded to the potential stigma associated with not meeting proficiency and/or the threat of sanctions imposed by ISAAP by elevating the gains of failing students in relation to those of

passing students. It is also plausible, however, that α_9 could represent evidence of regression to the mean. To account for this possible bias, models were rerun using a standardized fall-to-spring student gain score in math as the dependent variables. Results were robust across specifications using a standardized fall-to-spring gain score in math.

The value on α_{10}, identified by $F.ISAT.M_{ij(t-1)} \times HISPANIC_{ijt}$, represents whether the average test-score gain in math of failing Hispanic students enrolled in schools that met ISAAP's proficiency standard in math the previous year differed from the average test-score gain in math of failing non-White and non-Hispanic students also enrolled in passing schools. A negative sign on the α_{10} coefficient that is statistically different from zero suggests that fall-to-spring test score gains of failing Hispanic students are smaller, on average, than those of similarly situated, failing non-White and non-Hispanic students.

The value on α_{11}, identified by $F.ISAT.M_{ij(t-1)} \times WHITE_{ijt}$, represents whether the average test-score gain in math of failing White students enrolled in schools that met ISAAP's proficiency standard in math the previous year differed from the average test-score gain in math of failing non-White and non-Hispanic students also enrolled in passing schools. A negative sign on the α_{11} coefficient that is statistically different from zero suggests that fall-to-spring test score gains of failing White students in passing schools are smaller, on average, than those of non-White and non-Hispanic students also enrolled in passing schools.

A Student X Subject general linear model with fixed effects is the preferred specification for estimating schools' responses to ISAAP for several reasons. First, this model permits estimating relationships between a single dependent variable and more than one response variable simultaneously. In a single stage this model can isolate

whether a school, based on prior year's performance, has targeted resources to (a) students in a failing subgroup, (b) students in a failing subject, and/or (c) students failing math on a failing subgroup. This model can also detect strategic responses by nonfailing schools to elevate achievement of failing students.

Second, a school fixed effects estimator (μ_j) was selected to control for observed and unobserved time invariant characteristics of the school that could be correlated with student achievement gains. Whether a school failed to meet proficiency may be correlated with student achievement gains. Suppose that student achievement gains in failing schools were, on average, smaller than nonfailing schools. If this is true, omitting school effects would yield biased estimates of the parameters of interest. Additionally, a school effects model is a within group estimator that attributes only within school movement in coefficients on parameters of interest. Thus, estimates measure how achievement gains change within failing schools as the school environment changes.

Third, the model takes into consideration other confounding influences on the proposed general linear model specifications such as testing effects and peer effects. Previous research indicates that test difficulty may change from one year to the next (Heubert & Hauser, 1999; Klein, Hamilton, McCaffrey, & Stecher, 2003; Koretz & Barron, 1998). If test difficulty varies from year to year, and/or varies for different student population from year to year, estimates of tradeoffs will be biased. Because each grade and year represents a new test, a year by grade interaction ($\eta_g \times \gamma_t$) can be used to control for testing effects.

Recent research suggests that peer achievement exerts a positive effect on achievement growth, and that students throughout a school's test-score distribution

benefit from the presence of high-performing peers (Hanushek, Kain, Markman, & Rivkin, 2003).[26] Because intraschool peer composition is likely to change from year to year, thereby rendering a cohort fixed effect an inadequate strategy to control for peer composition, I utilized a vector of variables to capture and control for the effect of peer composition (X_{jt}). These variables include poverty status, as measured by eligibility for free or reduced price lunch, and school-level student background characteristics including percentage of students by race/ethnicity.

[26]Positive peer group effects on student outcomes have also been found at the college level (Zimmerman, 2003).

CHAPTER V

RESULTS AND DISCUSSION

The primary objective of this dissertation is to investigate strategic resource-allocation decision making by traditional public schools and public charter schools seeking to close the achievement gap among low- and high-performing students. This chapter reports findings from a series of general linear models using student-level data contained in the Northwest Evaluation Association's Growth Research Database to ask the following research questions:

1. Have Idaho schools responded to ISAAP by raising the achievement of failing students relative to the achievement of nonfailing students?

2. Have Idaho schools responded to ISAAP by raising the achievement of failing student subgroups relative to the achievement of nonfailing student subgroups?

3. Have Idaho schools responded to ISAAP by raising the achievement of failing students at the expense of high-performing students?

4. Do other programmatic features of NCLB explain strategic resource-allocation practices, including severity of sanction, students in overlapping categories, and students in terminal grades?

5. Does increased market sensitivity found in public charter schools impact strategic resource-allocation practices?

Have Idaho Schools Raised the Achievement of Failing Students Relative to the Achievement of Nonfailing Students?

This section reports results from models used to estimate if schools have responded to ISAAP by raising the achievement of low-performing students relative to that of high-performing students. The dependent variable across all reported models is a fall-to-spring measure of student test-score gain in mathematics as measured by the Idaho State Assessment Test. Two indicator variables denoting whether a student is likely to be targeted are used as regressors: (a) a binary variable indicating whether a student failed to meet the state-defined performance threshold on ISAT during the fall test administration, and (b) a continuous variable that measures the distance between a student's fall ISAT score and the state-defined performance threshold. Control variables include percentage of students in a school by race and ethnicity and percentage of economically disadvantaged students as determined by free and reduced price lunch status. Select model specifications include Grade X Year and school fixed effects.

Do Schools Target Failing Students?

Table 4 displays estimates of the effect of a student failing to meet ISAAP's grade-specific performance threshold on individual student growth. Model 1.1 demonstrates a strong response by schools to elevate the performance of students scoring below ISAAP's performance threshold. The estimate on the binary indicator variable, $F.ISAT.M$, is positive and statistically significant at $\alpha < .01$. In substantive terms, Model 1.1 reveals that students that failed to meet the state-defined performance standard on the fall ISAT test gained, on average, 3.03 points (SE = 0.0266) more than

Table 4

Effect of Failing Fall Math ISAT on Math Performance (General Linear Regression Model with Math Gain Score Fall-to-Spring as the Dependent Variable)

	Student Failed the Fall ISAT Math Test *(F.ISAT.M)*				Student Failed the Fall ISAT Math Test, Adjusted for Expected Gain *(F.ISAT.ADJ.M)*			
	None	School	School Grade X Year	School Grade X Year[a]	None	School	School Grade X Year	School Grade X Year[a]
Variable	Model 1.1	Model 1.2	Model 1.3	Model 1.4	Model 1.5	Model 1.6	Model 1.7	Model 1.8
F.ISAT.M	3.033*** (0.0266)	2.989*** (0.0258)	2.965*** (0.0255)	2.928*** (0.0249)	--	--	--	--
F.ISAT.ADJ.M	--	--	--	--	2.883*** (0.0293)	3.512*** (0.0284)	3.766*** (0.0280)	3.682*** (0.0273)

Note. School-level controls for race/ethnicity and free and reduced price lunch status included. Estimates are weighted by total number of students at each campus who take the math ISAT test. Robust standard errors reported in parentheses. R^2 for models: Model 1.1 = 0.0440, Model 1.2 = 0.1203, Model 1.3 = 0.1454, Model 1.4 = 0.1499, Model 1.5 = 0.0340, Model 1.6 = 0.1257, Model 1.7 = 0.1575, and Model 1.8 = 0.1612. Observations for models: Models 1.1-1.3 and 1.5-1.7 = 307,758; Models 1.4 and 1.8 = 307,136.

[a]Suspicious values removed; suspicious values defined by values 1.5 IQR above Q3 and 1.5 IQR below Q1.

***$p < .01$.

the average passing student when controlling for race, ethnicity, and free and reduced price lunch status at the school level.

However, Model 1.1 does not take into consideration time-invariant school characteristics such as school climate, teaching staff, school and district leadership, and peer climate, all of which exert a known influence on student achievement. Consequently, Model 1.2 includes school fixed effects to control for intercampus differences that are stable over time and likely correlated with student achievement gains. Inclusion of school fixed effects also takes into consideration any systematic variation across districts and communities (Hanushek, Kain, & Rivkin, 2004).

Estimates produced by Model 1.2 remained consistent with those reported in the parsimonious form; that is, students that scored below ISAAP's performance threshold on the fall ISAT administration demonstrated greater test-score growth throughout the academic year than similarly situated nonfailing students. Indeed, when controlling for school fixed effects and for race, ethnicity, and free and reduced price lunch status at the school level, the average difference between students that failed to meet the state-defined performance standard on the fall ISAT test and the average passing student was 2.99 points (*SE* = 0.0258). Inclusion of school fixed effects increased the proportion of variance explained to 12.03%.

Model 1.3 includes Grade X Year fixed effects to difference out any instability in ISAT. While NWEA reports attest to the stability, reliability, and validity of ISAT (Hauser & Kingsbury, 2004), empirical research frequently has documented instability in standardized test instruments (Ballou, 2002; Braun, 1988; Heubert & Hauser, 1999; Yen, 1986). Indeed, when controlling for potential changes in test difficulty through inclusion of Grade X Year fixed effects, the estimated effects of failing on student performance hold strongly in comparison to those reported in models 1.1 and 1.2. Specifically, Model 1.3 approximates a test-score gain difference of 2.97 points (*SE* = 0.0255), whereby failing students outperform their nonfailing student counterparts from the fall-to-spring administration of ISAT. In this specification the value on the coefficient of determination increased modestly to 14.54%.

Using Grade X Year effects to control for variation in calibration of test difficulty assumes potential instabilities do not differentially impact certain students and student subgroups. This assumption is potentially problematic in that all Idaho public school

students are required to take the math ISAT assessment two times per year–once in the fall and then again in the spring–yet ISAAP only sanctions schools and districts according to spring outcomes. Suppose a group of ISAT test takers consider the fall ISAT a nuisance and, as a consequence, underperform on the test, thereby causing observed fall-to-spring gain scores to be artificially inflated for these students. If such a scenario is nonrandom, and the magnitude of these test-score gains is not severe enough to be detected by the outlier algorithm employed in this study, then estimates of the effect of failing on student performance will be biased upward. A similar effect also might occur should teachers downplay the significance of the test in the fall.

Given such, Model 1.4 estimates the effect of failing to meet state-defined performance threshold on student performance after removing student observations with "suspicious" gain scores from the sample. Suspicious gain scores are defined as any value 1.5 IQR below Q1 or 1.5 IQR above Q3, but no less than 3 IQR below Q1 or no greater than 3 IQR above Q3, respectively. As indicated in column 5 of Table 4, the coefficient on $F.ISAT.M$ is statistically different from zero at the $\alpha < .01$. The estimated effect suggests that the average failing student's fall-to-spring gain score is 2.93 points (*SE* = 0.0249) greater than a similarly situated nonfailing student.

To put the magnitude of the estimated growth differentials between failing and nonfailing students in perspective, the weighted mean fall-to-spring gain score is 8.91 points with a standard deviation of 6.92 points. Estimates generated from Model 1.4 approximate that the average failing student gains two fifths of a standard deviation more than a nonfailing student in a single school year after controlling for the racial, ethnic, and economic composition of a student's peers as well as time-invariant school

characteristics. In substantive terms, in a single year this average difference is the equivalent of 8 percentile points in the achievement distribution for a below-average third grader.

Do Schools Target Students Expected to Fail?

A concern with the previous estimates stems from $F.ISAT.M$ not taking into consideration the expected test-score gains of Idaho public school students from the fall-to-spring administration of ISAT. Although marginal students just below the performance threshold are labeled failing according to $F.ISAT.M$, in reality these students are expected to meet, or even exceed, ISAAP's defined performance threshold given that the average student's test score increases by eight points from the fall to spring. If schools indeed are targeting resources to failing students, this omission likely biases downward the average test-score gain estimated by $F.ISAT.M$ for students expected to fail the spring ISAT after taking into consideration projected fall-to-spring gains.

Internal test-score reporting under ISAAP intensifies this supposition. NWEA furnishes classroom teachers and building principals with student-specific proficiency reports within 3 days of administration of the fall ISAT. These reports identify (a) students' performance against proficiency growth targets in prior years, (b) the quartile distribution of classrooms against norm groups, and (c) students' projected individual performance on spring administration of ISAT (Northwest Evaluation Association, 2006). Accordingly, teachers and principals are positioned both to make knowledge-based resource-allocation decisions in the short run to avoid ISAAP sanctions and to track progress of current and incoming students over time.

To explore whether reported estimates are negatively biased, models 1.1 through 1.4 are reestimated using $F.ISAT.ADJ.M$. $F.ISAT.ADJ.M$ captures meaningful differences in average student gains by grade given that average fall-to-spring test-score gains ranged from 6.24 points to 12.57 points. The largest observed gains occurred in grades 3 and 4, whereas the smallest observed gains occurred in grades 7 and 8. Indeed, a two-sample *t* test indicated that average gains in grades 3, 4, and 5 were statistically greater than gains in grades 7 and 8 at the $\alpha < .01$ level.

Using $F.ISAT.ADJ.M$ as a proxy for expected student test performance in the spring, the mean percentage of failing students across all years in the sample is calculated as 27.2%, an estimate similar to the 28.3% of students reported by Idaho as having failed the spring ISAT over the 3-year period under study. $F.ISAT.ADJ.M$ also reduces the total number of students identified as failing under the unadjusted indicator variable $F.ISAT.M$ by 50%.

The panel on the right-hand side of Table 4 reports the estimated difference on student test-score gains from fall-to-spring administration of ISAT, taking into account inclusion in the passing achievement distribution of those marginal failing students whose projected fall-to-spring test-score gain would result in their passing the performance threshold. The coefficients on $F.ISAT.ADJ.M$ are statistically different from zero at the $\alpha < .01$ level, and the sign on the coefficients is positive across all specifications. The magnitude of the estimated difference between failing and nonfailing student achievement gains increased in models 1.5 through 1.8, relative to those reported for $F.ISAT.M$ in models 1.1 through 1.4, suggesting that the unadjusted indicator biased downward the average estimated test-score gain for failing students. Indeed, Model 1.8

approximates a difference of 3.68 points (*SE* = 0.0273) between failing and nonfailing students, whereas Model 1.4 approximated a difference of 2.93 points (*SE* = 0.0249).

To test whether the reported differences in models 1.1 through 1.4 and in models 1.5 through 1.8 are statistically different, four unrestricted models were estimated in which $F.ISAT.M$ and $F.ISAT.ADJ.M$ were included. These models took form as:

$$\Delta Y_{ijt} = Y_{ijt} - Y_{ij(t-1)} = \alpha_0 + \alpha_1 F.ISAT.M_{ij(t-1)} + \alpha_2 F.ISAT.M_{ij(t-1)} \times F.ISAT.ADJ.M_{ij(t-1)} + \mathrm{X}_{jt} + \mu_j + \eta_g \times \gamma_t + e_{ijt}.$$

While full results are not reported, the coefficients on the difference-in-difference estimates were statistically significant across all models. $F.ISAT.ADJ.M$ offers strong evidence that Idaho public schools focused on elevating the performance of students expected to fail the spring ISAT test after taking consideration the projected average test-score gain in that grade.

Does a Student's Location Relative to ISAAP's Performance Threshold Matter?

Table 5 reports the estimated effect on student test-score gains of a student's location in the test-score distribution relative to ISAAP's grade-specific performance threshold. $ST.GAP.M$, the indicator variable constructed to capture distance from the performance threshold, is expressed as the additive inverse of the distance of a student's ISAT score from the state-defined performance threshold. $ST.GAP.M$ has a mean value of 1.86, standard deviation of 11.99, and range of 65.81 after stripping out extreme outlier values.

Results reported in Table 5 are similar to those identified and discussed in Table 4. The signs on $ST.GAP.M$ are positive, and the coefficients are statistically different

Table 5

Effect of Student Distance from ISAAP Performance Threshold on Math Performance (General Linear Regression Model with Math Gain Score Fall-to-Spring as the Dependent Variable)

	Student Distance from ISAAP Performance Threshold (*ST.GAP.M)*				Student Distance from ISAAP Performance Threshold, Adjusted for Expected Gain *(ST.GAP.ADJ.M)*			
	None	School	School Grade X Year	School Grade X Year[a]	None	School	School Grade X Year	School Grade X Year[a]
Variable	Model 2.1	Model 2.2	Model 2.3	Model 2.4	Model 2.5	Model 2.6	Model 2.7	Model 2.8
ST.GAP.M	0.1654*** (0.0010)	0.1696*** (0.0009)	0.1698*** (0.0009)	0.1633*** (0.0009)	--	--	--	--
ST.GAP.ADJ.M	--	--	--	--	1.142*** (0.0068)	1.201*** (0.0066)	1.221*** (0.0065)	1.1762*** (0.0064)

Note. School-level controls for race/ethnicity and free and reduced price lunch status included. Estimates are weighted by total number of students at each campus who take the math ISAT test. Robust standard errors reported in parentheses. R^2 for models: Model 2.1 = 0.0915, Model 2.2 = 0.1716, Model 2.3 = 0.1967, Model 2.4 = 0.1968, Model 2.5 = 0.0867, Model 2.6 = 0.1712, Model 2.7 = 0.1992, and Model 2.8 = 0.1992. Observations for models: Models 2.1-2.3 and 2.5-2.7 = 307,758; Models 2.4 and 2.8 = 307,136.

[a]Suspicious values removed; suspicious values defined by values 1.5 IQR above Q3 and 1.5 IQR below Q1.

***$p < .01$.

from zero at the $\alpha < .01$ level. The estimated effect in Model 2.4, for example, suggests that the average failing student's fall-to-spring gain score is 0.16 points (SE = 0.0009) greater for each unit a student is away from ISAT's grade-specific performance threshold when controlling for race/ethnicity, free and reduced lunch status, school fixed effects, and Grade X Year fixed effects. Furthermore, the percentage of variance explained around the mean reaches 19.7% in models 2.3 and 2.4.

Models 2.5 through 2.8 report estimated differences in test-score gains and losses for students across the achievement distribution after adjusting $ST.GAP.M$ for average expected fall-to-spring test score gains in math by grade. The adjusted variable is

identified as $ST.GAP.ADJ.M$. The signs on $ST.GAP.ADJ.M$ are positive, and the coefficients are statistically different from zero at the $\alpha < .01$ level. Once again, the percentage of variance explained around the mean reaches 19.9% in the complete model specifications, which include controls for race/ethnicity, free and reduced lunch status, school fixed effects, and Grade X Year fixed effects.

Even though model specifications using $ST.GAP.M$ and $ST.GAP.ADJ.M$ as indicator variables indicate a positive slope across their respective distributions with average gains increasing monotonically from the highest to lowest achieving students, these specifications impose a strong assumption on the relationship between students' test-score gains and their location in the achievement distribution; that being, the amount by which the gain increases for a student who starts at a particular point below the cut off will equal the amount by which the gain diminishes for a student who starts at the same point but above the cut off. By imposing such a strong assumption on the relationship of $ST.GAP.M$ and $ST.GAP.ADJ.M$ to student gains, the estimates are likely to be biased.

As such, Table 6 displays estimated differences in test-score gains and losses for students across the achievement distribution allowing for completely different outcomes when a student passes or fails the fall ISAT. The unadjusted passing and failing $ST.GAP.M$ variables, denoted as $ST.GAP.PASS.M$ and $ST.GAP.FAIL.M$, are reported in models 3.1 through 3.4, respectively. The adjusted passing and failing $ST.GAP.M$ variables, denoted as $ST.GAP.PASS.ADJ.M$ and $ST.GAP.FAIL.ADJ.M$, are reported in models 3.5 through 3.8, respectively. The coefficients across all model specifications are positive and statistically different from zero at the $\alpha < .01$ level.

Table 6

Effect of Student Distance from ISAAP Performance Threshold on Math Performance Using ST.GAP.FAIL.M and ST.GAP.PASS.M (General Linear Regression Model with Math Gain Score Fall-to-Spring as the Dependent Variable)

	Student Distance from ISAAP Performance Threshold *(ST.GAP.FAIL.M and ST.GAP.PASS.M)*				Student Distance from ISAAP Performance Threshold, Adjusted for Expected Gain *(ST.GAP.FAIL.ADJ.M and ST.GAP.PASS.ADJ.M)*			
	None	School	School Grade X Year	School Grade X Year[a]	None	School	School Grade X Year	School Grade X Year[a]
Variable	Model 3.1	Model 3.2	Model 3.3	Model 3.4	Model 3.5	Model 3.6	Model 3.7	Model 3.8
ST.GAP.FAIL.M	0.1968*** (0.0016)	0.2264*** (0.0015)	0.2329*** (0.0015)	0.2186*** (0.0015)	--	--	--	--
ST.GAP.PASS.M	0.1198*** (0.0021)	0.0861*** (0.0020)	0.0769*** (0.0020)	0.0832*** (0.0020)	--	--	--	--
ST.GAP.FAIL.ADJ.M	--	--	--	--	1.0084*** (0.0108)	1.4554*** (0.0107)	1.6907*** (0.0108)	1.5914*** (0.0107)
ST.GAP.PASS.ADJ.M	--	--	--	--	1.3374*** (0.0141)	0.8294*** (0.0140)	0.5379*** (0.0141)	0.5813*** (0.0138)

Note. School-level controls for race/ethnicity and free and reduced price lunch status included. Estimates are weighted by total number of students at each campus who take the math ISAT test. Robust standard errors reported in parentheses. R^2 for models: Model 3.1 = 0.0933, Model 3.2 = 0.1774, Model 3.3 = 0.2038, Model 3.4 = 0.2024, Model 3.5 = 0.0875, Model 3.6 = 0.1737, Model 3.7 = 0.2069, and Model 3.8 = 0.2054. Observations for models: Models 3.1-3.3 and 3.5-3.7 = 307,758; Models 3.4 and 3.8 = 307,136.

[a]Suspicious values removed; suspicious values defined by values 1.5 IQR above Q3 and 1.5 IQR below Q1.

***$p < .01$.

Allowing for completely different outcomes when $ST.GAP.M$ and $ST.GAP.ADJ.M$ are negative and positive alters the magnitude of the reported estimates. To illustrate, Figure 3 juxtaposes average differences in fall-to-spring test-score gains in math for $ST.GAP.ADJ.M$ against those for $ST.GAP.FAIL.ADJ.M$ and $ST.GAP.PASS.ADJ.M$. Coefficients on $ST.GAP.FAIL.M$ and $ST.GAP.FAIL.ADJ.M$

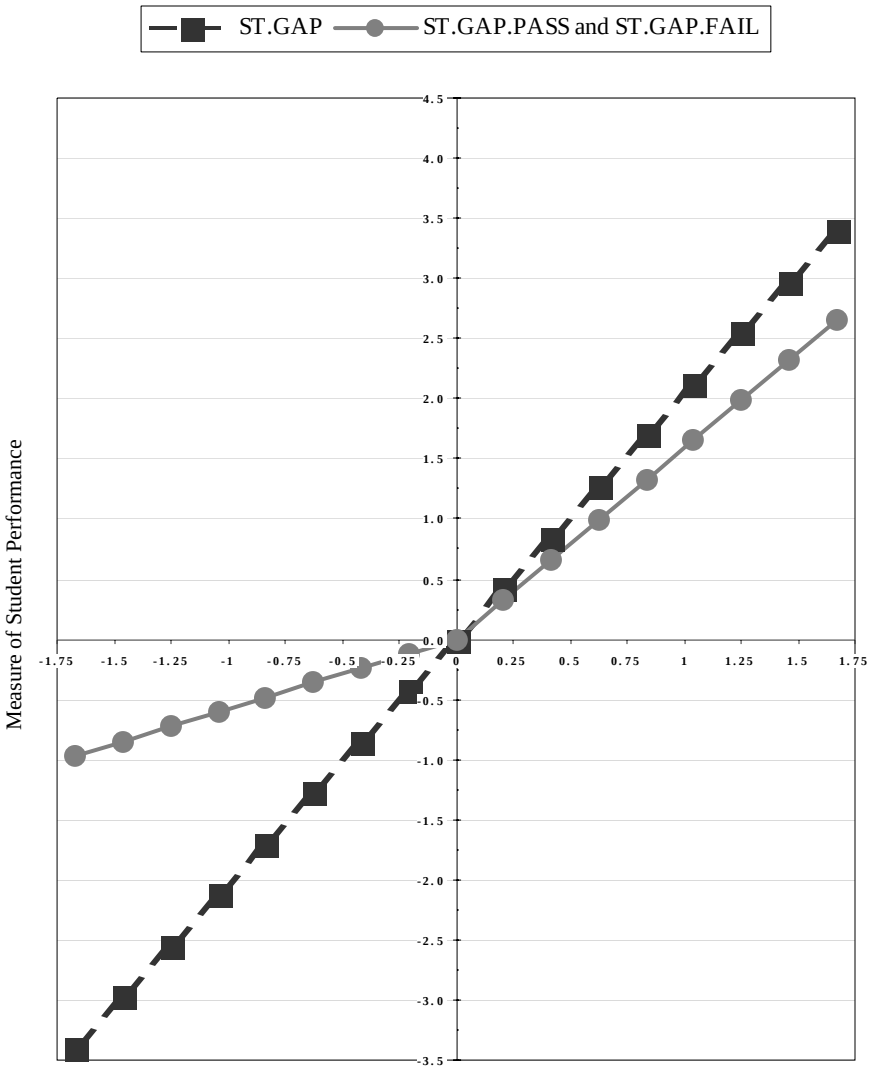

Figure 3. Effect of distance from ISAT performance threshold on math performance.

increased in size whereas the coefficient on *ST.GAP.PASS.M* and *ST.GAP.PASS.ADJ.M* decreased in size in relation to the magnitude of the coefficient values reported in Table 5. Nevertheless, these findings suggest that traditional public schools in Idaho are elevating the average performance of failing students relative to that of passing students.

As previously discussed, regression to the mean potentially confounds identification of achievement tradeoffs by leading to overstatement of schools' actual responses to Idaho's minimum competency accountability program. Accordingly, models were reestimated using a standardized fall-to-spring student gain score in math as the dependent variable. The coefficient estimates on $F.ISAT.M$, $F.ISAT.ADJ.M$, $ST.GAP.M$, $ST.GAP.ADJ.M$, $ST.GAP.FAIL.M$, $ST.GAP.PASS.M$, $ST.GAP.FAIL.ADJ.M$, and $ST.GAP.PASS.ADJ.M$ are robust across specifications using a standardized fall-to-spring gain score in math.

The preceding section explored several variations on the two primary indicator variables used to denote whether a student was likely to be targeted and specified a series of models thereon, culminating in a complete model specification that controlled for the impact of peer composition, school effects, and Grade X Year effects. In addition to the importance of including a full set of controls, I found $F.ISAT.ADJ.M$ and $ST.GAP.FAIL.ADJ.M$ and $ST.GAP.PASS.ADJ.M$ to be superior indicators of whether a student was likely to be targeted by schools responding to Idaho's minimum competency accountability program. As such, subsequent reporting of the Student X Subject interaction general linear models focus on estimates for indicator variables $F.ISAT.ADJ.M$, $ST.GAP.FAIL.ADJ.M$, and $ST.GAP.PASS.ADJ.M$ when controlling for peer composition, school effects, and Grade X Year effects.

Have Idaho Schools Responded to ISAAP by Raising the Achievement of Failing Students and Failing Student Subgroups?

The following section reports results from a more refined series of model specifications used to estimate if schools strategically targeted resources to failing

students and student subgroups. Table 7 displays estimates from the Student X Subject interaction model. The α_1 coefficient, denoted by $F.PROF.M$, is a key parameter of interest in detecting high-/low-achievement tradeoffs. The sign on the $F.PROF.M$ coefficient is positive, and the value is statistically different from zero at the $\alpha < .05$ level. This estimate indicates that passing non-Hispanic and passing non-White students enrolled in failing schools gained, on average, 0.16 points more (*SE* = 0.0738) than non-failing, non-Hispanic and non-failing, non-White students enrolled in passing schools.

As indicated in models 5.7 and 5.8 in Table 8, the estimates on the α_1 coefficient are not statistically significant when $ST.GAP.FAIL.ADJ.M$ and $ST.GAP.PASS.ADJ.M$ are used as indicator variables. These estimates indicate that average student gains for passing non-Hispanic and passing non-White students enrolled in failing schools are not statistically different from passing non-Hispanic and passing non-White students enrolled in schools that met ISAAP's proficiency standard.

The α_4 and α_5 coefficients, denoted by $F.PROF.M \times HISPANIC$ and $F.PROF.M \times WHITE$, are also key parameters of interest (Table 7). The sign on $F.PROF.M \times HISPANIC$ is positive, and the value is statistically different from zero at the $\alpha < .01$ level. Model 4.8 reports that Hispanic students who attended a school that failed to meet ISAAP due to poor performance by Hispanic students gained, on average, 0.24 points (*SE* = 0.0764) more than student who are not in the failing subgroup but who also attended a school that failed. This finding suggests that schools that failed to meet ISAAP proficiency standard due to poor performance by Hispanic students in math targeted resources to the Hispanic subgroup.

Table 7

Effect of ISAAP on Math Performance Using a Student X Subject Interaction Model and Failed ISAT (General Linear Regression Model with Math Gain Score Fall-to-Spring as the Dependent Variable)

	School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Failed Fall ISAT Math Test (*F.ISAT.M*), Student Part of Failing Subgroup (*F.SUB X M*)				School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Failed Fall ISAT Math Test (*F.ISAT.ADJ.M*), Student Part of Failing Subgroup (*F.SUB X M*)			
	None	School	School Grade X Year	School Grade X Year[a]	None	School	School Grade X Year	School Grade X Year[a]
Variable	Model 4.1	Model 4.2	Model 4.3	Model 4.4	Model 4.5	Model 4.6	Model 4.7	Model 4.8
F.PROF.M (α_1)	-0.1696**	0.6481***	0.0264	0.0453	-0.0567	0.8000***	0.1519**	0.1603**
	(0.0739)	(0.0826)	(0.0869)	(0.0847)	(0.0609)	(0.0709)	(0.0757)	(0.0738)
HISPANIC (α_2)	-0.8045***	-0.8307***	-0.7790***	-0.7341***	-0.6250***	-0.6609***	-0.6579***	-0.6386***
	(0.1286)	(0.1269)	(0.1260)	(0.1227)	(0.0977)	(0.0964)	(0.0955)	(0.0931)
WHITE (α_3)	0.1778*	0.1655*	0.1832*	0.1879*	0.1590**	0.1493*	0.1697**	0.1579**
	(0.1020)	(0.1006)	(0.0999)	(0.0973)	(0.0810)	(0.0798)	(0.0791)	(0.0771)
F.PROF.M X F.SUB.HISP.M (α_4)	-0.0398	0.0671	0.1088	0.0950	0.0692	0.2023***	0.2516***	0.2388***
	(0.0809)	(0.0908)	(0.0906)	(0.0882)	(0.0669)	(0.0785)	(0.0783)	(0.0764)
F.PROF.M X F.SUB.WHITE.M (α_5)	0.0537	1.0952***	0.7718***	0.7054***	0.2498**	1.3808***	1.0542***	0.9738***
	(0.1421)	(0.1551)	(0.1554)	(0.1514)	(0.1128)	(0.1289)	(0.1293)	(0.1260)
F.PROF.M X F.ISAT.M (α_6)	0.9792***	0.8511***	0.7847***	0.7415***	--	--	--	--
	(0.1006)	(0.1001)	(0.0994)	(0.0969)				
F.PROF.M X F.ISAT.ADJ.M ($\alpha_{6'}$)	--	--	--	--	1.4629***	1.2194***	1.0933***	1.0480***
					(0.1085)	(0.1074)	(0.1066)	(0.1039)
F.PROF.M X F.SUB.HISP.M X F.ISAT.M (α_7)	0.2025*	0.2136**	0.2096**	0.2380**	--	--	--	--
	(0.1082)	(0.1076)	(0.1068)	(0.1040)				
F.PROF.M X F.SUB.WHITE.M X F.ISAT.M (α_8)	0.1798	0.2403	0.2839	0.2587	--	--	--	--
	(0.1808)	(0.1790)	(0.1777)	(0.1731)				
F.PROF.M X F.SUB.HISP.M X F.ISAT.ADJ.M ($\alpha_{7'}$)	--	--	--	--	-0.1286	-0.0953	-0.1083	-0.0622
					(0.1156)	(0.1144)	(0.1134)	(0.1106)

(*table continues*)

Table 7 (continued)

	School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Failed Fall ISAT Math Test (*F.ISAT.M*), Student Part of Failing Subgroup (*F.SUB X M*)				School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Failed Fall ISAT Math Test (*F.ISAT.ADJ.M*), Student Part of Failing Subgroup (*F.SUB X M*)			
	None	School	School Grade X Year	School Grade X Year[a]	None	School	School Grade X Year	School Grade X Year[a]
Variable	Model 4.1	Model 4.2	Model 4.3	Model 4.4	Model 4.5	Model 4.6	Model 4.7	Model 4.8
F.PROF.M X F.SUB.WHITE.M X F.ISAT.ADJ.M ($\alpha_{8'}$)	--	--	--	--	-0.2927 (0.1798)	-0.3399** (0.1777)	-0.2786 (0.1761)	-0.2782 (0.1717)
F.ISAT.M (α_9)	2.7862*** (0.1308)	2.8972*** (0.1293)	2.9258*** (0.1284)	2.9384*** (0.1251)	--	--	--	--
F.ISAT.ADJ.M ($\alpha_{9'}$)	--	--	--	--	3.3285*** (0.1337)	3.4931*** (0.1322)	3.6529*** (0.1310)	3.6074*** (0.1278)
F.ISAT.M X HISPANIC (α_{10})	0.4111*** (0.1585)	0.4512*** (0.1565)	0.3894** (0.1554)	0.3201** (0.1514)	--	--	--	--
F.ISAT.M X WHITE (α_{11})	-0.2927*** (0.1326)	-0.2604** (0.1310)	-0.2900*** (0.1300)	-0.3370*** (0.1267)	--	--	--	--
F.ISAT.ADJ.M X HISPANIC ($\alpha_{10'}$)	--	--	--	--	0.0979 (0.1530)	0.1769 (0.1512)	0.1510 (0.1498)	0.1149 (0.1461)
F.ISAT.ADJ.M X WHITE ($\alpha_{11'}$)	--	--	--	--	-0.2355* (0.1357)	-0.1833 (0.1341)	-0.2174 (0.1329)	-0.2590** (0.1296)

Note. School-level controls for race/ethnicity and free and reduced price lunch status included. Estimates are weighted by total number of students at each campus who take the math ISAT test. Robust standard errors reported in parentheses. R^2 for models: Model 4.1 = 0.1089, Model 4.2 = 0.1351, Model 4.3 = 0.1476, Model 4.4 = 0.1520, Model 4.5 = 0.1177, Model 4.6 = 0.1447, Model 4.7 = 0.1598, and Model 4.8 = 0.1635. Observations for models: Models 4.1-4.3 and 4.5-4.7 = 307,758; Models 4.4 and 4.8 = 307,136.

[a]Suspicious values removed; suspicious values defined by values 1.5 IQR above Q3 and 1.5 IQR below Q1.

* $p < .10$. ** $p < .05$. *** $p < .01$.

Table 8

Effect of ISAAP on Math Performance Using a Student X Subject Interaction Model and Student Gap (General Linear Regression Model with Math Gain Score Fall-to-Spring as the Dependent Variable)

	School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Part of Failing Subgroup (*F.SUB X M*), Student Distance from ISAAP Performance Threshold (*ST.GAP.M*)				School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Part of Failing Subgroup (*F.SUB X M*), Student Distance from ISAAP Performance Threshold (*ST.GAP.ADJ.M*)			
	None	School	School Grade X Year	School Grade X Year[a]	None	School	School Grade X Year	School Grade X Year[a]
Variable	Model 5.1	Model 5.2	Model 5.3	Model 5.4	Model 5.5	Model 5.6	Model 5.7	Model 5.8
F.PROF.M (α_1)	-0.2501*** (0.0798)	-0.2501*** (0.0798)	-0.0675 (0.0903)	-0.0170 (0.0887)	-0.3294*** (0.0771)	0.5717*** (0.0847)	-0.1155 (0.0879)	-0.0631 (0.0863)
HISPANIC (α_2)	-1.3895*** (0.1187)	-1.3895*** (0.1187)	-1.4449*** (0.1160)	-1.3867*** (0.1145)	-1.1268*** (0.1104)	-1.2295*** (0.1089)	-1.2390*** (0.1079)	-1.1861*** (0.1063)
WHITE (α_3)	0.3498*** (0.1015)	0.3498*** (0.1015)	0.3504*** (0.0992)	0.3945*** (0.0980)	0.3663*** (0.0948)	0.3620*** (0.0934)	0.3621*** (0.0926)	0.3994*** (0.0912)
F.PROF.M X F.SUB.HISP.M (α_4)	0.2819*** (0.0868)	0.2819*** (0.0868)	0.4211*** (0.0946)	0.3829*** (0.0930)	0.3364*** (0.0841)	0.4744*** (0.0928)	0.4930*** (0.0924)	0.4512*** (0.0909)
F.PROF.M X F.SUB.WHITE.M (α_5)	0.0499 (0.1407)	0.1407 (0.1407)	1.1293*** (0.1524)	1.0895*** (0.1502)	-0.0843 (0.1364)	1.3722*** (0.1491)	1.0637*** (0.1487)	0.9995*** (0.1464)
F.PROF.M X ST.GAP.FAIL.M (α_6)	0.0833*** (0.0059)	0.0833*** (0.0059)	0.0673*** (0.0058)	0.0607*** (0.0058)	--	--	--	--
F.PROF.M X ST.GAP.PASS.M (α_7)	-0.0230*** (0.0077)	-0.0230*** (0.0077)	-0.0291*** (0.0076)	-0.0242*** (0.0075)	--	--	--	--
F.PROF.M X ST.GAP.FAIL. ADJ.M ($\alpha_{6'}$)	--	--	--	--	0.8175*** (0.0426)	0.7023*** (0.0422)	0.6190*** (0.0420)	0.5659*** (0.0420)
F.PROF.M X ST.GAP.PASS. ADJ.M ($\alpha_{7'}$)	--	--	--	--	-0.2336*** (0.0535)	-0.2446*** (0.0536)	-0.1955*** (0.0534)	-0.1575*** (0.0526)
F.PROF.M X F.SUB.HISP.M X ST.GAP.FAIL.M (α_8)	-0.0161*** (0.0062)	-0.0161*** (0.0062)	-0.0158*** (0.0061)	-0.0118* (0.0061)	--	--	--	--

(*table continues*)

Table 8 (continued)

	School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Part of Failing Subgroup (*F.SUB X M*), Student Distance from ISAAP Performance Threshold (*ST.GAP.M*)				School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Part of Failing Subgroup (*F.SUB X M*), Student Distance from ISAAP Performance Threshold (*ST.GAP.ADJ.M*)			
	None	School	School Grade X Year	School Grade X Year[a]	None	School	School Grade X Year	School Grade X Year[a]
Variable	Model 5.1	Model 5.2	Model 5.3	Model 5.4	Model 5.5	Model 5.6	Model 5.7	Model 5.8
F.PROF.M X F.SUB.WHITE.M X ST.GAP. FAIL.M (α_{10})	-0.0213** (0.0090)	-0.0213** (0.0090)	-0.0211** (0.0088)	-0.0239*** (0.0089)	--	--	--	--
F.PROF.M X F.SUB.HISP.M X ST.GAP.PASS.M (α_{9})	0.0235*** (0.0083)	0.0235*** (0.0083)	0.0226*** (0.0082)	0.0200** (0.0081)	--	--	--	--
F.PROF.M X F.SUB.WHITE.M X ST.GAP. PASS.M (α_{11})	0.0056 (0.0158)	0.0056 (0.0158)	0.0126 (0.0155)	0.0182 (0.0152)	--	--	--	--
F.PROF.M X F.SUB.HISP.M X ST.GAP.FAIL. ADJ.M (α_{8})	--	--	--	--	-0.2155*** (0.0451)	-0.1732*** (0.0447)	-0.1738*** (0.0443)	-0.1410*** (0.0444)
F.PROF.M X F.SUB.WHITE.M X ST.GAP.FAIL. ADJ.M (α_{10})	--	--	--	--	-0.0594 (0.0656)	-0.1035 (0.0650)	-0.0828 (0.0644)	-0.0893 (0.0649)
F.PROF.M X F.SUB.HISP.M X ST.GAP.PASS. ADJ.M (α_{9})	--	--	--	--	0.2156*** (0.0581)	0.1768*** (0.0581)	0.1758*** (0.0576)	0.1541*** (0.0567)
F.PROF.M X F.SUB.WHITE.M X ST.GAP.PASS. ADJ.M (α_{11})	--	--	--	--	-0.0736 (0.1140)	0.0149 (0.1131)	-0.0087 (0.1122)	0.0212 (0.1100)
ST.GAP.FAIL.M (α_{12})	0.2004*** (0.0068)	0.2004*** (0.0068)	0.2114*** (0.0067)	0.2031*** (0.0067)	--	--	--	--
ST.GAP.PASS.M (α_{13})	0.0656*** (0.0102)	0.0656*** (0.0102)	0.0743*** (0.0099)	0.0754*** (0.0099)	--	--	--	--

(*table continues*)

Table 8 (continued)

	School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Part of Failing Subgroup (*F.SUB X M*), Student Distance from ISAAP Performance Threshold (*ST.GAP.M*)				School Failed to Meet ISAAP Proficiency Standard (*F.PROF.M*), Student Part of Failing Subgroup (*F.SUB X M*), Student Distance from ISAAP Performance Threshold (*ST.GAP.ADJ.M*)			
	None	School	School Grade X Year	School Grade X Year[a]	None	School	School Grade X Year	School Grade X Year[a]
Variable	Model 5.1	Model 5.2	Model 5.3	Model 5.4	Model 5.5	Model 5.6	Model 5.7	Model 5.8
ST.GAP.FAIL. ADJ.M (α_{12})	--	--	--	--	1.3412*** (0.0457)	1.4105*** (0.0452)	1.5194*** (0.0448)	1.4611*** (0.0449)
ST.GAP.PASS. ADJ.M (α_{13})	--	--	--	--	0.5667*** (0.0695)	0.6171*** (0.0684)	0.4959*** (0.0678)	0.5061*** (0.0677)
ST.GAP.FAIL.M X HISPANIC (α_{14})	0.0409*** (0.0077)	0.0409*** (0.0077)	0.0443*** (0.0075)	0.0420*** (0.0075)	--	--	--	--
ST.GAP.FAIL.M X WHITE (α_{15})	0.0007 (0.0069)	0.0007 (0.0069)	0.0020 (0.0068)	-0.0075*** (0.0068)	--	--	--	--
ST.GAP.PASS.M X HISPANIC (α_{16})	-0.0371*** (0.0141)	-0.0371*** (0.0141)	-0.0452*** (0.0138)	-0.0414* (0.0137)	--	--	--	--
ST.GAP.PASS.M X WHITE (α_{17})	0.0158 (0.0103)	0.0158 (0.0103)	0.0141 (0.0101)	0.0196 (0.0101)	--	--	--	--
ST.GAP.FAIL. ADJ.M X HISPANIC ($\alpha_{14'}$)	--	--	--	--	0.2120*** (0.0516)	0.2672*** (0.0510)	0.2614*** (0.0505)	0.2459*** (0.0505)
ST.GAP.FAIL. ADJ.M X WHITE ($\alpha_{15'}$)	--	--	--	--	-0.0171 (0.0466)	0.0023 (0.0460)	0.0069 (0.0455)	-0.0593 (0.0456)
ST.GAP.PASS. ADJ.M X HISPANIC ($\alpha_{16'}$)	--	--	--	--	-0.0402 (0.0975)	-0.1072 (0.0961)	-0.1788* (0.0952)	-0.1530 (0.0944)
ST.GAP.PASS. ADJ.M X WHITE ($\alpha_{17'}$)	--	--	--	--	0.1990* (0.0703)	0.1209* (0.0692)	0.1067 (0.0685)	0.1418** (0.0683)

Note. School-level controls for race/ethnicity and free and reduced price lunch status included. Estimates are weighted by total number of students at each campus who take the math ISAT test. Robust standard errors reported in parentheses. R^2 for models: Model 5.1 = 0.1675, Model 5.2 = 0.1675, Model 5.3 = 0.2081, Model 5.4 = 0.2049, Model 5.5 = 0.1673, Model 5.6 = 0.1969, Model 5.7 = 0.2114, and Model 5.8 = 0.2079. Observations for models: Models 5.1-5.3 and 5.5-5.7 = 307,758; Models 5.4 and 5.8 = 306,922.

[a]Suspicious values removed; suspicious values defined by values 1.5 IQR above Q3 and 1.5 IQR below Q1.

* $p < .10$. ** $p < .05$. *** $p < .01$.

The sign on $F.PROF.M \times WHITE$ is positive, and the value is statistically different from zero at the $\alpha < .01$ level. Model 4.8 reports that White students who attended a school that failed to meet ISAAP due to poor performance by White students, on average, gained almost 1 point (*SE* = 0.1260) more than students who are not in the failing subgroup but who also attended a school that failed. This finding suggests that schools that failed to meet ISAAP proficiency standard due to poor performance by White students in math targeted resources to White subgroup.

The magnitude of the differential test-score gain estimates on the α_4 and α_5 coefficients is remarkably similar to that produced in models when $ST.GAP.FAIL.ADJ.M$ and $ST.GAP.PASS.ADJ.M$ are used as indicator variables (Table 8). The magnitude of the Hispanic subgroup interaction increased modestly in size from 0.24 to 0.45, whereas the estimate on the White subgroup interaction remained virtually identical to that reported earlier (i.e., value of 0.99 with *SE* = 0.1464). These findings provide strong evidence that schools that failed to meet ISAAP proficiency standard responded by elevating the achievement of students within that subgroup.

The value on α_6, identified by $F.PROF.M \times F.ISAT.ADJ.M$, is the failing student interaction (Table 7). The sign on the failing student interaction is positive, and the value is statistically different from zero at the $\alpha < .01$. The estimate suggests that the average student who is expected to fail the spring ISAT and who attended a school that failed to meet ISAAP proficiency standard in the year prior gained 1.05 points more (*SE* = 0.1039) than passing students that attended a school that failed. While this provides strong evidence that schools targeted resources to students expected to fail spring ISAT test, this strategic resource targeting does not appear to work to the detriment of students

expected to pass spring administration of ISAT. These results suggest that traditional Idaho public schools are operating in the absence of resource constraints. That is, in the face of increased accountability and threat of sanction, there is sufficient slack in the operation of traditional public schools in Idaho enabling schools elevate outcomes of low-performing students without diverting attention and resources away from other students.

Models using $ST.GAP.FAIL.ADJ.M$ and $ST.GAP.PASS.ADJ.M$ as indicator variables (Table 8) provide further illustration of the conclusion that gains have been realized through increased accountability and threat of sanction. Take, for example, the value on $\alpha_{7'}$, identified by $F.PROF.M \times ST.GAP.PASS.ADJ.M$. The sign on the passing student conditioning effect is negative, and the value is statistically different from zero at the $\alpha < .01$. The estimate suggests that the average student who is least expected to fail the spring ISAT and who attended a school that failed to meet ISAAP proficiency standard in the year prior will have greater gains than other failing students that are closer to ISAT performance threshold.

Nevertheless, estimates on the student interaction and subject interaction do not fully explain the issue of interest, namely if schools responded to ISAAP's incentive structure by targeting resources to failing students in failing subgroups. To provide a more refined specification of resource targeting, a three-way interaction is introduced. This three-way interaction permits examination of whether Idaho schools responded to ISAAP by targeting resources to failing students in failing subgroups at the expense of similarly situated peers who met proficiency.

The coefficient on α_7, identified by $F.PROF.M \times F.SUB.HISP.M \times F.ISAT.M$, is the failing student by Hispanic subgroup interaction. The sign on the α_7 coefficient is negative. However, the value on α_7 coefficient is not statistically different from zero. This finding suggests that Hispanic students who failed math ISAT enrolled in a school that failed to meet ISAAP proficiency standard due to math performance by Hispanic subgroup do not differ from non-Hispanic, failing students in this type of failing schools. That is, schools did not target resources to failing students in the subgroup for which a school failed to meet ISAAP proficiency standard.

The coefficient on α_8 is the failing student by White subgroup interaction, identified by $F.PROF.M \times F.SUB.WHITE.M \times F.ISAT.M$. The sign on the α_8 coefficient is negative. However, the value on α_7 coefficient is not statistically different from zero. This suggests that White students who failed math ISAT enrolled in a school that failed to meet ISAAP proficiency standard due to math performance by Hispanic subgroup do not differ from non-White, failing students in this type of failing schools. That is, schools are not targeting resources to failing students in the subgroup for which a school failed to meet ISAAP proficiency standard.

While estimates reported in models 4.7 and 4.8 (Table 7) for the failing student by White subgroup are consistent with those reported in models 5.7 and 5.8 (Table 8), the failing student by Hispanic subgroup interaction effect is statistically significant when $ST.GAP.FAIL.ADJ.M$ and $ST.GAP.PASS.ADJ.M$ are used as indicator variables. The estimated effect on α_8 is negative, suggesting that failing Hispanic students enrolled in

schools that failed to meet ISAAP proficiency standard do not gain, on average, less than passing peers.

Do Estimated Effects Persist After Standardization of Students' Fall-to-Spring Gain Scores?

Each of the models reported have produced substantively and statistically significant indications of strategic resource allocation decision making by Idaho public schools in response to the state's minimum competency accountability plan. Evidence has been relatively consistent across model specifications. Recognizing that previous research on achievement tradeoffs has questioned whether estimates are robust to influence of mean reverting bias (Reback, 2006), this study also modeled all specifications using a standardized fall-to-spring gain score in math.

To test for robustness of the complete model to mean reversion, I followed a standardization procedure similar to that employed by Hanushek et al. (2005), whereby the initial achievement distribution in Idaho is defined by fall ISAT score and divided into 20 equal intervals for each unique combination of year and grade. The mean and standard deviation test-score gain for all students starting in a particular interval for each unique combination of year and grade are then computed. The standard fall-to-spring test-score gain for each student is calculated as the difference between that student's nominal gain and the mean gain of all students in the interval over the standard deviation of all student gains in the interval. Thus, gains in each interval are distributed with a mean of zero and standard deviation of one.

Results of subsequent analyses using standardized fall-to-spring gain score in math as the dependent variables suggest that initial findings on the parameters of interest

are not an artifact of mean reversion. This finding likely is due to the fact that the estimate on α_9 censored the influence of mean reversion. The magnitude of the estimate on α_9 reduced relative to the magnitude of other coefficients, suggesting that mean reversion was a source of bias on the α_9 coefficient in earlier estimates. The estimates on α_9 , as expected, were 0.

Does Increased Market Sensitivity Found in Public Charter Schools Impact Resource-Allocation Practices?

Economist Milton Friedman asserted that introducing school choice into the public education system would encourage competition, which, in turn, would provoke traditional public schools to become more productive. Some 40 years after Friedman's free-market arguments, an increasing number of parents now can choose from a number of school choice options, including magnet schools, charter schools, cyber-schools, private or independent schools, school vouchers, tax credits and deductions, and home schooling. Indeed, between 1993 and 2003, the percentage of students in 1st through 12th grade attending their "assigned" public school decreased from 80% to 74%, in comparison with an increase of 2.7 million students attending public schools of choice (National Center for Educational Statistics, 2004).

Charter schools are held accountable under NCLB and, hence, subject to the same sanctions as traditional public schools. Thus, there are incentives for charter schools to engage in the same strategic tradeoffs as regular public schools. However, charter schools, by virtue of funding and enrollment policy, are more market sensitive than their

traditional public school counterparts, a condition that may attenuate allocation of resources to particular groups of students.

This potential effect could manifest itself in a number of ways. For instance, a charter school may be more motivated than traditional public schools to avoid the stigma associated with AYP designation and any sanctions levied thereafter given the expense associated with operating a relatively new institution and the potentially crippling implications of an exodus of students that might, in comparison, represent no more than natural attrition to a larger, better-funded, and more stable public school. In a similar vein, a charter school is more likely to be resource constrained (Speakman, Finn, & Hassel, 2005), at least in the near term, thereby increasing the likelihood that resources could and would be diverted from certain students and student subgroups in order to elevate the performance of low-performing peers.

Nevertheless, it is clear that school choice remains in its incipient stages in Idaho. There are only 17 charter schools in Idaho, of which only 6, 10, and 13 met ISAAP's minimum *n* requirement during the 2002-03, 2003-04, and 2004-05 school years, respectively. Moreover, only two schools failed to meet AYP across all years and the total enrollment between the two schools was 156 students (113 and 43). As such, any meaningful comparisons between traditional public schools and public charter schools will require greater market penetration by charter institutions.

CHAPTER VI

FINDINGS AND POLICY IMPLICATIONS, DIRECTIONS FOR FUTURE RESEARCH, AND CONCLUSIONS

This dissertation investigated the strategic resource-allocation decision making of traditional public schools and public charter schools seeking to close the achievement gap among low- and high-performing students in response to the No Child Left Behind Act of 2001, as well as factors hypothesized to explain this decision making. Both student-level, longitudinal test-score data furnished by the Northwest Evaluation Association and school-level accountability data collected from the Idaho State Department of Education were analyzed.

Recognizing that no formal accounting system tracks allocation of resources at the student or classroom level, distributional inequities in student achievement among low- and high-performing student achievement were used to infer a reprioritization of intraschool resources. Specifically, an NCLB-induced resource-allocation decision was detected if a greater than expected increase in the achievement of traditionally low-performing students occurred in tandem with a less than expected increase in the achievement of traditionally high-performing students. Key findings, implications for education policy, and directions for future research are summarized below.

Findings and Policy Implications

There was strong evidence that Idaho public schools responded to NCLB by raising the achievement of failing students relative to that of passing students. Findings

were consistent across all indicator variables constructed to capture the likelihood a student was targeted with resources. Moreover, all estimates were robust across model specifications when a standardized fall-to-spring test-score gain in math was introduced as the dependent variable to gauge the potentially confounding influence of mean reverting measurement error on estimates.

There was strong evidence that Idaho public schools responded to ISAAP's threat of sanction by raising the achievement of students in failing subgroups. Hispanic students who attended a school that failed to meet ISAAP's proficiency standard due to poor performance by Hispanic students gained, on average, between 0.24 and 0.45 points more than students who are not in the failing subgroup but also attended a school that failed. White students who attended a school that failed due to poor performance by White students in math gained, on average, between 0.97 and 0.99 points more than students who are not in the failing subgroup but who also attended a school that failed. All estimates were robust across model specifications when a standardized fall-to-spring gain score in math was used as the dependent variable.

Evidence on whether failing schools target failing students in failing subgroups was inconsistent. While failing White students in schools that failed to meet ISAAP's performance standard due to poor performance in math by White students gained no more or less than non-White, failing students in failing schools, estimated results on the failing student by Hispanic subgroup conditioning effect were split. One set of estimates suggested that the average test-score gain difference between failing Hispanic students in schools that did not meet ISAAP proficiency standard due to poor performance by Hispanic students and nonfailing students were statistically insignificant. However, when

ST.GAP.PASS.ADJ.M and *ST.GAP.FAIL.ADJ.M* were used as the indicator variables of interest, estimates indicated that failing Hispanic students gained less than failing non-Hispanic students.

There was mixed evidence on the differential effect of ISAAP on passing students in nonfailing subgroups enrolled in failing schools. These estimates were either statistically different from zero with a positive sign on the estimated coefficient, or not statistically significant. Despite mixed evidence, it is worth noting that these estimates refute the supposition that Idaho's minimum competency accountability program compromises educational needs and opportunities of high-performing, academically accelerated students. Indeed, these results provide evidence that there was slack in the operation of traditional public schools in Idaho, and that heightened accountability and threat of sanction spurred operational efficiency.

This dissertation also explored whether other programmatic features of NCLB, such as severity of sanction, overlapping student categories, and students in terminal grades further explained strategic resource-allocation decision making. No evidence was found of the severity of sanctions impacting strategic resource-allocation decision making.[27] However, investigation into the effect of severity of sanctions was limited given, first, that the data panel encompassed only 2 years of potential responses by schools to Idaho's minimum competency accountability program and, second, that NCLB sanctions truly only kick in after a school fails proficiency for 3 consecutive years. This

[27]The least squares mean difference using the Tukey-Kramer adjustment for multiple comparisons between schools that failed to meet ISAAP proficiency standard for no years, 1 year, and 2 consecutive years was used to determine that the mean difference between schools failing to meet ISAAP proficiency standard for 1 and 2 years, respectively, was not statistically different.

dynamic will be revisited when additional data become available and sanctions of increasing severity are imposed upon serially failing schools.

There was no evidence of schools differentially targeting students whose performance influenced proficiency determination of more than one subgroup. Most student subgroups in Idaho for which NCLB requires states to hold schools accountable do not meet the state-defined minimum required *n* of 34 students. Furthermore, even if a school does satisfy the minimum *n* requirement for a marginally sized student subgroup, there is a good chance that particular student subgroup will fall below the minimum *n* the following year due to natural attrition and, hence, be exempted from AYP calculations. As such, homogeneity of Idaho's student population renders the issue of students in overlapping categories largely irrelevant.

Directions for Future Research

Future iterations of this achievement tradeoff research will explore (a) additional measures of the degree to which students are likely to be targeted, (b) whether the task of meeting ISAAP performance targets is feasible, and (c) whether schools substitute resources across outcomes.

Because ISAAP-induced resource-allocation decision making implies that schools target resources to students offering a greater return on expenditure for a given increase in test scores, subsequent techniques for estimating a school's short-run incentive to target resources should take into consideration the probability that a school overlooks a student. Future research will explore the impact of $RAT.GAP_{ijt}$ on student gains. $RAT.GAP_{ijt}$ is the ratio of $ST.GAP_{ijt}$ and the gap between the passing threshold and the

marginal student whose improved performance is needed to meet state-defined proficiency standards under AYP, assuming that students who scored better than the marginal student but still below passing improve their test performance at the expected rate. If a school only needs three more students to pass the threshold to meet ISAAP's proficiency standard in math, $RAT.GAP_{ijt}$ quantifies the ratio between $ST.GAP_{ijt}$ and the gap between the proficiency threshold and the marginal student, as expressed in terms of the gap between the passing score and the test score of the marginal student. Furthermore, $RAT.GAP_{ijt}$ advances $ST.GAP_{ijt}$ by taking into consideration how important a particular student is for a particular school to meet AYP and how far that school is from the state-defined proficiency standard.

Future research will also consider the magnitude of the task of improving failing subgroups in terms of the extent to which a particular subgroup is failing relative both to the size of that subgroup and to the total number of high-stakes students tested in the school. Specifically, a ratio variable will be introduced, and defined by the number of failing students in a particular subgroup that need to pass in order for that subgroup to meet AYP over the number of students tested in that subgroup. A second ratio variable will accompany the first, and be defined by the same number of failing students required to pass in that subgroup over the total number of high-stakes students tested in the school. In a resource-constrained environment, it is hypothesized that the likelihood of resources being allocated to a particular subgroup grows as the number of marginal students in that subgroup increases, but is tempered concomitantly by the size of that marginal subgroup in relation to the number of high-stakes students tested in the school. Fewer resources

available to nonfailing students and/or nonfailing student subgroups would suggest amplification of the low-/high-achievement tradeoff.

Model specifications estimated in this study were restricted to a school failing to meet ISAAP's proficiency standard due to poor math performance. A similar indicator variable could be created for reading and language arts both in and across subgroups for which ISAAP holds schools accountable. If schools failed to meet proficiency due to poor performance in reading are outcomes in mathematics affected? Or, do schools focus more resources on outcomes tied accountability?

Specifically, I intend to examine if Idaho public schools reduce instruction in low-stakes subjects. In 2005, Idaho entered into a flexibility agreement with the US Department of Education that permits districts to alter accountability program features required under NCLB. Specifically, the language arts portion of the ISAT was a high-stakes exam in Idaho for 3 years (i.e., 2002-03, 2003-04, 2004-05 school years). Starting in the 2005-06 school year, however, the language arts test was replaced by state-defined student growth measures. To test for a subject tradeoff one can compare achievement before and after Idaho entered into flexibility agreement.

Conclusion

This dissertation set out to provide empirical evidence on achievement tradeoffs and NCLB. Recognizing that ideological predispositions have dominated burgeoning public and scholarly interest in distributional effects under NCLB's accountability system, these findings provide much-needed evidence as Congress approaches reauthorization of NCLB in 2007. There is strong evidence in Idaho that NCLB's threat

of sanctions increased incentives for schools and school districts to elevate learning opportunities for traditionally low-performing students and student subgroups, but that the increased performance by traditionally low-performing students and student subgroups did not occur at the expense of traditionally high-performing students. Indeed, it appears that Idaho's response to NCLB is one of improved efficiency and not achievement tradeoffs, in that traditional public schools in the state did more with the same level and distribution of resources as in years past.

REFERENCES

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis, 23*(2), 171-191.

Amrein, A. L., & Berliner, D. C. (2002a). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18). Retrieved September 8, 2006, from http://epaa.asu/epaa/v10n18/

Amrein, A. L., & Berliner, D. C. (2002b). *The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP test results in states with high school graduation exams.* Tempe: Educational Policy Research Unit, College of Education, Arizona State University.

Amrein-Beardsley, A. A. & Berliner, D. C. (2003). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Education Policy Analysis Archives, 11*(25). Retrieved September 8, 2006, from http://epaa.asu.edu/epaa/v11n25/

Baker, G. (1992). Incentive contracts and performance measurement. *Journal of Political Economy, 100,* 598-614.

Ballou, D. (2002). Sizing up test scores. *Education Next*. Retrieved September 8, 2006, from http://www.educationnext.org/20022/10.html

Ballou, D., Liu, K., & Rolle, E. (2005). *Response to No Child Left Behind among Tennessee schools.* Unpublished working paper, Peabody College of Vanderbilt University.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231-268.

Booker, K., Gilpatric, S., Gronberg, T., & Jansen, D. (2006). *The effect of charter schools on traditional public school students in Texas: Are children who stay behind left behind?* Unpublished working paper, Texas A&M University.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*(1), 1-18.

Burgess, S., Propper, C., Slater, H., & Wilson, D. (2005). *Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools* (Working Paper No. 05/128). Bristol, England: Centre for Market and

Public Organisation. Retrieved September 8, 2006, from http://www.bris.ac.uk/cmpo/workingpapers/wp128.pdf#search=%22Burgess%2C%20S.%2C%20Propper%2C%20C.%2C%20Slater%2C%20H.%2C%20%26%20Wilson%2C%20D.%20(2005).%20Who%20wins%20and%20who%20loses%20from%20school%20accountability%3F%22

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305-331.

Center for Education Policy. (2005). *From the capital to the classroom: Year 3 of the No Child Left Behind Act.* Washington, DC: Author.

Chakrabarti, R. (2006). *Do public schools facing vouchers behave strategically? Evidence from Florida* (Working Paper Series). Nashville, TN: National Research and Development Center on School Choice.

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (2004a). *A nation deceived: How school's hold back America's brightest students* (The Templeton National Report on Education, Vol. 1). West Conshohocken, PA: Templeton Foundation Press.

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (2004b). A nation deceived: How school's hold back America's brightest students (The Templeton National Report on Education, Vol. 2). West Conshohocken, PA: Templeton Foundation Press.

Cronin, J., Kingsbury, G. G., McCall, M. S., & Bowe, B. (2005). *The impact of the No Child Left Behind Act on student achievement and growth*. Boise, ID: Northwest Evaluation Association.

Cross, R. W., Rebarber, T., & Torres, J. (2004). *Grading the systems: The guide to tests, standards, and accountability policies.* Washington, DC: Thomas B. Fordham Foundation.

Cullen, J., & Reback, R. (2002). *Tinkering towards accolades: School gaming under a performance accountability system.* Unpublished manuscript, University of Michigan.

Davidson, J., & Davidson, B. (2005). *Genius denied: How to stop wasting our brightest young minds*. New York: Simon & Schuster.

Deere, D., & Strayer, W. (2001). *Putting schools to the test: School accountability, incentives, and behavior.* Unpublished manuscript, Texas A&M University.

Diamond, J. B., & Spillane, J. (2004). High stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record*, *106*(6), 1145-1176.

Education Commission of the States. (2004). *State implementation of the No Child Left Behind Act: Respecting diversity among states (ECS Report to the Nation).* Denver, CO: Author.

Figlio, D. N. (2005). *Testing, crime and punishment* (NBER Working Papers No. 11194). Cambridge, MA: National Bureau of Economic Research. Retrieved September 8, 2006, from http://www.nber.org/papers/w11194.pdf

Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability and disability: Gaming the system?* (NBER Working Papers No. 9307). Cambridge, MA: National Bureau of Economic Research. Retrieved September 8, 2006, http://www.nber.org/papers/w9307.pdf

Figlio, D. N., & Winicki, J. (2005). Food for thought? The effects of school accountability plans on school nutrition. *Journal of Public Economics*, 89, 381-394.

Gallagher, J. J. (2004). No Child Left Behind and gifted education. *Roper Review, 26*(3), 121-123.

Goodnough, A. (1999, December 8). Answers allegedly supplied in effort to raise test scores. *New York Times*, n.p.

Greene, D. R., Trimble, C., & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice, 1,* 22-32.

Greene, J. P., & Forster, G. G. (2003). Guest editorial: Burning high-stakes testing at the stake. *Education Gadfly*, *3*(1). Retrieved September 8, 2006, from http://www.edexcellence.net/foundation/gadfly/issue.cfm?id=6&edition=#412

Grissmer, D. W., & Flanigan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas.* Washington, DC: National Education Goals Panel.

Guthrie, J.W. (2006, October 19). *Data systems linking resources to actions and outcomes: One of the nation's most pressing education challenges.* Paper presented at the meeting of the University of Arkansas Department of Education Reform Technical Board of Advisors Conference.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved September 8, 2006, from http://epaa.asu.edu/epaa/v8n41/

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2005). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18*(5), 527-544.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). The revolving door. *Education Next, 4*(1), 77-82.

Hanushek, E. A., Kain, J. F., Rivkin, S. G., & Branch, G. F. (2005). *Charter school quality and parental decision making with school choice* (NBER Working Papers No. 11252). Cambridge, MA: National Bureau of Economic Research. Retrieved September 8, 2006, from http://www.nber.org/papers/w11252

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297-327.

Hauser, C., & Kingsbury, G. (2004). *Differential item functioning and differential test functioning in the Idaho Standards Achievement Tests.* Lake Oswego, OR: Northwest Evaluation Association.

Heubert, J. P., & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation* (Report of the National Research Council). Washington, DC: National Academy Press.

Holmes, G. M. (2003). *On teacher incentives and student achievement*. Unpublished working paper, East Carolina University, Department of Economics.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organizations, 7,* 24-52.

Idaho State Board of Education. (2006). *State of Idaho consolidated state application accountability workbook for state grants under Title IX, Part C, Section 9302 of the Elementary and Secondary Education Act (Public Law 107-110).* Washington, DC: U. S. Department of Education Office of Elementary and Secondary Education.

Jacob, B. (2005). Accountability, incentives, and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5-6), 761-796.

Jacob, B., & Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, *118*(3), 843-877.

Kentucky Department of Education. (2001). *Standard setting: Synthesis of three procedures and findings.* Frankfurt, KY: Author.

Kingsbury, G. G. (2003, April 24). *A long-term study of the stability of item parameter estimates.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kirst, M. W. (1990). *Accountability: Implications for state and local policymakers*. Washington DC: US Department of Education, Office of Educational Research.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2003). *What do tests scores in Texas tell us?* Santa Monica, CA: RAND.

Koretz, D., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS).* Washington, DC: Educational Resources Information System.

Koretz, D., Barron, S. I., Mitchell, K. J., & Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high stakes conditions* (Technical Report 551). Los Angeles: University of California, Center for the Study of Evaluation.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly, 38*(4), 494-529.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance.* Englewood Cliffs, NJ: Prentice-Hall.

MacPhail-Wilcox, B. (1988). Motivation, satisfaction, and productivity: A psychological perspective on resource allocation in education. In D. H. Monk & J. Underwood (Eds.), *Microlevel school finance: Issues and implications for policy* (pp. 233-269). Cambridge, MA: Ballinger Press.

Mohrman, S. A., & Lawler, E. E. (1996). Motivation for school reform. In S.A. Fuhrman & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 115-143). San Francisco: Jossey-Bass.

National Center for Educational Statistics. (2004). *Digest of education statistics.* Washington, DC: US Department of Education.

Neufeld, S. (2000, October 2). Backlash fermenting against school tests: Groups organize to complain about STAR. *San Jose Mercury News,* n.p. Retrieved September 8 2006, from http://www.fairtest.org/arn/Backlash%20Article.html

Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA achievement level test and measures of academic progress.* Portland, OR: Author.

Northwest Evaluation Association (2006). *Assessment system classroom reports.* Portland, OR: Author.

Peabody, Z., & Markley, M. (2003, June 14). State may lower HISD rating: Almost 3,000 dropouts miscounted, report says. *Houston Chronicle*, p. A1.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature, 37,* 7-63.

Raymond, M. E., & Hanushek, E. A. (2003). Shopping for evidence against school accountability. In W. J. Fowler (Ed.), *Developments in school finance: 2003* (pp. 119-130). Washington, DC: National Center for Education Statistics.

Reback, R. (2006). *Teaching to the rating: School accountability and the distribution of student achievement.* Unpublished manuscript, Barnard College, Columbia University.

Renzulli, J. (2005). Commentary: A quiet crisis is clouding the future of R&D. *Education Week, 24*(38), 32-33, 40.

Resnick, D. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies, Part II* (pp. 173-194). Washington, DC: National Academy Press.

Ruf, D. (2005). *Losing our minds: Gifted children left behind.* Scottsdale, AZ: Great Potential Press.

Sausner, R. (2005, October 21). Gifted education: Deceived, denied and in crisis. *District Administration*. Retrieved September 8, 2006, from http://districtadministration.ccsct.com/page.cfm?p=1230

Simon, R. (2006). *No Child Left Behind: Disaggregating student achievement by subgroups to ensure all students are learning.* Washington, DC: Committee on Education and the Workforce.

Speakman, S., Finn, C. E., Jr., & Hassel, B. C. (2005). *Charter school funding: Inequity's next frontier.* Washington, DC: Thomas B. Fordham Institute.

Stecher, B. M. (2002). Consequences of large-scale, high stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79-100). Santa Monica CA: RAND.

Stecher, B. M., Hamilton, L., & Naftel, S. (2005). *Introduction to first-year findings from the Implementing Standards-Based Accountability (ISBA) Project* (RAND Working Paper No. WR-255-EDU). Santa Monica, CA: RAND. Retrieved September 8, 2006, from http://www.rand.org/pubs/working_papers/2005/RAND_WR255.pdf#search=%22Stecher%2C%20B.%20M.%2C%20Hamilton%2C%20L.%2C%20%26%20Naftel%2C%20S.%20(2005).%20Introduction%20to

%20first-year%20findings%20from%20the%20 Implementing%20Standards-Based%20Accountability%20(ISBA)%20Project.%20RAND%20Working%20Paper.%22

Swanson, C., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis, 24*(1), 1-27.

Thompson, B. (2003). *High stakes testing and student learning: A response to Amrein and Berliner.* Unpublished manuscript, Milwaukee School of Engineering.

Vroom, V. (1964). *Work and motivation*. New York: Wiley Press.

Wanker, W. P., & Christie, K. (2005). State implementation of the No Child Left Behind Act. *Peabody Journal of Education, 80*(2), 57-72.

Winter, G. (2002, December 28). More schools rely on tests, but study raises doubts. *New York Times,* p. A1. Retrieved September 8, 2006, from http://www.ccebos.org/timestestingdoubts.html

Yen, W. M. (1986). Normative growth expectations must be realistic: A response to Phillips and Clarizio. *Education Measurement: Issues and Practice, 7,* 16-30.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics, 85*(1), 9-23.