ABSTRACT

Title: EVALUATING SCHOOLS BASED ON THE PERFORMANCE OF STUDENTS WITH DISABILITIES: A COMPARISON OF STATUS AND VALUE-ADDED APPROACHES

Sandra Dee Embler, Doctor of Philosophy, 2006

Directed By: Professor Philip J. Burke
Department of Special Education

The purpose of this study was to describe, analyze, and compare the results of five commonly used approaches to measuring school-level performance for the subgroup of students with disabilities. Using the reading and mathematics scale scores of students in grades two, four, and six on the Comprehensive Test of Basic Skills (CTBS) and the California Achievement Tests (CAT/5) within a large school district in the mid-Atlantic, five approaches were applied to classify schools as high-performing or low-performing based on the subgroup students with disabilities. The approaches applied included three status approaches (cross-sectional, cross-sectional with confidence interval, and three-year rolling average) and two value-added approaches (unadjusted and adjusted for student demographics). The characteristics of schools classified as high-performing and low-performing based on the performance of students with disabilities using each of the approaches were explored. Each approach was also examined for its reliability, fairness, inclusiveness, and usefulness.

Significant differences in the performance of students with disabilities based on socioeconomic status were observed across all grade levels, but no differences in gain scores were consistently observed. No significant differences in reading and mathematics performance were consistently found across grade levels based on the disability group and LRE of students.

Overall, none of the accountability approaches employed reliably rated schools based on the performance of students with disabilities. Even within the same subject area and using the same approaches, schools labeled as high-performing for students with disabilities one year were labeled as low-performing the following year and vice versa.

Schools classified as high-performing using the cross-sectional and three-year averaging approaches demonstrated some bias against high-poverty schools and schools with large percentages of minority students. Schools classified as high-performing using the cross-sectional with confidence approach disproportionately identified schools with small numbers of students as high-performing. The value-added approaches were least biased in terms of socioeconomic status and the percentage of minority students, but were limited in their inclusiveness. The usefulness of all the approaches was limited by complicated assessment and accountability policies and the use of non-standard accommodations. All analyses were affected by the small number of students with disabilities in the subgroup.

EVALUATING SCHOOLS BASED ON THE PERFORMANCE OF STUDENTS WITH DISABILITES: A COMPARISON OF STATUS AND VALUE-ADDED APPROACHES

By

Sandra Dee Embler

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
2006

Advisory Committee:
Professor Philip J. Burke, Chair
Dr. Robert Croninger
Dr. Paula Maccini
Dr. Kimber Malmgren
Dr. Margaret McLaughlin

DEDICATION

This work is dedicated to my best friend and cheerleader, my husband, Michael. You were always there for me when the tables were uneven, or when I was searching desperately for just the right word. You were understanding when I spent my weekends in front of the computer and constantly reminded me of the great faith and pride you had in me when I needed it most. I couldn't have done it without you!

I would also like to thank my sister, Pat who let me ramble endlessly at times and understood when I just needed to vent. To my parents, Ruby and LeRoy, thank you for always believing that I could accomplish whatever I set my mind to do.

Finally, to all my friends and colleagues who continually asked "Are you finished yet?" In your own way you encouraged me to finally finish.

ACKNOWLEDGEMENTS

I would like to thank Philip Burke, my committee chair and advisor for his support throughout the dissertation process and for always having an open door.

To Maggie McLaughlin, my mentor, I owe my deepest appreciation for her guidance and input throughout my doctoral program. She constantly challenged my interpretations and constantly reminded me of the "so what" of educational research that is essential. The experiences she provided could never be gained through a class or from a book. She truly knows what it means to be a mentor, and without her this dissertation would not have been possible.

My thanks also go to Dr. Robert Croninger, who taught me that the real story exists beyond the numbers and that true meaning isn't always related to statistical significance. I'll carry your lessons with me always.

To my remaining dissertation committee members, Polly Maccini and Kimber Malmgren, thank you for your great ideas and feedback.

TABLE OF CONTENTS

Dedication ii

Acknowledgements iii

Table of Contents iv

List of Tables vi

List of Figures vii

Chapter 1: Statement of the Problem 1
- Accountability Reforms 2
- Theory of Action 4
- Summary 8
  - Purpose of Study 9
  - Research Questions and Hypotheses 10
  - Limitations 12
  - Definition of Terms 13

Chapter 2: Conceptual Framework and Review of Relevant Literature 18
- The Road to Accountability in Education 19
  - Shift from Inputs to Outputs 21
  - Accountability and Students with Disabilities 23
  - Current Accountability Mandates: IDEA and NCLBA 25
- Framework and Theory of Action 28
  - Goals: Content and Performance Standards 29
  - Assessments: Requirements and Uses 32
  - Incentives: Rewards and Sanctions 34
- State Accountability 37
- Accountability Approaches 40
  - Cross-Sectional Approach 41
  - Quasi-Longitudinal Approaches 41
    - Student Mobility 44
    - Heterogeneity 46
  - Longitudinal Approaches 48
    - Adjusted vs. Unadjusted Value-added Approaches 49
    - Value-added Approaches for Accountability 52
- Assumptions and Standards of Accountability Approaches 53
  - Validity and Reliability 54
  - Fairness 57
  - Inclusiveness 58
  - Usefulness 59
- Comparisons of Accountability Approaches 61
- Summary 75

Chapter 3: Data and Methodology 78
Data Sources 78
Dependent Variables 90
Independent Variables 91
Data Analysis and Management 94
Independent *t-tests* 95
ANOVAs 96
Cross-Sectional Approach 96
Cross-Sectional with Confidence Interval Approach 97
Three-year Averaging Approach 99
Value-added Approaches 99
School Classification Matrix 102

Chapter 4: Analyses and Results 103
Performance of Students with Disabilities 104
Socioeconomic Status 104
Disability Group 108
Least Restrictive Environment 114
School Classifications Using Status Approaches 120
Cross-sectional Approach 121
Cross-sectional with Confidence Interval Approach 123
Three-Year Rolling Average Approach 128
School Classifications Using Value-added Approaches 130
Unadjusted Value-added Approach 132
Adjusted Value-added Approach 133
Reliability of School Classifications 137

Chapter 5: Discussion 140
Performance of Students with Disabilities 140
Classification of Schools 147
Fairness 147
Inclusiveness 150
Usefulness 151
Reliability 152
Implications for Policy and Research 155
Minimum *n* 157
Hybrid Accountability Approaches 158
Theory of Action 159
Summary 160

Footnotes 163
Appendices 164
A. Disabilities as Defined in State Code of Regulations 165
B. Description and Location of Cluster Programs 166
C. School Classification Matrices 168
References 178

LIST OF TABLES

Table 1. Approaches Used in Accountability Systems 42
Table 2. Comparisons of Accountability Approaches 63
Table 3. Number of Students Excluded from Sample 81
Table 4. Race of Subsamples b y Year and Subject Area 82
Table 5. Gender and Socioeconomic Status of Subsamples by Year and Subject Area .. 83
Table 6. Disability Classification of Subsamples by Year and Subject Area 84
Table 7 Least Restrictive Environment of Subsamples by Year and Subject Area 85
Table 8. Average, Minimum, and Maximum Number of Students with Disabilities in Elementary, Middle, and Special Schools 86
Table 9. Average, Minimum, and Maximum Percentage of Students with Disabilities by Race and Socioeconomic Status 87
Table 10. Average, Minimum, and Maximum Percentage of Students with Disabilities in Elementary and Middle Schools by Disability Category 88
Table 11. Average, Minimum, and Maximum Percentage of Students with Disabilities in Special Schools by Disability Category 89
Table 12. Mean Standardized Reading Scores and Gain Scores by Socioeconomic Status and Grade 106
Table 13. Mean Standardized Mathematics Scores and Gain Scores by Socioeconomic Status and Grade 107
Table 14. Mean Standardized Reading Scores by Disability 110
Table 15. Mean Standardized Reading Gain Scores by Disability 111
Table 16. Mean Standardized Mathematics Scores by Disability 112
Table 17. Mean Standardized Mathematics Gain Scores by Disability 113
Table 18. Mean Standardized Reading Scores by LRE 116
Table 19. Mean Standardized Reading Gain Scores by LRE 117
Table 20. Mean Standardized Mathematics Scores by LRE 118
Table 21. Mean Standardized Mathematics Gain Scores by LRE 119
Table 22. Characteristics of Schools Classified as High-performing and Low-Performing in Reading Using Status Approaches 124
Table 23. Characteristics of Schools Classified as High-performing and Low-performing in Mathematics Using Status Approaches 125
Table 24 Students Excluded from Value-added Analyses 131
Table 25 Characteristics of Schools Classified as High-performing and Low-performing in Reading Using Value-added Approaches 134
Table 26. Characteristics of Schools Classified as High-performing and Low-performing in Mathematics Using Value-added Approaches 135
Table 27. Schools Most Frequently Classified as High-performing in Reading and Mathematics 139

LIST OF FIGURES

Figure 1. Elements of the No Child Left Behind Accountability Model 31

Figure 2. State Reading and Mathematics Target AMOs 98

Chapter I

Recent educational reforms have called for increased accountability and an emphasis on educational results for all students. One of the central components of these reforms is the administration of large-scale assessments (Doran, 2003; Kifer, 2001; Mazzeo, 2002). The data from assessments are used to provide feedback on the effectiveness of programs and practices and produce evidence upon which policy decisions are made (Kean, 2004). The recently enacted No Child Left Behind Act ("NCLBA," 2001) also requires that assessment results be used to evaluate the quality and effectiveness of schools (Doran, 2003).

The use of assessments for accountability is plagued with many challenges (Goldhabar, 2001; Kifer, 2001; Ladd, 2002). While some of these challenges involve test quality and the pragmatics of assessments, significant questions surround which of the available approaches used to analyze and interpret assessment data is the most reliable and informative (Doran, 2003; Ladd, 2002; Moe, 2003). Several methodological approaches can be used to interpret assessment results, and the approaches can vary in whether they measure the current level, or status, of performance or whether they focus on measuring change. Approaches can also differ in whether they make adjustments for individual and school characteristics (Linn, 2001).

The information and resulting inferences about school quality depend heavily on which of the available approaches and their corresponding methods are used to interpret assessment results (Barton & Coley, 1998; Clotfelter & Ladd, 1996; Hanushek, Kain, & Rivkin, 2002; Heistad & Spicuzza, 2000; Hill, 2001; Ladd, 1999; Linn, 2000). Research has shown that schools identified as high-performing using one approach may be

identified as low-performing using a different approach (Clotfelter & Ladd, 1996; Rubenstein, Stiefel, Schwartz, & Hadj Amor, 2004).

When rating schools the No Child Left Behind Act (NCLBA) requires that assessment results not only be analyzed collectively for the school, but separately for specific subgroups, including students with disabilities. Student characteristics as well as assessment policies can significantly impact interpretations of performance for the subgroup students with disabilities and have complicated the use assessments for accountability (Almond, Lehr, Thurlow, & Quenemoen, 2002; McLaughlin & Thurlow, 2003). Ysseldyke and Nelson (2002) identified more than twenty factors critical to the accurate interpretation of assessment data for students with disabilities. Among these is the small number and heterogeneity of this subgroup, the movement of students in and out of special education, and assessment accommodation policies.

Under the NLCBA judgments about school quality based on assessment results have serious consequences. While schools judged to be high-performing may receive monetary rewards, schools deemed low-performing face increasing sanctions. Therefore, obtaining accurate interpretations of assessment results is critical (Crane, 2002). Misinterpretation of performance trends could result the erroneous labeling of schools, misguided actions, and inaccurate policy decisions (Stevens, Estrada, & Parkes, 2000; Ysseldyke & Bielinski, 2001).

Accountability Reforms

Until recently assessments and accountability were only loosely connected (Carnoy & Loeb, 2004; Mazzeo, 2002). Assessments were used primarily for classroom planning or as measures of placing students in academic tracks. In contrast,

accountability for schools has traditionally been based on changes in governance structure, administrative processes, or the fiscal resources available to schools and districts (Ladd, 1996).

Increased availability of information and reports of the declining performance of American students led to public and political pressure for policymakers to 'do something' about student achievement and prompted calls for greater accountability from those in charge of education. In addressing these pressures, policymakers and reformers shifted their focus from measuring inputs to measuring student outcomes and began using incentives as a means to increase performance (Fuhrman & Elmore, 2004; Mazzeo, 2002; O'Day, 2004).

Parallel to the calls for accountability came increasing pressure to include students with disabilities in general education reforms, including assessment and accountability systems. While the inclusion of this subgroup in broader school and system accountability reforms is viewed as an important step in increasing their academic performance, it conflicts with the traditional accountability for this subgroup, which has historically focused on the legal rights of the student and procedural compliance (McLaughlin & Thurlow, 2003). Outcomes for students with disabilities have typically been measured through the Individualized Education Program (IEP), with results remaining at the school or individual teacher level (McLaughlin & Thurlow, 2003). As a result, policymakers have little experience in interpreting large-scale assessment results and using the results for accountability with this subgroup.

Theory of Action

The theory of action of accountability systems is based on a means-end relationship between assessments and academic achievement (Ladd, 2002; Mazzeo, 2002; O'Day, 2004). This theory of action assumes that when actors are confronted with information about student performance and motivated by incentives (both positive and negative), they will respond appropriately by making changes in policies, practices, and instruction leading to increased student performance (Hanushek & Raymond, 2002; Linn, 2001; O'Day, 2004). The core of this theory of action lies in the relationship between information regarding student performance generated by assessments and improvement (O'Day, 2004). Information is seen as the catalyst that guides actions and leads to academic improvement.

How best to interpret assessment results to obtain the information necessary to drive the theory of action is a contentious issue (Hanushek & Raymond, 2002; Kifer, 2001; Ladd, 2002). The NCLBA mandates that states use a cross-sectional approach that compares the mean proficiency rates of schools and subgroups from year to year and allows no adjustments for individual or school characteristics. Researchers find this approach useful in determining the level at which students are performing and in identifying schools where students are performing below standards (Baker & Linn, 2004; Bracey, 2000a; Doran, 2003; Fuhrman & Elmore, 2004; Linn, 2001; Meyer, 1996). However, this approach is not deemed capable of producing accurate judgments about instructional practices, the effectiveness of programs, or the overall quality of schools (Meyer, 1996; Stevens et al., 2000).

The cross-sectional approach may be especially misleading when used to make statements about schools based on the performance of students with disabilities (US Department of Education, 1994; Ysseldyke & Bielinski, 2002). The reliability and validity of this approach is dependent on the stability from year to year in the characteristics of students (Hanushek & Raymond, 2003), which can be significantly impacted by student mobility (Bryk, 2003; Meyer, 1997; Ysseldyke & Bielinski, 2000). For example, researchers have estimated that 15-20% of students discontinue receiving special education services each year and return to general education, while similar numbers of students enter special education each year (Walker, Singer, Palfrey, Orza, Wenger, & Butler, 1988; Ysseldyke & Bielinski, 2000). Although the NCLBA allows states to average assessment data over two or three years to minimize differences in cohorts and to help stabilize results for small groups (Hill & DePascale, 2003; Kane & Staiger, 2002; Linn, 2004), the Joint Standards on assessments (AERA, APA, & NCME, 1999) warn that a failure to consider changes in the populations being measured may lead to serious misinterpretation of academic performance trends.

In contrast to cross-sectional approaches, researchers typically employ "value-added" approaches. Although there are variations in value-added approaches, most employ a longitudinal design in which the progress of individual students is tracked from year to year to determine increases or decreases in performance (Doran & Izumi, 2004; Meyer, 1996; Stone, 1999). Supporters of value-added approaches assert that when evaluating schools, it is necessary to examine changes in the performance of the same students over time and how the school or program may have contributed to this improvement or 'added academic value.' To make equal comparisons between schools

value-added approaches may also factor in student demographic characteristics, such as race and poverty (Goldhabar, 2001; Linn & Baker, 2002).

Researchers consider value-added approaches a fair and effective approach in analyzing trends in academic performance and evaluating schools (Bryk, 2003; Elmore, 2004; Ladd, 2002; Linn, 2001; O'Day, 2004). Linn and Haug (2002) suggest that value-added measures are in fact the "most direct and valid" way to evaluate student performance and the effects of schools (p.36). Because they focus on measuring change in academic achievement and not the current level of performance, value-added approaches may be fairer in evaluating schools with very low-performing students (Choi, Seltzer, Herman, & Yamachiro, 2004; Education Commission of the States, 2004; Goldschmidt & Choi, 2004; Gong, 2004).

Despite the conceptual appeal and support of value-added measures in the research community, developing and maintaining longitudinal datasets can be expensive and time-consuming. Because they often involve complex statistical analyses, the results of value-added analyses can also be difficult to explain to educators, parents, and the public. If student characteristics are accounted for, there are concerns over which demographic variables should be used. While not controlling for differences in student characteristics may lead to unfair comparisons among schools, adjusting for them may imply that lower expectations are being set for some students and schools (Fuhrman, 2003). The use of value-added measures may be further limited due to attrition. Value-added approaches typically include only students with prior achievement scores, making their use problematic when applied to subgroups that have small numbers or those groups that are highly mobile, such as students with disabilities.[1]

The interpretations obtained from the different approaches can lead to very different conclusions regarding school quality (Clotfelter & Ladd, 1996; Linn, 2004; Rubenstein et al., 2004). Clotfelter and Ladd (1996) individually ranked 571 schools using nine different approaches and found that the correlations between the rankings schools received ranged from a low of zero to a high of .94. Barton and Coley (1998) examined the performance of students on the National Assessment of Educational Progress (NAEP) and concluded that the academic performance of students in the US has either remained level or decreased depending on the approach used to measure progress.

To assist policymakers, educators, and researchers in reconciling differences between the various accountability approaches Baker, Linn, Hermann, and Koretz (2002) developed a set of standards or guidelines to use when comparing approaches. Baker et al. suggest that these standards be used to determine the extent to which different approaches "help, are indifferent to, or undermine" the overall goals of the system (p.70). The standards pertaining to the interpretation of assessment data include reliability, validity, fairness, usefulness, and inclusiveness. Baker et al. recognize that all of these can be theoretically subsumed under the validity of the system, but assert that they are important enough to examine separately.

Baker et al. (2002) acknowledge that different approaches will satisfy these standards to varying degrees and that no single approach is likely to meet all the standards. Trade-offs between standards will most likely be necessary. However, the researchers suggest that evaluating accountability approaches in light of these standards will help clarify the goals and interpretations of accountability system and better guide the process of evaluating the quality of schools.

Summary

The NCLBA requires states to develop accountability systems to measure and report on the academic progress of students and to evaluate schools. The Act further requires that the results of assessments be the primary evidence used to make judgments about how successful or effective each school is in increasing the academic achievement of students.

A well-designed accountability system is a valuable tool in providing the information and signals regarding student achievement that are necessary to improve academic performance. Well-designed accountability systems can strengthen the link between information and improvement. They can further assist educators in evaluating and improving their teaching practices, help school administrators alter policies and practices, lead states and districts to correctly evaluate programs, and shift resources where they are most needed (Linn & Haug, 2002; O'Day, 2004; Stecher, Hamilton, & Gonzalez, 2003).

However, a poorly designed accountability system may do more harm than good (Kane, Staiger, & Geppert, 2002). A poorly designed system has the potential to break the link between the actors and actions and can undermine the success of the accountability model as a whole (Linn, 2001; O'Day, 2004) A poorly designed accountability system can also lead to the false identification of schools and programs, the adoption and continuation of programs that do not really work, or the discontinuation of policies and practices that are truly beneficial. Stecher et al. (2003) report that when choosing approaches and evaluating accountability

systems, policymakers must:

> … be confident that they are reasonably accurate indicators of student achievement and that changes in the results over time reflect real changes in achievement. If not, the accountability system is compromised. For example, if the scores and gains are due to factors other than instruction, ...the wrong incentives will be applied, the risks of inappropriate action will be substantial, and the utility of the scores for decision making will be limited (p.3).

As states struggle to establish accountability systems, it is imperative that they have the information necessary to guide their decision-making. Although policymakers and researchers have recently begun to compare methodological approaches used in accountability systems, the majority of this research has focused on the overall student population. None have examined the consequences of the different approaches for specific subgroups of students. Approaches have also not been examined in light of the standards proposed by Baker et al. (2002).

Given the goal of raising the academic achievement of all students and the stakes for not meeting this goal, it is essential that accountability approaches be examined to determine their strengths and weaknesses and how each approach supports or fails to support the underlying theory of action of accountability reforms. With the required application of accountability approaches to specific subgroups, it is also crucial that their effects on various subgroups also be examined.

Purpose of Study

The purpose of this study was to describe, analyze, and compare the results of five commonly used approaches to measuring school-level performance for the subgroup of

students with disabilities in grades two, four, and six. Using the reading and mathematics scale scores of students on the Comprehensive Test of Basic Skills (CTBS) and the California Achievement Tests (CAT/5) within a large school district in the mid-Atlantic, I applied five approaches to estimate the performance of schools: cross-sectional, cross-sectional with confidence interval, three-year rolling average, unadjusted value-added, and adjusted value-added. The similarities and differences among the results obtained using each approach were explored, and I analyzed how the approaches differed, both in their theoretical underpinnings and in their resulting inferences about schools. Each approach was also examined against the standards identified by Baker et al. (2002).

This study will extend current knowledge regarding students with disabilities and large-scale assessments by: (a) adding to the current literature base on performance trends for students with disabilities, (b) illustrating the effects of approaches on interpretations of performance for students with disabilities, and (c) examining the effect of student characteristics on interpretations of performance measures.

Research Questions and Hypotheses

There are four main questions that guided the current research:

Research Question 1: What is the mathematics and reading performance of students with disabilities over two years? Does the academic performance of students in this subgroup differ by: (a) socioeconomic status, (b) disability group, or (c) least restrictive environment (LRE)?

Research Question 2: What are the characteristics of schools labeled high-performing and low-performing using the following status approaches: (a) cross-

sectional approach; (b) cross-sectional with 95% confidence interval, and, (d) three-year rolling average?

Research Question 3: Are value-added approaches practical for rating schools based on the subgroup of students with disabilities? If so, what are the characteristics of schools labeled high-performing and low-performing based on this subgroup using (a) the value-added approach unadjusted for student demographic characteristics, and (b) the value-added approach adjusted for socioeconomic status (SES) and least restrictive environment (LRE)?

Research Question 4: Across target years what is the reliability of the five approaches in classifying schools based on the performance of students with disabilities (a) within approaches and subjects areas, and (b) across approaches and subject areas?

Because this was an exploratory study, no hypotheses were proposed. While prior research has shown that students of lower socioeconomic status have lower achievement scores than students of higher socioeconomic status, this relationship has not been examined for students with disabilities. It is plausible that academic performance for students with disabilities is more closely related to the disability group of students rather than socioeconomic status.

Due to the small number and mobility of students with disabilities, I expected that the percentage of students scoring proficient would be unstable from year to year using the status approaches, and that schools housing cluster programs for more severely disabled students would be more likely rated low-performing. I also anticipated that schools would be categorized differently based on the approach used. Conceptual underpinnings suggest that because the various approaches measure different aspects of

academic achievement (i.e. status vs. change), they would produce varying lists of high-performing schools. Although there is no prior research examining this specifically for schools based on students with disabilities, I anticipated that the results would be the same as those reported for non-disabled students (Barton & Coley, 1998; Clotfelter & Ladd, 1996; Rubenstein et al., 2004).

Limitations

This study had several limitations. First, analyses were limited to students with disabilities from a single public school district located in the mid-Atlantic. Although this district is large and diverse, the findings may not generalize to other districts, regions, or states. In addition, this study was limited to students in second, fourth, and sixth grades. While the importance of accountability systems across all grades is recognized, they were not addressed here because data were not available. Findings related to secondary schools may have produced different results.

The Title I state assessment for accountability in the state in this study is not the CTBS/CAT5. Although the CTBS and CAT5 are well-established and widely used assessments, they are not officially aligned with the state's curriculum and no cut scores for proficiency have been established by the state. The need to use vertically scaled and consistent longitudinal outcome measures outweighed the option of using the current state accountability assessment.

In addition, state accountability systems are complex and constantly in flux. The federal government continues to issue new guidance and states are responding by continually revising their accountability systems and policies. This study did not attempt to replicate the state's current accountability system under NCLBA or the state

system for measuring yearly progress. While I applied some of the policies currently employed in the state's accountability system, this study does not reflect all current policies.

Finally, there are uncertainties surrounding the quality of the extant data used in this study. All student-level data were reported by the schools and transferred into the district-wide database. In turn, the data used in this study are subsets extracted from the district-wide database. Each of these steps presents room for inconsistency and error and may have affected the overall quality and consistency of the data across schools.

Definition of Terms

Accommodation. An accommodation is an alteration or change in the standard administration of a test. This change can be in the manner in which test items are presented, the manner in which they are responded to, or in the testing environment itself. To be appropriate, accommodations should be listed in the student's Individualized Education Program (IEP) and used during classroom instruction and assessment. [2]

Accountability system. A predefined system designed to collect and analyze information for the specific purpose of holding schools, educators, and others responsible for students' academic performance. Accountability systems typically include the setting of standards, measurement of progress toward standards, and the distribution of rewards and/or sanctions based on measured outcomes.

Adequate Yearly Progress (AYP). Under the No Child Left Behind Act, schools must be able to demonstrate students are making adequate yearly progress in academic achievement and one other indicator (i.e., attendance or graduation rates). AYP is the

minimum level of progress that schools, districts, and states must achieve on these indicators each year.

Cluster programs. Cluster programs are a method of delivering services to students in which specialized programs are established in clusters, as opposed to being offered in every school.

Cohort. A cohort is a group that shares statistical similarities. The identified groups might share a common factor such as the same age or the same grade.

Cohort static. A sample in which group membership is defined by the individual's status in the first year and membership remains constant in following years.

Cohort dynamic. A design in which successive groups are compared to each other, with group membership redefined every year. This year's fourth graders, for example, may be compared to last years' fourth graders.

Cross-sectional. A research approach in which data are collected at one point in time. While data can be collected multiple times and at varying times, they do not require the same group members for each data collection.

Cut score(s). Cut scores are specified points along a continuum of scale scores delineating various performance levels, such as proficient or advanced.

Effect-size. Effect size refers to a family of indices that provide a measure of the magnitude of differences between groups, as opposed to statistical significance. Effect sizes as classified by Cohen are: small - .20, moderate - .50, and large -.80 (Huck, 2000).

Fully unconditional model. The simplest model in Hierarchical Linear Modeling (HLM), with no predictors at any level. The fully unconditional model provides an

estimate of how the variance is distributed across the levels (such as between students and schools).

Intraclass Correlation (ICC). The ICC represents the proportion of variance that can be explained by the level two variable in an HLM model. The ICC is computed as the variance of the level-two variable divided by the total variance. In school effects studies, the ICC is the proportion of variance in the model that is attributable to the school.

Longitudinal approach. A research method that follows and measures growth in the same students over time.

Matrix sampling. A measurement format in which a large set of test items is organized into a number of relatively short item sets, each of which is randomly assigned to a subsample of test takers, thereby avoiding the need to administer all items to all examinees in a program evaluation. Matrix sampling is used to gain more information about student performance in a fixed amount of time over a broader area of content.

Minimum *n*. Under NCLBA, states were required to set the minimum number (n) of students required for reporting and for determining AYP. For determining AYP, minimum *n* represents the smallest number of students in each group necessary to produce statistically reliable results. The minimum *n* is set individually by each state, and at present ranges from five in Maryland to 200 (or 10% of the population) in Texas.

No Child Left Behind Act of 2001 (NCLBA). The No Child Left Behind Act of 2001 was signed into law by President Bush on Jan. 8, 2002, and is the reauthorization of the Elementary and Secondary Education Act originally enacted in 1965. The NCLBA is

the main K-12 education law and implemented a series of accountability measures for public and charter schools in the US.

Nonpersistent factors. Unique occurrences that influence teaching and learning during one year, but may not be present in another (e.g., staff turnover, teacher strike, extremely disruptive cohort of students).

Performance levels. Qualitative descriptors of students' performance determined by cut scores designed to provide evidence of the level at which students have met the content standards. Under NCLBA states were required to distinguish three performance levels: advanced, proficient, and basic.

Reliability. In assessments the reliability is the degree to which test scores are dependable or relatively free from measurement error. In accountability systems, reliability refers to the degree of accuracy and consistency with which schools are rated from year to year (Hill & DePascale, 2003).

Rolling averages. NCLBA allows states to average test scores across three years in order to define AYP. In this manner, a state could use the average student data from 2002, 2001 and 2000 as the gauge of student performance in 2002. In 2003 scores from 2000 would be dropped. As this continues during subsequent years, a three-year rolling average is established.

Student with a disability. A student " with mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), serious emotional disturbance (hereinafter referred to as emotional disturbance'), orthopedic impairments, autism, traumatic brain injury, other health

impairments, or specific learning disabilities; and who, by reason thereof, needs special education and related services " [20 U.S.C §3(A)(i)].

Unadjusted value-added approach. A research approach for measuring student or school performance in which no student or schools demographic variables are used.

Validity. In assessments validity refers to the degree to which the assessment measures what it purports to measure. In accountability systems, validity refers to the degree to which the system permits accurate inferences.

Value-added approach. A methodology that attempts to calculate the contribution a school makes to the education of its students. If students score higher on an assessment than would have been predicted on the basis of their known earlier level of achievement, the difference is determined to be the 'value-added' by the school. Value-added scores can be unadjusted or adjusted, depending on whether they include controls for student or school characteristics.

Vertically equated. Tests that are vertically equated are developed so that the skills measured and the scoring of different levels of the assessment is along a continuum.

Volatility. Volatility is the amount of variation in test scores from one year to the next due to measurement error, sampling error, non-persistent factors, or changes in cohorts. In accountability systems, volatility is often referred to as the reliability or the consistency with which the system classifies schools.

Chapter II

Conceptual Framework and Review of Literature

Current accountability reforms are based on the premise that student performance will increase if those responsible for their performance are held more accountable. By rewarding schools that are performing well and sanctioning those schools that are not performing well, accountability systems are designed to motivate and change the behavior of individuals and to increase the academic achievement of students. Central to the theory of action of these reforms is the role of information regarding student performance, which O'Day (2004) refers to as the lifeblood of current accountability mechanisms.

While the importance of information about student achievement is widely recognized as essential to accountability, how best to interpret and use this information to make judgments about schools is highly controversial (Doran, 2003; Hanushek & Raymond, 2002; Kifer, 2001; Ladd, 2002). Research has shown that different methodological approaches used to measure student performance differ not only in their conclusions, but in the degree to which they satisfy the underlying assumptions and design principles of accountability systems.

This chapter draws on existing literature to examine educational accountability approaches and systems and is divided into four major sections. In the first section the evolution of the accountability movement in education is described, particularly as it applies to the expanding role of assessments and the inclusion of students with disabilities. In the second section the structure and function of current educational accountability systems are discussed, with a focus on the interaction of the components

that form the theory of action. Design principles and assumptions of these systems are also presented, with an emphasis placed on those underlying the interpretation of assessments. The components of accountability systems are also detailed as they are conceptualized in the NCLBA and implemented in the state in this study. In the third section the current approaches used to measure performance, both mandated and proposed, are reviewed. The strengths and weaknesses of each approach are presented, especially as they relate to the underlying assumptions and their use with students with disabilities. In the final section, the potential effects of the varying approaches on interpretations of school performance are described relative to the current literature.

The Road to Accountability in Education

The administration of assessments to measure student progress has been a vital part of American education since the nineteenth century (Mazzeo, 2002; Ravitch, 2002). However, the use of assessments to measure the quality and effectiveness of education, as well as a means of holding schools accountable, is a relatively recent phenomenon (Mazzeo, Ravitch). Understanding the origins of this shift and the forces behind it are central to understanding the increased emphasis on accountability and the issues currently facing policymakers implementing accountability reforms.

Prior to the mid 1960's schools administered assessments primarily for student level decision-making. Assessments were used largely to motivate students or as means of guidance (Carnoy & Loeb, 2004; Mazzeo, 2002; Ravitch, 2002). The results of assessments were used as the basis for entry into high school and college or to place students in academic tracks and programs. Accountability for academic achievement occurred almost exclusively at the student level. The prevailing theory was that academic

performance was the responsibility of the individual student not of the school or educational system (Ravitch, 2002).

Beginning in the late 1950's, several events initiated changes in the political and educational landscape in the United States (Mazzeo, 2002; Ravitch, 2002). The launching of Sputnik in 1957 generated national concern about the overall quality of American education. Specifically, concern grew over the ability of American students to compete academically with students in other countries. The Civil Rights movement several years later drew attention to the inequities in American education, particularly for disadvantaged groups such as minorities and students living in poverty. These events led to a decrease in the desire to hold individual students accountable and an increase in the desire to hold schools accountable for providing a quality education to all students (Mazzeo).

The predominant response by policymakers to this shift was the establishment of fiscal accountability systems and the regulation of inputs and educational processes (Hanushek & Raymond, 2002). In 1965 the federal government, historically absent from education policy, enacted the Elementary and Secondary Education Act (ESEA) as part of the War on Poverty. Title I of this Act provided federal funds to schools with high concentrations of low-income students. In addition to tying federal aid to national policy concerns, the ESEA also initiated the use of assessments for program evaluation. All schools receiving Title I funds under ESEA were required to evaluate the effectiveness of their programs using objective measurements, such as large-scale assessments.

Individual states also began to develop accountability systems. However, the indicators used in these initial accountability systems were measured almost exclusively

at the state or district level and were based on inputs as opposed to outputs (Carnoy & Loeb, 2004; Fuhrman & Elmore, 2004). Inputs frequently measured included the number of books and computers, class size, fiscal resources, certification and salary of teachers, the presence of curricula, and building conditions (Carnoy & Loeb, Mazzeo, 2002; O'Day, 2004). The underlying assumption of these accountability systems was that higher teachers' salaries, better facilities, and ample textbooks were positively related to achievement, and that increases in these resources would remedy whatever ailed the nation's schools (Ravitch, 2002).

Shift from Inputs to Outputs

With the expanding availability of information about educational inputs and student performance, educational reformers began to question the link between resources and student performance (Ravitch, 2002). Despite increased spending on education, scores on the National Assessment of Educational Progress (NAEP) remained stagnant, and scores on international tests consistently placed US students at the median of the international distribution (Greene, 2002; Hanushek & Raymond, 2002). Also, while schools with greater inputs demonstrated higher test scores, this relationship weakened after accounting for school characteristics such as poverty (Carnoy & Loeb, 2004).

The report "A Nation at Risk" (National Commission on Excellence in Education, 1983) drew national attention to the poor academic performance of US students and concluded that the economic strength of the United States was threatened by the poor quality of America's schools. Of five major recommendations the Commission urged for the development of higher standards and a nationwide system of state and local tests designed to evaluate student progress.

Policymakers at both the state and federal level came under increasing pressure to improve student performance and looked to other government and private organizations for systems to accomplish this goal (Fuhrman & Elmore, 2004). Particularly appealing to policymakers were those accountability systems used in the corporate world that centered around setting goals, measuring outputs, the use of rewards, and holding individuals accountable for outcomes and improvement (Elmore, 2004; Ravitch, 2002).

Policymakers also began viewing schools, as opposed to districts and states, as the unit of improvement (Fuhrman & Elmore, 2004; Mazzeo, 2002). The effective schools research of the 1970's and the large-scale studies of the 1980's had shown that some schools performed better than others. This research provided evidence that the school was the primary unit responsible for improving performance. It also showed that schools and what happens in them does indeed matter (Fuhrman & Elmore; Hanushek & Raymond, 2002).

During the 1990's, the federal government continued its expansion into education policy and increased the focus on academic performance, although federal mandates, such as the ESEA still relied heavily on fiscal accountability. Federal mandates were expanded to cover not only individual students receiving Title I services, but schools, local education agencies (LEAs), and states serving these students. In 1994 Congress enacted Goals 2000, a voluntary program which specified national goals including increased high school graduation rates, competency in subject matter, and literacy for all adults. In exchange for funds, states that chose to participate were required to "implement yearly assessments to determine the performance of each LEA and school in achieving these outcomes" (20 U.S.C. § 5801 et. seq.).

In the same year Congress reauthorized the Elementary and Secondary Education Act and renamed it the "Improving America's Schools Act" ("IASA," 1994). The IASA adopted many of the principles in Goals 2000, such as the development of standards and the use of assessments to measure progress. The IASA stipulated that each state receiving Title I funds must develop challenging content standards in reading and math. The IASA also required that states adopt yearly assessments to determine the "performance and progress of each LEA and school in enabling all children served under this law in meeting the state's performance standards" [34 C.F.R § 1111(b)3)].

Unlike previous mandates that allowed states to use locally determined assessments with no central accountability system for results, the IASA required states to implement state-wide assessments and to use incentives and corrective actions based on the results of these assessments. However, unlike Goals 2000, the IASA was not a voluntary program as the mandates applied to all schools receiving Title I funds.

In addition to extending the federal role in accountability reforms, the IASA included significant changes regarding the participation and performance of students with disabilities. The IASA mandated that all students participate in the state assessments and that the results for all students be publicly reported. In defining "all" the IASA specifically refers to students with disabilities as well as students with limited English proficiency [34 C.F.R. § 111(b)(3)(F)].

Accountability and Students with Disabilities

Despite directives in Goals 2000 and the IASA to include students with disabilities, the majority of these students were still not included in statewide assessment and accountability systems after the laws were enacted (McDonnell, McLaughlin, &

Morison, 1997; Thurlow, 2000). Many states seemed to pick and choose which students to include, while other states obtained waivers to exempt students with disabilities, and still others ignored this subgroup altogether (Taylor & Piche, 2002).

In order to receive funding under the IASA, the US Department of Education required states to develop a consolidated plan detailing how the state would meet the requirements of the statute. To meet the requirements of the IASA, states had to define how and to what extent they would include all students in their assessment and accountability systems, including students with disabilities. A review by the US Department of Education found that 36 out of 41 state plans originally not approved were cited for problems with the inclusion of students with disabilities in their state systems (Thurlow, 2004). In 1995 the Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act (IDEA) reported that the percentage of students with disabilities participating in state assessments ranged from 10% to 90%, with a national estimate placed below 50% (US Department of Education, 2001).

Even national studies commissioned by the federal government and designed to report on the performance of students failed to fully incorporate students with disabilities. Prior to 1996 the National Assessment of Educational Progress (NAEP) excluded any student who was mainstreamed less than 50 percent of the time in academic subjects or was judged incapable of participating meaningfully in the assessment (Olson & Goldstein, 1997). Mazzeo et al. (2000) reported that half of all students with disabilities originally sampled were excluded from the NAEP in 1992 and 1994.

The exclusion of students with disabilities from assessments can be attributed in part to the individualized focus of education for these students. When Congress passed

the Education for All Handicapped Children Act in 1975 ("EAHCA," 1975), a major provision in the law dictated the development of an Individualized Education Plan (IEP) for each student identified as having a disability. The IEP is the primary legal document that outlines the services a student is entitled to receive. The IEP also lists the goals the student is to master and spells out how progress in reaching these goals will be measured.

Historically educators have believed that the best way to measure the progress of students receiving special education services was through the IEP, by measuring the level and number of IEP goals and objectives mastered (Thurlow, 2000). Because IEPs are designed around the individual needs of each student, using the IEP as the measure of performance was considered preferable to the use of large-scale assessments which are based on broad skill areas. Therefore, the IEP has served as the foremost method of evaluating the progress of students with disabilities and the primary method of accountability for this subgroup (McDonnell et al., 1997).

While the IEP may provide practical information on the achievement of each individual student, it does not allow for aggregation of results or standardization in reporting the progress of students (McLaughlin & Thurlow, 2003). Thus, no data on group performance are generated by the IEP. The IEP also does not provide aggregate data on the performance of students with disabilities at the state, local, or even school level. Therefore, as a method for system and school accountability the IEP has many shortcomings (McDonnell et al., 1997; Thurlow, 2000).

Current Accountability Mandates: IDEA and NCLBA

During the reauthorization of the Individuals with Disabilities Education Act ("IDEA," 1997), increasing the participation of students with disabilities in assessments

and accountability systems was a priority for both policymakers and advocates (Pizzuro, 2001). Policymakers were concerned by increased spending for special education and supported more accountability for students with disabilities as a means of increasing efficiency. Advocates believed that "he who gets tested gets taught" and supported increased inclusion of students with disabilities in assessments as a means of achieving equality between students receiving special education services and their peers in general education. Both policymakers and advocates agreed that the inclusion of students with disabilities in state assessments was necessary if these students were to reap the benefits of educational accountability reforms (Pizzuro, 2001; Thurlow, 2004).

To help ensure the inclusion of students with disabilities in accountability reforms already in place in general education, Congress included new provisions in the reauthorization of IDEA. Like the IASA, the 1997 IDEA amendments required states to include students with disabilities in state and local assessments and to report separately the scores of these students with the "same frequency and detail used to report the performance for non-disabled students" (34 C.F.R § 300.138). In an attempt to close loopholes in previous legislation regarding participation, the IDEA specified that all students identified with a disability, regardless of severity, must participate in state assessments (34 C.F.R § 300 et seq). For students who are not able to participate in the regular assessments, even with appropriate accommodations, states were required to develop alternate assessments. The IDEA required states to report the number of students participating in the regular as well as alternate state assessments. Unlike Title I mandates that applied only to schools receiving Title I funds, the provisions in the 1997 IDEA applied to all states serving students with disabilities.

While the IDEA increased the requirements for participation of students with disabilities in assessments and reporting, it did not specifically mandate their inclusion in formal state accountability systems. The IDEA did not specify how many or which students should take the alternate assessments (Thurlow, 2004). The IDEA continued to refer to the IEP as the method of accountability and did not require "any agency, teacher, or other person be held accountable if a child does not achieve the growth projected in the annual goals and benchmarks or objectives" [34 C.F.R § 300.350 (b)]. There were also no incentives for states to focus on the outcomes for students with disabilities, as the IDEA included no provisions dictating rewards or sanctions for the disaggregated performance of this subgroup.

With the reauthorizations of the Elementary and Secondary Education Act in 2001 ("NCLBA") and the Individuals with Disabilities Education Improvement Act in 2004 ("IDEA "), the full inclusion of students with disabilities in accountability measures became formalized. In crafting the NCLBA, Congress dictated that accountability policies apply equally to specific subgroups, including economically disadvantaged students, students from major racial groups, students with limited English proficiency, and students with disabilities. The IDEA mandated that all students be included in all state and district assessments and that the performance goals for students with disabilities be the same as those for all students(U.S.C. 20 § 612(a)(16)(A). From the development of standards to establishing requirements for participation and yearly progress objectives, the NCLBA and the IDEA placed students with disabilities equally in the framework of accountability systems.

Framework and Theory of Action

Although there are variations in their specific elements, current accountability systems are centered around three primary components: goals (or standards), assessments, and incentives (Hanushek & Raymond, 2002; Linn, 2001; Stecher et al., 2003; West & Peterson, 2003). Because of their heavy reliance on academic performance and standards, these systems are often referred to as standards-based or performance-based systems (Mazzeo, 2002; Stecher et al., 2003).[3]

The components of accountability systems and their relationship as conceptualized within the NCLBA are illustrated in Figure 1 (Stecher et al., 2003). Within this model the horizontal bands indicate the responsible agency, while the arrows between components indicate the "flow of information, responsibility, or consequences" (Stecher et al., p.3). It is this interaction between components that forms the basis for the theory of action of accountability systems. As evidenced by the placement and direction of the arrows, the components are linked so that no single part acts alone. Rather, the components form a continuous feedback loop, with information or actions in one area affecting actions in adjacent areas. Stecher et al. provide a summary of how these components, in theory, work together and how they synthesize the theory of action:

> The goals of the system are embodied in a set of content or performance standards that schools and teachers use to guide curriculum and instruction. Tests are developed to measure student learning and determine if students have mastered the standards. Improved performance on the tests leads to rewards that reinforce effective behavior; poor performance on the tests leads to sanctions and improvement efforts that modify ineffective behavior (p.3).

While Figure 1 illustrates the generic accountability components that are codified in the NCLBA, the elements are common to standards based reforms and the majority are not unique to NCLBA. What sets this model apart from previous accountability systems is the prescriptiveness of the components and the requirements that they be applied separately to specific subgroups.

Goals: Content and Performance Standards

In order to improve student achievement it is necessary to first define the goals, or what it is that students are to know, and the level at which they must demonstrate that knowledge (Cowan, 2003; Kifer, 2001; Stecher et al., 2003). Within accountability systems these goals are accomplished through the development of academic content and performance standards.

Academic content standards reflect the required content for each subject and delineate "what all students are expected to know and be able to do" (Cowan, 2003, p. 6). At the classroom level, content standards are essential because they guide instruction by clarifying what is to be taught at each level and in each subject (Cowan, 2003; Stecher et al., 2003). At the district and state levels, content standards can influence the selection of textbooks and materials, as well as the development of curricula. The NCLBA does not mandate the specific topics content standards should emphasize or the level at which topics should be covered. However, the Act does specify that content standards be "challenging and encourage the teaching of advanced skills" [34 C.F.R §200.1(b)(ii)]. Furthermore, the NCLBA mandates that content standards be the same for all students and all schools, which ensures that some students are not held to lower expectations due to race, socioeconomic status, or ability.

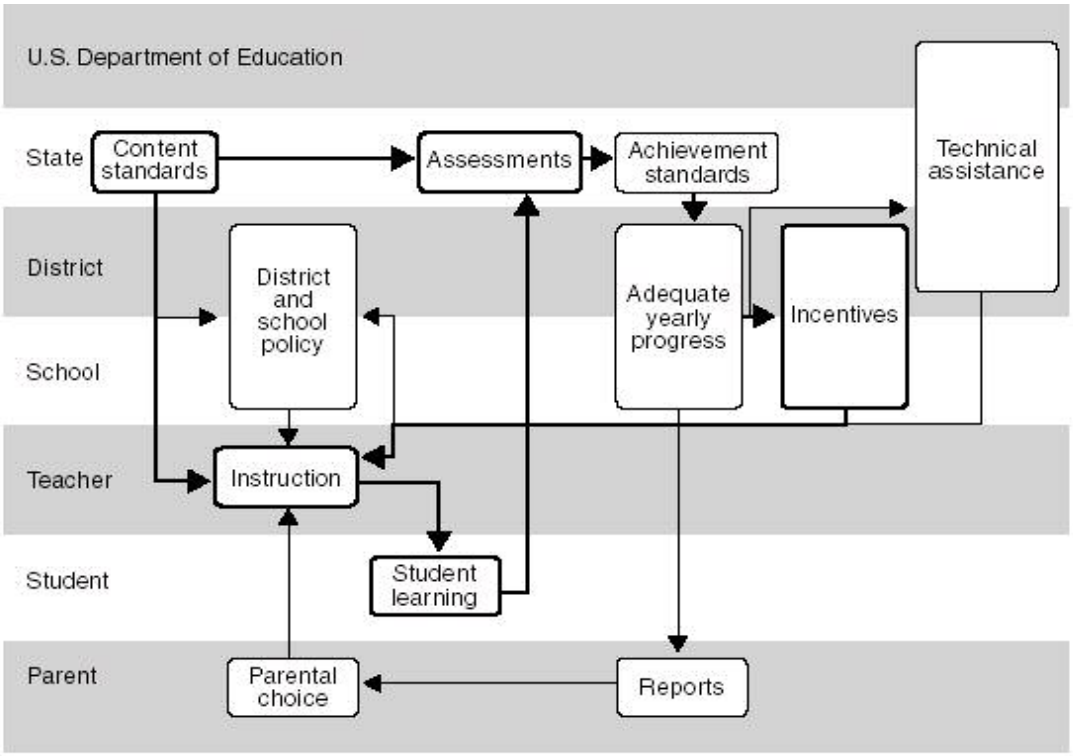
While content standards define what students are expected to know, achievement standards describe the level at which students have mastered the required content. Achievement, or proficiency, standards translate assessment scores into qualitative descriptors, such as proficient and nonproficient, that are meaningful to policymakers, educators, parents, and students (Cowan, 2003; Stecher et al., 2003).

In translating academic performance into achievement standards, a student's numerical score is placed into one of several set categories or achievement levels. Under the NCLBA, states were required to establish cut scores which delineate three achievement levels: basic, proficient, and advanced.[4] The proficient and advanced levels are meant to include those students who have mastered the content standards, while the basic category is meant to include those students who have not yet mastered the standards for the subject and grade being assessed.

As with content standards, the NCLBA requires states to develop separate achievement standards for each subject and grade level. While the Act originally mandated that achievement standards be the same for all students and schools, the US Department of Education has made two exceptions to this provision for students with disabilities. In December 2003, the Department announced that states could develop alternate achievement standards for students with the most significant cognitive disabilities. In May 2005, Secretary Spellings announced that states could also develop "modified achievement standards" for students who can make progress, but who may not master grade-level achievement standards in the same time frame as other students. States are not limited in the number or percentage of students that can be judged against alternate and modified achievement standards. However, only the scores

Figure 1:

Elements of the No Child Left Behind Accountability Model



SOURCE: Stecher et al., "Working Smarter to Leave No Child Behind," Santa Monica, CA: RAND Corporation, 2003

of 1% of all students in the grades tested can be considered proficient using alternate achievement standards, while 2% can be considered proficient using modified achievement standards (US Department of Education, 2005a). Any students exceeding these percentages must be counted within the basic category for purposes of determining AYP.

Assessments: Requirements and Uses

Determining if students have mastered content standards requires measurement (Hanushek & Raymond, 2002). As illustrated in Figure 1, measurement of student learning in relation to content standards occurs through the administration of assessments. Within accountability systems assessments are used to determine the progress of the school, district, and state in helping all students master the content standards (Hanushek & Raymond, 2002; Linn, 2001). Although the NCLBA requires states to use one other measure in addition to assessments, schools are to be judged primarily on the results of state assessments. Accountability determinations and subsequent judgments about schools are built around progress on assessments, defined in NCLBA as adequate yearly progress (AYP).

The NCLBA requires states to administer assessments annually in reading/language arts and mathematics to students in grades 3-8, and at least once to students in grades 10-12. Beginning in 2005-06 states must also assess students in science. To fulfill the requirements of the NCLBA, states were allowed to develop or choose their own assessments; however, the assessments must be aligned with the content standards. That is, the knowledge and skills measured on state assessments must be the same as those specified in the content standards. In addition, the state assessments must

be valid and reliable for the purposes they are being used and provide "coherent information" regarding student achievement [34 C.F.R § 1111(b)(3)(C)(ii)]. States were also required to determine the minimum number ("minimum *n*") of students in each subgroup necessary to produce statistically reliable and valid results. States were required to provide the US Department of Education with evidence from the test publisher, or other reliable sources, that the assessments the state has chosen are of adequate technical quality for each purpose under the NCLBA. State assessments must also be accessible to a wide range of students and incorporate universal design to the extent possible (PL 108-446 § 612(a)(b)(A).

In crafting the NCLBA, Congress set an end goal of having 100% of all students proficient by the 2013-2014 school year. In order to achieve this goal, states were to establish annual measurable objectives (AMOs). The AMOs are the minimum percentage of students required to meet or exceed the advanced and proficient levels on the academic assessments in each school. As an alternate route to making AYP, states can employ the safe harbor provision. Safe harbor allows a school to make AYP if the percentage of students not reaching proficient decreases by at least ten percent and the school meets the other performance indicators of graduation or attendance rates.

To prevent states from excluding large numbers of students from assessments, the NCLBA requires that 95% of students in each subgroup participate in the state assessments in order for the school, district, or state to make AYP. While the original law required the participation rate to be calculated on an annual basis, an amendment was made in 2004 allowing states to average participation rates over a two or three-year

period. The amendment also specifies that students who do not participate due to a significant medical emergency do not count against the school's participation rate.

Within accountability systems it is the results of assessments that drive the theory of action. Assessments serve as the single measure of student learning and the primary indicator of the effectiveness of instruction and district and school policies. The results of assessments provide the basis for judgments about schools and determine whether schools are subsequently rewarded or sanctioned. It is the results of assessments that provide policymakers with the information they need in order to increase student achievement.

Incentives: Rewards and Sanctions

Incentives in the form or rewards or sanctions are the third component of accountability systems and serve multiple functions (Herman, 2004; Stecher et al., 2003). As illustrated in Figure 1, incentives are a vital component of accountability systems. Incentives are meant to affect the behavior of teachers and administrators by providing a "strong signal to teachers and schools about what they should be teaching and what students should be learning" (Herman, 2004, p.142). Incentives are also meant to alter the behavior of policymakers by leading to changes in leadership and policies. Within accountability systems incentives are evidenced through public reporting, public recognition, the distribution of monetary rewards, or through sanctions such as requiring public school choice and threats of restructuring.

In distributing rewards, the NCLBA provides some flexibility to states. The NCLBA only requires states to specify how they will include rewards, such as bonuses and recognition, in holding schools accountable for student achievement. States vary in

the rewards they offer schools. While some states have opted to provide financial bonuses to schools that meet annual goals, others provide only public recognition of such schools.

Although the NCLBA allows states flexibility in distributing rewards, the Act requires a mandatory sequence of increasingly stiff sanctions for schools not making AYP for each sugroup.[5] States were required to determine whether sanctions applied to all schools, or only Title I schools. Schools that do not meet their AMOs for any subgroup for two consecutive years must be identified for "school improvement." Within three months of being identified for school improvement, schools must devise a school improvement plan (SIP) that outlines not only the strategies to improve academic performance, but also addresses the reasons for the school's failure (34 C.F.R § 200.11 et seq). Specifically, the SIP must incorporate strategies based on scientifically-based research that address the academic areas that the school was identified for school improvement. For example, a school that failed to meet annual goals in reading must develop a SIP that lists specific strategies the school will use to increase reading. The plan must also incorporate a teacher mentoring and professional development program.

Beginning the following school year, the state must provide parents of students in schools identified for improvement with written notice regarding the schools' identification. The school must also offer students the option to transfer to a school not identified for improvement. Where there is more than one school in the district not identified for improvement, students must be offered the choice between at least two schools. For students with disabilities, the Act does not require that the choice schools be

the same as for students without disabilities, only that choice schools for students with disabilities be able to offer a free and appropriate education (FAPE).

If a school fails to meet annual goals for three years, the school must continue offering choice to students. However, eligible students who do not take advantage of school choice must be provided with supplemental services. Supplemental services must be provided in addition to instruction provided during school hours and must be designed to increase the academic achievement of students on the state's assessment.

After four consecutive years of failing to make AYP for the school overall or for any one subgroup, districts and states can take "corrective action." Schools identified for corrective action must continue to implement school improvement plans, public school choice, and supplemental services. In addition, the district or state may institute a "restructuring plan" for the school, defined as a "major reorganization of a school's governance arrangement" (34 C.F.R § 200.43(a). The NCLBA provides a list of eight possible options for states under corrective actions. These include: replacing teaching and administrative staff, implementing a new curriculum, decreasing management authority, or restructuring the internal organization of the school. While the LEA may institute additional actions, it must choose at least one of the eight options specified in NCLBA.

After five consecutive years of failing to make AYP, LEAs must prepare a plan to impose alternative governance. Options for alternative governance include reopening the school as a charter school, replacing all or most of the school staff, and turning the operation of the school over to the state. Within the theory of action of accountability systems, it is the application of rewards and sanctions that drives the actions of the major

actors. The desire for rewards and the fear of sanctions lead the actors to actions that will increase the performance of students.

State Accountability

The No Child Left Behind Act, although mandating many aspects of school accountability, does provide some flexibility in implementation for states. For example, states are free to develop their own content and achievement standards and to choose their own assessments, although these components must go through a peer review approval process before implementation. For students with disabilities, states are also free to develop alternate assessments and make decisions regarding the application and inclusion of accommodations.

The state in this study developed their state accountability system in 1988 and created state standards in 1990. Although the standards were used by schools since their inception, these standards were not formally adopted by the State Board until 1999. To measure progress toward the state standards, the state developed state assessments and a state accountability system. The state assessments were criterion referenced performance tasks linked to content standards in the areas of reading, math, writing, social studies, and science. The state assessments were administered to students in grades 3, 5, and 8.

The state assessments were designed for school accountability purposes and not to measure individual student achievement. The assessments employed a matrix sampling procedure and individual students completed only one-third of each assessment (Kifer, 2001). Individual student assessments were then combined by grade and subject to calculate an overall school rating. However, states were still required under the 1994 ESEA to administer assessments that were nationally norm-referenced in order to allow

comparisons with other assessments. To satisfy this requirement, the state administered the California Test of Basic Skills (CTBS) and California Achievement Test (CAT/5) to students in grades two, four, and six.

With the passage of NCLBA and its requirement that assessments produce individual student scores, the state developed and began administering a new state assessment in the spring of 2003 to students in grades three, five, and eight in the areas of mathematics and reading /language arts. Several districts in the state continue administering the CTBS to students in second grade as an indicator of how these students are achieving in the areas of reading, math, and language.

For students with severe cognitive disabilities who are not able to participate in the regular state assessment, the state administers the state's alternate assessment, which uses a portfolio measure. The state is currently in the process of developing their state assessment based on modified achievement standards.

The current state assessment is combination of criterion and norm referenced components and contains multiple choice and short essay questions. Individual students' scores on the state assessments are computed as scale scores. Using cut scores developed by the state, the scale scores are then translated into three performance levels for reporting and calculating yearly progress: basic, proficient, and advanced. Neither the reading/language arts nor mathematics tests of the state assessment are vertically equated. This means that the scores between the reading and mathematics assessment can not be equally compared, and that gain scores on each assessment from year to year can not be computed.

To fulfill the requirement that students with disabilities be provided with accommodations, the state developed policies defining which accommodations would be counted as standard and those that would be considered non-standard on all assessments given in the state. The state defines standard accommodations as those that do not change the construct of the assessment; and counts these scores equally with all other scores. In contrast, the state defines non-standard accommodations as those that change the construct of the assessment and counts the scores obtained with non-standard accommodations as basic, or nonproficient.

The state developed separate accommodation policies for each of their state assessments and these accommodation policies changed several times between 2000 and 2005. In 1999-2000, calculators on the math assessment and extended time on the reading assessments were considered non-standard accommodations. However, in 2001 the state changed calculators to standard and extended time to non-standard accommodations. In 2003 all accommodations except the read aloud for students below 4th grade were reclassified as standard.

To assure that at least 95% of students participate in the state assessments, the state assigns all students enrolled on the date of testing a score. Students who do not physically participate (such as those absent or those not completing an assessment) are automatically assigned the score of basic. To assure the statistical reliability and validity of calculations required by the NCLBA, the state requires that all subgroups have a minimum of five individuals and uses a 95% confidence interval for purposes of calculating AYP.[6]

Accountability Approaches

Using assessment scores for accountability requires not only that the tests be valid, but the interpretations regarding school quality generated by the assessments must also be valid. In order to serve as the catalyst for actions within accountability systems, the data obtained from assessments must be interpreted (O'Day, 2004). It is the interpretation of assessment results that answers questions regarding the level at which students are performing, how the academic performance of students differs, and if the academic performance of students is increasing or decreasing. The interpretation of aggregated assessment results is also used to provide information regarding the quality and effectiveness of schools.

There are several approaches that can be used within accountability systems to draw conclusions on student performance. The approaches can differ significantly in the unit of analysis employed, the data used, and in the students included in the analyses. The approaches can also differ in their underlying definition of a high-performing or effective school (Raudenbush, 2004). While some approaches define the quality and effectiveness of a school by the average achievement of students, other approaches define quality by the level of change in students' achievement.

Researchers use various terms to refer to the different approaches used to interpret assessment results (Table 1). However, each approach can be classified into one of three broad categories: (1) cross-sectional approaches, (2) quasi-longitudinal approaches, and (3) longitudinal approaches (Carlson, 2000; Hamilton & Koretz, 2002; Linn, 2001). Within each of these broad categories are specific approaches with diverse methodological designs.

Cross-Sectional Approach

Using the cross-sectional approach, student performance is measured at a single point in time. Accountability systems that employ the cross-sectional approach typically measure the average performance of students or the percentage of students scoring above or below a set proficiency level. The cross-sectional approach defines school quality by the average score or percentage of students scoring proficient at one point in time. Schools that meet or exceed certain levels are defined as high performing or effective schools, while those not meeting the set levels are determined to be low-performing.

Because the cross-sectional approach provides only a single "snapshot" of the performance of students, it cannot provide information on changes in achievement across years. Prior to the NCLBA the cross-sectional approach was the most common approach used by states to evaluate the quality of schools. However, with the passage of the NCLBA states were forced to abandon the cross-sectional approach for approaches capable of providing information on achievement over time.

Quasi-longitudinal Approaches

The second group of approaches used in accountability systems are referred to as quasi-longitudinal approaches (Carlson, 2000; Hamilton & Koretz, 2002; Linn & Baker, 2002). In quasi-longitudinal approaches the achievement of all tested students in a single year are compared to the achievement of all tested students in preceding years (Linn, 2001). In quasi-longitudinal approaches comparison groups can be partially independent (e.g. third graders last year compared to fourth graders this year) or totally independent (e.g. third graders this year compared to third graders in previous years). Although quasi-longitudinal approaches measure and compare achievement over time, they are not

Table 1

Approaches Used in Accountability Systems

Category	Definition	Approaches
Cross-sectional	Measures performance at a single point in time. Example: The percentage of third graders scoring proficient in the current year.	point in time measure cross-sectional with confidence interval
Quasi-longitudinal	Measures the performance of the same "group" over time. Group membership is not static; members are not required to be the same from year to year. Example: A comparison of all third graders last year and all fourth graders this year.	average score trend cohort dynamic cross-sectional trend repeated cross-sectional successive cohorts three-year averaging
Longitudinal	Measures the same group over time, but group membership is static, the same students comprise the comparison groups. Example: Third graders the previous year and the same students as fourth graders in the current year.	cohort static cohort study value-added

considered a true longitudinal approach because comparison groups do not contain the same students (Figlio, 2002; Linn, 2000). Quasi-longitudinal approaches are mandated in the NCLBA, schools are required to measure the percentage of all tested students scoring proficient each year and compare these percentages across years. As an alternative, schools may average assessment data over three years and use these averages for comparisons over time.

The major advantage of both cross-sectional and quasi-longitudinal approaches is that they are easy to calculate and are easily understood by policymakers and the general public (Goldhabar, 2001; Heistad & Spicuzza, 2000; Raudenbush, 2004). In addition, because they only require the calculation of average achievement scores or the percentages of students scoring at a set performance level, extensive collection of additional data and sophisticated statistical analyses are not utilized. Cross-sectional and quasi-longitudinal approaches are also advantageous in that the scores of all tested students can be included.

However, the validity of the interpretations from cross-sectional and quasi-longitudinal approaches is dependent on the stability of the comparison groups from year to year (Bracey, 2000a; Ladd, 2002; Linn, 2001; Ysseldyke & Bielinski, 2002). For example, this year's third graders are assumed similar in the aggregate to last year's third graders. To fulfill this assumption, consistent comparison groups are necessary from year to year (Bracey, 2000a; Ysseldyke & Bielinski, 2002). If the composition of the groups changes significantly, "…then one is really comparing apples to oranges" making the resulting inferences from these analyses severely flawed (Ysseldyke & Bielinski, 2002, p. 2). For students with disabilities, the characteristics of groups can change significantly

across grade levels and years, due in part to the mobility of these students in and out of special education and the extreme heterogeneity of this population (Hanushek et al., 2002; Linn & Haug, 2002; Walker et al., 1988; Ysseldyke & Bielinski, 2000).

Student Mobility. Several researchers have documented the mobility of students in and out of special education (Carlson & Parshall, 1996; Hanushek et al., 2002; Walker et al., 1988; Ysseldyke & Bielinski, 2002). Walker et al. (1988) examined the mobility of a sample of 1,184 students receiving special education services in three large urban cities: Milwaukee, Wisconsin; Charlotte-Mecklenburg, North Carolina; and Rochester, New York. The authors selected a stratified random sample of students with disabilities in 1982 and followed these students for two years to determine the mobility and changes in disability classification that occurred during this period. Over the two-year period, 71% of the total sample remained in special education with no change in their disability category and 17% were terminated from special education services. The remaining 12% stayed in special education but were classified under a different disability category.

At the national or state level, there is limited research on the mobility of students in and out of special education. The federal government under IDEA requires and collects such data only on students over the age of 14, and only three studies examined the issue at the state level. Carlson and Parshall (1996) analyzed data from over 51,000 students in Michigan, while Hanushek et al (1998) and Ysseldyke and Bielinski (2002) analyzed data on approximately 200,000 students in Texas.

In all three studies the percentage of students in the states receiving special education services ranged from 11%-14%. Of those students receiving special education services in Michigan, Carlson and Parshall (1996) reported that 7% of students across

grades 1-12 discontinued receiving services and returned to general education annually. Transition rates were highest for students in fourth through sixth grade. Ysseldyke and Bielinski (2002) and Hanushek et al. (1998) found that approximately 10%-13% of students in special education in fourth grade in Texas were not in special education the following year. Similarly, 17% of those receiving special education services in 5th grade had not received special education services the previous year. And overall, approximately 20% of the population in special education in Texas changed from year to year (Ysseldyke & Bielinski, 2000).

Ysseldyke and Bielinski (2002) also found that in Texas the turnover rate decreased and the population became more stable as students progressed in grade. By seventh grade only 10% of students in special education in Texas exited and only 8.0% of students had not previously received special education services. Researchers in all three studies found that the students most likely to exit special education in the respective states were identified as speech and language impaired, while those entering special education were disproportionately identified as learning disabled or emotionally disturbed (Carlson & Parshall, 1996; Walker et al., 1988; Ysseldyke & Bielinski, 2002).

Students also transitioned between general and special education more than once. Carlson and Parshall (1996) reported that after three years, 4% of those who had exited special education in Michigan had returned. Ysseldyke and Bielinski (2002) found that over a five year period, 16% of students in Texas who had exited special education by the end of 4th grade returned after the 5th grade. Over the five-year period, 16% of those students who had moved in and out of special education had done so at least twice.

While mobility occurs with all populations, the movement of students between general and special education can be especially troublesome because it is directly related to academic achievement (Hanushek, Kain, & Rivkin, 1998; Ysseldyke & Bielinski, 2002). According to the IDEA, placement in special education is based on the presence of a disability that "adversely affects academic performance" (34 C.F.R § 300.7). Therefore, by definition students are placed in special education when they are diagnosed with a disability and when they demonstrate lowered academic performance. Likewise, students exit special education when they no longer meet these criteria.

In examining the relationship between mobility and academic achievement, Ysseldyke and Bielinski (2002) found that the achievement scores of students in Texas who exited special education were roughly .50 standard deviations higher than those who remained in special education. In contrast, the achievement scores of students who entered special education averaged .75 standard deviations below those who exited special education. This movement of students in and out of special education results in unstable groups from year to year. Moreover, the relationship of this movement to academic achievement leads to a subgroup of increasingly lower-achieving students.

Heterogeneity. The requirement for consistent cohort groups is also likely to be violated due to the extreme heterogeneity of the subgroup of students with disabilities. The IDEA specifies thirteen categories of disabilities (34 C.F.R § 300.7) ranging from students with mild speech impairments to those with severe physical and cognitive disabilities. While there is variation in the cognitive and academic functioning of students from year to year within all groups, there can be extreme variance in the characteristics of students with disabilities. For example, a school's population of

students with disabilities one year could be comprised mainly of students with mild speech and language disabilities, while the following year the majority of students could have moderate to severe cognitive disabilities. Yet all students diagnosed with a disability, regardless of severity, are included in the subgroup of students with disabilities in accountability systems.

Due to inconsistent cohorts, researchers are particularly critical of using cross-sectional or quasi-longitudinal approaches for accountability (Barton & Coley, 1998; Heistad & Spicuzza, 2000; Meyer, 1997; Stevens et al., 2000). When changes in the student population occur, legitimate changes in assessment performance can not be separated from changes due to differences in cohorts (Raudenbush, 2004; Tindal, 2002; Ysseldyke & Bielinski, 2002). Changes in achievement levels from year to year may reflect changes in the composition of groups rather than changes in instructional effectiveness or programs (Doran & Izumi, 2004; McDonnell et al., 1997; Raudenbush, 2004). Kane, Staiger, and Geppert (2001) find that:

> Even when a school is on the right track, the path to improved student performance is rarely a straight path. Each two steps forward is often followed by one step back. The cause is often not a lack of resolve among school administrators or a waning desire among teachers and students. Rather, it is the natural fluctuation in performance that comes with the passing of successive cohorts of children through a school (p.3).

Researchers caution that cross-sectional and quasi-longitudinal approaches in accountability systems may lead policymakers, researchers, and the general public to severely misinterpret the performance of students and erroneously judge the quality of

schools. Researchers and policymakers alike contend that in order to measure the progress of students and evaluate the quality of schools it is necessary to track the same students over time (Ladd, 2002; Linn, 2001; Meyer, 1997; Raudenbush, 2004).

Longitudinal Approaches

Approaches in the third category of accountability approaches are referred to as longitudinal approaches. Longitudinal approaches are similar to quasi-longitudinal approaches in that they compare student achievement over time. However, as opposed to cross-sectional and quasi-longitudinal approaches that compare different students over time, longitudinal approaches base comparisons on the same students. For example, the scores of students of students in third grade one year would be compared to the scores of the same students as fourth graders the following year.

One of the most common and widely used longitudinal approaches is the value-added approach. The value-added approach is grounded in economics and seeks to defines an organization's effectiveness by increases in productivity or outputs (Greene, 2002; Meyer, 1997). The use of value-added approaches in accountability systems developed out of the realization that individual students enter school with differences in their educational experiences, and that these individual differences in academic experience account for much of the variance between schools (Baker & Linn, 2004; Bryk, 2003; Clotfelter & Ladd, 1996; Fuhrman & Elmore, 2004; Linn, 2001; Raudenbush, 2004). When used within accountability systems, the value-added approach is a means of analyzing and reporting student performance based on improvement ("growth") in assessment scores over two or more points in time (Betebenner, 2004; Crane, 2002; Doran, 2003; McDonnell et al., 1997; Meyer, 1997). The aggregated

increases or decreases in students' performance are used to rate the quality and effectiveness of schools. Schools where students' academic performance has increased are considered high-performing, while those where students' academic achievement has not increased are considered low-performing. Because the value-added approach focuses more directly on student learning, this approach may be especially beneficial in analyzing the performance of students with disabilities (McDonnell et al., 1997).

Although all value-added approaches measure the change in academic performance, they can differ in whether student or school characteristics, such as poverty, are used to adjust or equate performance among groups. Value-added approaches that do not factor in student or school characteristics are considered "unadjusted," while those that factor in the demographics of individual students and schools are referred to as "adjusted."

Adjusted vs.Unadjusted Value-Added Approaches

William Sanders, one of the pioneers of value-added approaches, believes that socioeconomic and demographic characteristics are consistent over a student's schooling and are thus automatically factored into students' test scores (Sanders, Saxton, & Horn, 1997). Sanders theorizes that achievement gains are affected only by the teacher and school, not by characteristics of students themselves. Sanders and his followers believe in the unadjusted value-added approach in which all students and schools are judged equally.

However, in the now famous Coleman Report (1966), the author concluded that schools "bring little to bear on a child's achievement that is independent of his background and general social context" (p. 325). Researchers have demonstrated that

achievement and achievement gains are not determined solely by current school and ability (Barton & Coley, 1998; Heistad & Spicuzza, 2000; Meyer, 1997; Rowan, Correnti, & Miller, 2002; Stevens et al., 2000; Ysseldyke & Bielinski, 2002) and are strongly related to family background and previous educational experiences (Hanushek & Raymond, 2003; Raudenbush, 2004). Based on these facts, unadjusted value-added approaches are criticized because they do not take into account the "non-educational determinants" of learning, such as students socioeconomic status (Koretz, 1996, p.171).

Value-added approaches that are adjusted for student demographics attempt to statistically isolate and control for outside factors that may affect student achievement (Heistad & Spicuzza, 2000; Stone, 1999) in order to establish "reasonable expectations" (McDonnell et al., 1997, p. 273). Supporters of adjusted value added models for accountability assert that when evaluating schools, accounting for differences between individual students and schools provides a more valid and fair assessment of the quality of schools (Fuhrman, 2003; Gaddy, McNulty, & Waters, 2002; Ladd, 2002; Linn, 2004).

Because value-added measures are perceived as more valid indicators of the effects of schools, they are thought to provide clearer signals to educators and policymakers regarding the effects of instruction and programs. In accountability systems, adjusted value-added approaches are viewed as being fairer because they can help distinguish between schools and programs that are producing results because of their efforts and instruction, as opposed to judging schools based on the characteristics of the children they serve.

Yet, despite the advantages of value-added approaches, they have several drawbacks. First, value added measures, especially those that are adjusted, are difficult

to understand. Therefore, they may be less useful than other approaches (Goldhabar, 2001). Critics of value-added approaches argue that accountability approaches must be clearly understood by all major stakeholders in order to be useful, and that value-added measures may be too complex for many parents, teachers, and the general public to understand (Ballou, 2002; Raudenbush, 2004). Walberg (2003) contends that when we employ complicated approaches such as value-added approaches and their variants, we place the accountability system in the hands of statisticians, thus giving up transparency and comprehensibility.

Second, value-added approaches require extensive data collection and can be difficult to calculate (Ladd, 2002). The calculations employed in value-added models require not only personnel who can perform these analyses but also programs that can translate the results into reports. These requirements can result in substantial costs for states and districts (Koretz, 1996; Raudenbush, 2004; Sanela, 2002).

Third, the inclusiveness of value added analyses can be affected by student mobility (Hanushek & Raymond, 2003; Ladd, 2002; Linn, 2001). Because prior achievement scores are required in value-added approaches, this approach has the potential to exclude large numbers of students from analyses and result in comparison groups with small numbers of students. Raudenbush (2004) reports that the reliability of value-added approaches is severely compromised when the number of students in each subgroup is less than five.

Finally, for conceptual, practical, and political reasons, there is significant debate over the proper control variables to include in adjusted value-added analyses (Ladd, 2002). Although researchers have consistently established a relationship between

academic achievement, socioeconomic status, and race, critics argue that adjusting for these factors in essence sets a lower standard for some students (Ladd, 2002). Is it fair, for example, to accept smaller gains for poor and minority students? Should programs that produce smaller gains with these populations be regarded equally successful as those that produce larger gains with more advantaged populations; or should programs be expected to achieve the same results, regardless of the population?

Value-added Approaches for Accountability. The NCLBA does not specifically address the use of value-added measures for accountability. However, certain provisions of the Act appear to preclude its use for rating and classifying schools. The Act specifically states that annual progress is to be measured by increases in the percentage of students scoring proficient and that all students are to be included. Although the NCLBA does allow states to measure annual progress under the safe harbor provision (if the percentage of students who are nonproficient decreases by ten percent), there is no current provision in the law to allow credit for schools based only on progress. Furthermore, the NCLBA requires that achievement goals apply to all schools and subgoups equally. Thus, adjusting for race, poverty, or disability status does not appear to be permitted. The Department of Education denied a recent request by Tennessee to use the value-added approach in their accountability system for students with disabilities and English Language Learners on the grounds that value-added measures do not require proficiency, and that using a separate measure for these subgroups does not meet the requirement that all groups be treated equally.

However, the US Department of Education has demonstrated support for the use of value-added approaches. In a letter to states, former Secretary Rod Paige (2002)

encouraged states to develop systems or in some manner recognize schools that are making improvements. In November 2005, the Department announced that grants would be awarded to ten states to develop and pilot growth models for accountability under the NCLBA. While the chosen states will be allowed to incorporate growth into their accountability systems, they must still assure that all students score proficient in all subjects by 2014.

Assumptions and Standards of Accountability Approaches

On the surface, the theory of action and framework of accountability systems appears relatively simple. There is an appealing logic to the theory that student achievement will increase when educators are judged based on students' performance and when consequences are attached to these judgments (Stecher et al., 2003). However, in reality accountability systems are complex and are composed of numerous interacting principles and assumptions that are necessary for their success (Elmore, 2004; Gong, 2002; Kifer, 2001; Linn, 2004; Walker et al., 1988; West & Peterson, 2003).

To aid policymakers, researchers, and educators in comparing and evaluating accountability systems, researchers from the Consortium for Policy Research in Education (CPRE) and National Center for Research on Evaluation, Standards, and Student Testing (CRESST) identified over twenty standards that should be met if accountability systems are to function as intended within the theory of action (Baker et al., 2002).

Baker et al. (2002) acknowledge that all the standards are necessary for accountability systems to function as intended, and that a failure to meet any of these can cause a failure in the system as a whole (Baker et al., 2002; Hanushek & Raymond, 2002;

Linn, 2001). However, due to the focus of this study, only those standards related to the interpretation of performance measures will be discussed. In addition, while an attempt is made to present each of these assumptions separately, they are closely related and their goals frequently overlap (Kifer, 2001; Ladd, 2002; Linn, 2004).

Validity and Reliability

Traditional definitions of validity are based on the technical properties of assessments, including content-related, construct-related, and criterion-related validity (Brualdi, 1999). However, Messick (1989) argued that not only should assessments themselves be examined for validity, but that the uses and interpretations of assessments should also be assessed for validity. Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13: italics in original). Baker et al. (2002) report that the validity of interpretations is the most important standard for an accountability system.

Current theories of validity have incorporated Messick's framework and define validity not only as a quality of a test per se, but the capacity of users to interpret and make accurate inferences from assessments (Baker & Linn, 2004). The most recent version of the Joint Standards for Assessment reflects Messick's theory and defines validity as the "degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (AERA et al., 1999, p.184). Fuhrman (2003) further summarizes the validity of an accountability system as the extent to which the system focuses on student learning and proposes that

accountability approaches should be valid to the extent that the resulting interpretations provide the information necessary for improvement in student achievement.

A primary threat to the validity of assessments for students with disabilities is the use of accommodations. Accommodations are changes in the standard administration of an assessment, such as the way that a test is given (Elliott, Ysseldyke, & Thurlow, 1998) or in the testing environment (Thurlow, McGrew, Tindal, Thompson, Ysseldyke, & Elliott, 2000). While validity and accommodations are typically discussed concerning the technical aspects of assessments, accommodations also have the potential to affect the interpretation of assessments for students with disabilities (McLaughlin & Thurlow, 2003).

Several federal laws entitle students with disabilities to accommodations. Section 504 of the Vocational Rehabilitation Act (29 U.S.C. § 612.17a) as well as the Americans with Disabilities Act (28 C.F.R § 35.130(b) (7) require that 'reasonable accommodations' be provided individuals with known physical or mental limitations. The IDEA and the NCLBA both contain language specific to a student's right to accommodations in assessments. The IDEA stipulates that students with disabilities be provided "appropriate accommodations and modifications in administration" (34 C.F.R § 200.138a), while the NCLBA requires states to provide "reasonable adaptations and accommodations for students with disabilities necessary to measure the academic achievement of such students" (34 C.F.R § 200.6(a) (i).

Accommodations are meant to level the playing field so that performance is a result of an individual's achievement and not the result of their disability (Elliott et al., 1998; Shriner, 2000). An accommodation is intended to "minimize the impact of test-

taker attributes that are not relevant to the construct that is the primary focus of the assessment" (AERA et al., 1999, p. 101). A Braille version of a math test, for example, would minimize an individual's visual impairment and allow the individual to demonstrate their math skills without being penalized for their disability.

While providing accommodations to students with disabilities has not proven difficult, the interpretation and treatment of scores obtained with certain accommodations has proven to be a major challenge in accountability systems (Koenig & Bachman, 2004; Thurlow et al., 2000). Section 504 states that a reasonable accommodation may include "…the provision of readers or interpreters, and other similar actions" (34 C.F.R § 104.12b (2). None of the federal laws mandating accommodations defines those accommodations that are considered "reasonable." Federal regulations only specify that the accommodation(s) must not cause an undue hardship (e.g. produce excessive cost, be extensive, or disruptive) on the individual or agency.

Because current accountability systems require that all students' scores be included, states must decide how to report assessment scores obtained with accommodations. Some states, such as Massachusetts allow all accommodations and treat all scores equally. Other states assign students who use certain accommodations (e.g., a reader for elementary reading tests) a minimum score regardless of the student's actual performance. When students' scores are excluded or counted as basic due to accommodations, the validity of interpretations about a school's performance is compromised. It is impossible to determine if schools' scores are low due to actual performance or due to the use of certain accommodations.

Performance trends can be affected by changes in the number and types of accommodations used as well as changes in accommodation policies from year to year (McLaughlin, Embler, Hernandez, & Caron, 2005; Thurlow, 2004; Ysseldyke & Nelson, 2002). For example, half of a group of students may use a non-standard accommodation one year and receive scores of basic, while no students use a non-standard accommodation the following year. Likewise, accommodations that are considered standard one year may not be considered standard the following year. As a result, observed year to year changes in performance can be reflective of the number and types of accommodations as opposed to true changes in performance (McLaughlin et al., 2005).

In addition to validity, Baker et al. (2002) maintain that the accountability system should also be reliable. Although schools' assessment scores will vary from year to year, the degree of this variation (often referred to as volatility) should not be such that the ratings schools receive differ vastly from year to year (Hill & DePascale, 2003). Even the most seemingly valid accountability system is flawed if there is so much volatility in the system that the ratings schools receive change drastically from year to year (Hill & DePascale, 2003; Kane & Staiger, 2002).

Fairness

Closely related to validity is the standard of fairness (Fuhrman, 2003; Kifer, 2001; Linn, 2001; Stecher et al., 2003). In accountability systems, fairness means that all individuals and groups are treated equally. Fuhrman (2003) also finds that fairness demands that the rating schools receive be based on the efforts and effectiveness of individuals within the school (Elmore, 2004; Ladd, 2002;

Stecher et al., 2003) and that schools be rewarded and sanctioned based on equal effort. Fairness in accountability systems also requires that judgments regarding school quality be based upon outcomes over which educators have control or factors they can change. An accountability system is fair if schools exerting equal efforts receive equal rewards or sanctions. However, a system is considered unfair if it holds teachers or others accountable for factors not directly under their control (Fuhrman, 2003; Ladd, 2002). Fuhrman (2003) proposes that a system that allocates consequences for teachers or schools based only on students' current levels of performance is neither valid nor fair if variations in students' scores are influenced by factors other than the efforts of the individuals or the quality of instruction.

Fairness is an important principle in the overall theory of action of accountability systems, because it affects the motivation of the major actors as well as their responses to incentives (Ladd, 2002). If the major actors, such as educators, see the accountability system and corresponding system for rating schools as unfair, or the goals as infeasible, the effects of incentives are negligible (Hanushek & Raymond, 2002). The actors are not likely to work toward goals seen as unreachable, even when incentives are attached to their effort.

Inclusiveness

The third standard of accountability systems is inclusiveness (Fuhrman & Elmore, 2004; Linn, 2004). Inclusiveness requires that the accountability system count the scores of all students and that all students' scores count equally. The

scores for some students should not be selectively excluded or counted to a lesser degree (Fuhrman, 2003; Fuhrman & Elmore, 2004; Linn, 2004)).

Thurlow, Quenemoen, et al. (2001) identified six characteristics of inclusive accountability systems as they apply to students with disabilities. The researchers propose that to be inclusive: 1) all students with disabilities should be included in the assessment system; 2) decisions regarding participation are the result of clearly articulated decision-making processes; 3) all students with disabilities are included in public reports, and in the same format and frequency as all other students, regardless of whether they participate with or without accommodations, or in an alternate assessment; 4) the performance of students with disabilities has the same impact on the final accountability system as the performance of other students, regardless of how the students participate in the assessment system (i.e., with or without accommodations, or in an alternate assessment); 5) there is monitoring, ongoing evaluation, and systematic training based on emerging research and best practice; and 6) every policy and practice is based on the premise that all students, regardless of achievement level, must be included.

Baker et al. (2002) propose that an accountability system that does not include all students equally is neither valid nor fair. Accountability approaches that are not inclusive may also give a "distorted and usually exaggerated view of overall performance" (Baker & Linn, 2004, p. 64), because the performance of all students is not represented.

Usefulness

Finally, in order for the accountability system to serve as a guide for improving instruction, programs, and policies, the approach used must produce information that is understandable and useful for all stakeholders (Baker et al., 2002; Fuhrman, 2003). The

NCLBA states that assessments should be used by local education agencies, schools, and teachers to improve the educational achievement of individual students and should be able to identify schools in need of assistance (34 C.F.R § 1111 (B) (10) (a) (b). Therefore, to be useful the accountability system must produce information regarding student and school performance that can be interpreted and used by educational practitioners, policymakers, as well as parents (Doran & Izumi, 2004; Ladd, 2002).

Parents, teachers, and policymakers must clearly understand not only the results of accountability systems but also the basis for conclusions about the quality of schools. For educational practitioners the accountability system must be capable of providing information on the effectiveness of classroom instruction and should serve to support appropriate classroom action. For policymakers the approach must be useful in providing feedback on the effectiveness and quality of programs and policies, as well as identify those schools where additional supports or assistance are needed. For parents the accountability system must be useful in making decisions regarding school choice.

Accountability systems must also be useful in motivating schools and generating appropriate incentives to improve the performance of their students (Ladd, 2002). Because incentives are designed to affect the effort put forth by educators and policymakers, the accountability system must be able to provide a direct link between the incentives and individuals' efforts. In order to make this link, the system must reflect the actions of the actors and be directly related to their level of contribution to students' education (Hanushek & Raymond, 2002).

Accountability approaches can vary significantly in the extent to which they satisfy the standards of validity, reliability, fairness, inclusiveness, and usefulness proposed by Baker et al (2002). However, by using these standards when designing and evaluating accountability systems, policymakers can evaluate the strengths and weaknesses of each system in light of the desired results. Using these standards will also help assure that the theory of action will be fulfilled and that the goal of increased academic achievement will be realized (Baker et al., 2002).

Comparisons of Accountability Approaches

Research has shown that the choice of accountability approach has a significant impact on conclusions regarding the performance of students and schools (Clotfelter & Ladd, 1996; Doran & Izumi, 2004; Goldstein, 1991; Hanushek & Raymond, 2002; Linn, 2001). Schools rated high-performing using one approach have frequently been classified as low-performing or ineffective using a different approach. These findings together with an increased emphasis on accountability reforms and pressure from policymakers have led researchers to examine the effects of different accountability approaches on judgments about student performance and inferences obtained regarding the quality of schools.

Researchers in six studies compared interpretations of student performance and school ratings using different accountability approaches (Table 2). Two (Barton & Coley, 1998; Tindal, 2002) compared quasi-longitudinal approaches, while Ysseldyke and Bielinski (2002) compared results using cross-sectional, quasi-longitudinal, and longitudinal approaches. Two compared adjusted versus unadjusted measures (Clotfelter & Ladd, 1996; McDonnell et al., 1997; 2004), and two (Clotfelter & Ladd, 1996;

McDonnell et al., 1997) utilized cross-sectional, quasi-longitudinal, and longitudinal approaches. However, only McDonnell et al. and Ysseldyke and Bielinski disaggregated results for students with disabilities. A final study (Rubenstein et al., 2004) is included in this review even though it measures school efficiency and is primarily concerned with financial inputs; because it provides the foundation for the matrix utilized in this study and the framework for evaluating schools across approaches.

Tindal (2002) found important differences in the conclusions regarding academic performance when comparing results obtained using a cross-sectional trend approach to those obtained using a longitudinal approach. Tindal examined the reading and math scores of approximately 1,900 students who had taken the Oregon state assessment for three or four consecutive years (3^{rd}-5^{th} grade, 5^{th}-8^{th} grade, or 8^{th} – 10^{th} grade). While the state had reported "small improvements" in the percentage of students passing the state assessment using a quasi-longitudinal approach, the author reached a different interpretation of the performance of students when he analyzed data for a students who had who had been enrolled in the state an taken the assessment across all years.

Tindal (2002) divided students into four categories: those who failed the assessment repeatedly, those who failed then passed, those who passed then failed, and those who passed the assessment repeatedly. Tindal found that approximately 60% of the students passed the state reading and math assessments in all years, while the percentage of those failing repeatedly ranged from 9% (for 3^{rd}-5^{th} grade students) to 24% (for 8^{th}-10^{th} grade students). However, the percentage of students who passed the assessment in the earlier years, but failed the assessment in the later years increased, leading Tindal to

Table 2

Summary of Reviewed Studies

Researchers	Sample	Performance Measure	Approaches Analyzed
Barton & Coley (1998)	Students from across the United States; N = not provided	NAEP	b, c
Clotfelter & Ladd (1996)	General and special education students in South Carolina N = 47,000 for quasi-longitudinal approaches N = 41,650 for longitudinal analyses	CTBS	b, c,
McDonnell, Morison, & McLaughlin (1997)	Students with disabilities from 337 schools across the US N = 10,333 for cross-sectional approaches; N = 7,906 for longitudinal analyses	CTBS	a, b, c
Rubenstein, et al. (2004)	Third through fifth grade general and special education students in New York City, N = 602 schools Fourth and sixth grade general and special education students in Ohio; N = 783 schools	CTBS and CAT NY state reading and math assessment Ohio state math and writing assessment	a, b
Tindal (2002)	General and special education students in Oregon N = 1,900 students in cross-sectional trend analyses N = 1,097 students in longitudinal analyses	Oregon state reading and math assessment	b, c
Ysseldyke & Bielinksi (2002)	General and special students in Texas, N = 217,519	Texas Assessment of Academic Skills (TAAS)	a, b, c

Note. a = cross-sectional approach, b = quasi-longitudinal approach, c = longitudinal approach

conclude that the performance of students in the state had actually declined.

Because the author did not disaggregate results for students with disabilities, it is uncertain if these results would be replicated with this subgroup. It is plausible that as both the level of the assessment and minimum criteria for passing increase, the percentage of students with disabilities failing would be higher than that observed in the population as a whole. Likewise, it is also possible that due to remediation the percentage of students passing would remain level or even increase. Although Tindal's study is inconclusive with regard to students with disabilities, it illustrates the differences in interpretations that can be reached between quasi-longitudinal and longitudinal approaches.

Barton and Coley (1998) analyzed scores from all 4th – 8th grade US students participating in the National Assessment of Educational Progress (NAEP) using two different two quasi-longitudinal approaches. Barton and Coley investigated the change in reading, math, and science scores of students between 1973-1977 and 1992-1996, and change in writing scores between 1984-1988 and 1992-1996. The authors reported results for students overall and disaggregated by race and state.

When using a quasi-longitudinal approach, comparing different students across years, the researchers found that scores in all subjects had increased or remained level between the two periods. In science, mathematics, and reading, scores for fourth and eighth graders in 1996 had increased from scores in 1973. In writing, scores for both grades in 1996 had not significantly changed from those in 1984.

Barton and Coley then examined the change in scale scores from fourth to eighth grade for the earlier group (1973-1977), compared to changes in scale scores from fourth

to eighth grade for the later group (1992-1996). While average scores had increased or remained level during this time, scores for the same students had remained level or decreased for both groups in all subjects. For example, in math, the average scores increased by 50 points from fourth to eighth grade in 1973-1977. In the same subject and grades, however, longitudinal analyses of the same students revealed an average scale score gain of only 45 points from 1992-1996, a statistically significant difference.

A comparison of scores disaggregated by state and race also revealed significant differences. For example, in 1992, the average mathematics scale score for the lowest scoring state, Arkansas, was 210, compared to Maine, the highest state, with an average scale score of 232. In 1996, the scale scores for Arkansas and Maine had risen to 262 and 284 respectively. While the mean score of Arkansas was below that of Maine in both years, both states had the same growth of +52 points between 1992 and 1996. In fact, Barton and Coley found that Nebraska, the state with the highest growth of 57 points, was only statistically higher than thirteen other states; and there were no significant differences in the gain scores of 37 of the 50 states.

The authors obtained similar results when comparing assessment scores disaggregated by race. When examining average scores for Black students in 1992 and 1996, the authors found that in both years, Black students scored significantly below White students. In math, Black students had an average scale score of 200 in fourth grade and 243 in eighth grade. In contrast, White students had an average scale score of 232 in fourth grade and 282 in eighth grade. While the performance of Black students was significantly lower than the performance of White students in both grades, growth

scores for the two groups were not statistically different. The scores for Black students increased by 43 points while the average for White students increased by 50 points.

Barton and Coley (1998) concluded that the current method of analyzing NAEP scores, by comparing averages from year to year, yields interpretations that are "quite apart" from those obtained by measures that follow a group over time (p. 15). While the authors did not attempt to determine which is the better approach, they stressed the importance of looking at both measures in order to obtain the information necessary to make definitive conclusions regarding student performance.

In a study similar to that of Barton and Coley, Ysseldyke and Bielinski (2002) analyzed assessment results for students in Texas, disaggregated by disability status. Their sample included all students with scores (N =217,519) on the Texas Assessment of Academic Skills (TAAS) for five consecutive years, beginning in 1993 as fourth graders. The authors calculated the achievement gap between general and special education students as an effect size using a quasi-longitudinal approach as well as the longitudinal approaches. Effect sizes were calculated so that positive values indicated that the mean for students receiving special education services was above that of students in general education, while negative values indicated that their scores were below that of students in general education.

Using a cross-sectional approach, Ysseldyke and Bielinski found that the average reading and math scores of the special education students in fourth grade were significantly below the scores of students in general education. This finding was evident in all grades and the effect size increased as students progressed in grade. In reading, the

effect size for special education grew from .64 in fourth grade to -1.16 in eighth grade, an increase of approximately .50 standard deviations.

The authors then used quasi-longitudinal and longitudinal approaches to analyze student performance. Results obtained with both of these approaches revealed smaller gaps between students in special and general education than was evident under the cross-sectional approach.

The mean for students with disabilities using the quasi-longitudinal approach was at the 18th percentile of the general education score distribution, and the effect size increased from -.48 in fourth grade to -.93 in eighth grade. Under the longitudinal approach, the effect size actually decreased slightly from -.48 in 4th grade to -.42 in 8th grade, and the mean for students with disabilities was at the 28th percentile of the general education distribution. Findings were similar in math, with an achievement gap of .52 standard deviation units under the cross-sectional trend approach and .56 standard deviation units under the quasi-longitudinal approach. In contrast, the achievement gap decreased by .04 standard deviation units when using a quasi-longitudinal approach and when examining student performance over time.

Although their sample was large, Ysseldyke and Bielinski (2002) cautioned that their results may not be representative of students with disabilities as a whole due to possible bias in their sample. Only 40% of special education students in fourth grade and up to 61% of special education students in eighth grade had available test scores. In addition, both general and special education students with low assessment scores one year were more likely to be missing scores the following year. Therefore, their sample

most likely represents general and special education students with higher academic achievement.

Of the three studies that included unadjusted and adjusted measures, only one reported results at the student level (McDonnell et al., 1997). The remaining two compared results at the school level and examined the changes in school's rankings under various approaches (Clotfelter & Ladd, 1996; Rubenstein et al., 2004). Although Clotfelter and Ladd (1996) used individual student level data, they aggregated and reported results at the school level.

Clotfelter and Ladd (1996) analyzed fourth and fifth grade reading and math scores on the Comprehensive Test of Basic Skills (CTBS) from South Carolina and compared results obtained using nine different analytical methods. The nine methods included cross-sectional and quasi-longitudinal approaches (i.e., two measures of simple gain scores), and six different longitudinal measures. For the quasi-longitudinal analyses, the authors used assessment data for all students in fifth grade in 1994, which included 45,872 students in 575 schools.

Using each of the nine different approaches, the authors calculated individual rankings for each school. In the first method, Clotfelter and Ladd calculated the mean or average score for 45,872 fifth grade students in 575 schools. In the two quasi-longitudinal measures, the researchers used the scores for all students in fourth grade in 1993 and all students in fifth grade (N = 45,872) in 1994. Clotfelter and Ladd then calculated the difference between the average fifth grade and average fourth grade scores, and the percentage change in scores between fourth and fifth grade. The final six longitudinal measures included only those students who had been in the same school in

fourth and fifth grade with assessment data in both years. This sample included 41,650 students in 571 schools. The fourth approach was based on the difference between the average fourth grade score in 1993 and the average fifth grade score in 1994, but included only those students for whom data were available in both years. The fifth approach used the school gain index (SGI), calculated by taking the difference between the students' fifth grade score, minus the predicted score for that the student. In the sixth measure, an adjusted SGI, the researchers used the residual from the SGI calculated in the fifth measure. The seventh and eighth measures were similar to the SGI measures, but added control variables. In measure seven, adjustments were made for socioeconomic status (percentage of students receiving free and reduced lunch), while in measure eight, adjustments for the racial composition of the school were added. Both measures used linear regression to calculate the effectiveness of the school, measured by the difference between the actual test scores and the adjusted predicted scores. Measure nine was based on the mean of residuals, after controlling for the socioeconomic characteristics of all students in the school.

The researchers calculated the correlation matrix of the schools' rankings under the nine measures. Clotfelter and Ladd report that the correlations ranged from a high of .94 to a low of 0, with the highest correlations observed between measures that were variants of each other, such as the gain measures in methods two through four, and those based on residuals, measures seven and eight. Method one, the mean or average score, was least correlated with the other measures, and evidenced a correlation of zero with measure six, the adjusted performance score.

The authors then calculated the correlation between the schools' rankings and measures of socioeconomic status and the percentage of Black students in the school. Average scores as well as measures of gain scores were negatively correlated to both the percentage of students receiving free and reduced lunch, as well as the percentage of the student body that was Black. Furthermore, these correlations were higher in reading than in math.

Clotfelter and Ladd (1996) also reported the characteristics of schools in the top and bottom 25th percentiles under each of the nine different methods. Using average scores, schools in the top 25th percentile had an average of only 23% of their students receiving free/reduced meals and 22% who were Black, compared to those below the 25th percentile that averaged 51% of their student population in each category. Using adjusted measures, the school compositions between those in the top and bottom groups were relatively equal.

The researchers concluded that only those measures that adjust for both prior learning and demographic characteristics treat schools fairly. They also advise that adjustments for prior achievement alone do not provide fairness. To be considered fair, adjustments must also be made for socioeconomic status. However, the researchers caution that in using complex measures to make adjustments, transparency may be lost.

Only one study disaggregated results for students with disabilities and used data from individual students for their analyses. McDonnell et al. (1997) analyzed data from the Prospects study to report on the achievement of students with disabilities. Prospects was a national longitudinal study mandated by Congress in 1988 to examine the effects of Title I on the academic achievement of students.

The original sample for Prospects contained three grade cohorts of students and included individual student level data, as well as information from parents, teachers, and principals. All students within selected schools and grades were included in the study, and no students were excluded due to disability or English proficiency. For their analyses, the authors utilized only the third grade cohort, which included 337 schools with 10,333 students.

The authors used more than twelve different methods including cross-sectional scores; estimation models controlling for individual, district, and family characteristics; and models with and without controls for prior achievement. In all models, outcome variables were the reading and math normal curve equivalent scores (NCEs) on the Comprehensive Test of Basic Skills (CTBS).

First, the authors used mean changes in scores from third grade to fourth grade. They discovered that the overall scores for students decreased from third to fourth grade (-1.1 NCEs), but that decreases in scores were evident for the majority of groups disaggregated by ethnicity and disability. Only three disaggregated groups, Asian Americans, Hispanics, and students with speech disabilities increased their scores. In contrast, the scores of all other major racial groups, as well as all other groups of students with disabilities decreased, with the largest decreases were evidenced by "Other American" which declined –2.2 NCEs, African Americans (-1.7 NCEs), and students with emotional disabilities (-1.6 NCEs).

While the authors did not observe significant variance around the mean, the overall variance in all but three of the twenty groups ranged between 18 and 20 NCEs. This variance within groups led the authors to find simple cohort analysis "…misleading

for assessing both achievement levels and educational progress" (p.265). However, the sample was not drawn to be representative of students with disabilities, only those included in the school and tested were included in the analyses. Thus, the results for students with disabilities must be interpreted with great caution.

In order to obtain a more detailed evaluation of educational achievement, various student, family, and district level controls were incorporated into the regression models. For each subject area (reading and math), regression models were developed both with and without controls for prior achievement, resulting in twelve regression analyses for each subject area. After controlling for various racial, socioeconomic, and family characteristics, but without controlling for prior achievement, there were significant effects for all ethnic groups. Except for "Other American" in mathematics, all ethnic groups scored significantly below Whites ($p<.001$). However, students with learning disabilities were the only group that scored significantly ($p<.001$) below their non-disabled peers, at –17.49 NCEs in reading and –16.12 NCEs in math.[7]

Prior achievement was positively ($p<.001$) related to present achievement. After controlling for prior achievement, the effects of disability, ethnicity, and socioeconomic status, were much smaller. For African Americans, the effect in reading decreased to –1.68 NCEs ($p<.05$) and became non-significant in math. Likewise, the effect in reading for students with learning disabilities decreased to –4.09 ($p<.05$) but was also non-significant in math. However, in all models, family education and expectations remained positively related to achievement. In their final recommendations, McDonnell et al. concluded:

> …policymakers undertaking standards-based reforms still need to compare student achievement over time, across populations, and between organizations … value-added models, which control for prior achievement, offer promise as a valid method for reporting achievement scores and should be considered by policy makers (p. 275).

A final study by Rubenstein et al (2004) is reviewed even though the primary focus of the study was to compare economic measures of school efficiency and not academic performance. This study is included because it illustrates the variability that can occur when different approaches are used to measure the same outcome variable. This study also provided the basis for analyzing school rankings across approaches and is the basis for the matrix used in the current study.

Using school level data from 783 schools in Ohio and 602 schools in New York City, Rubenstein et al. (2004) applied four different approaches to measure the effects of student and school level factors on school rankings. Using each of the four approaches, Rubenstein et al. computed the schools' percentile rankings. Achievement measures for New York included third through fifth grade reading and math scores on the Comprehensive Test of Basic Skills (CTBS) and English Language Assessment (ELA), and the California Achievement Test (CAT). In Ohio, outcome measures were the passing rates for fourth through sixth grade students on the state writing and math assessments. The four approaches used included adjusted performance measures (APMs), education production functions (EPFs), data envelopment analysis (DEA), and a cost function analysis (CFA).

APM and DEA methods are considered cross-sectional approaches because they measure efficiency at a single point in time and do not make comparisons across observations. APM approaches are based on a linear regression that uses academic achievement as the dependent variable. School and grade level aggregate information as used as predictors. School rankings are determined by the residual value. DEA is similar to APMs in that is uses the residuals from a linear regression equation to calculate a school's ranking. However, DEAs predict a school's performance based on a school's distance from the highest result obtained as opposed to its distance from the mean. The DEAs also included multiple dependent variables in the same equation (such as reading and math) as opposed to APMs that use only one dependent variable.

EPF and CFA approaches are characterized as quasi-longitudinal because they employ two or more years of data. The EPF measure the change in the outcome measure use the residual to determine a school's ranking. The CFA approach calculates the minimum cost of producing a certain output, and bases school rankings on their distance from the minimum cost.

Rubenstein et al. divided the schools into five categories based on percentile rankings: 1) schools below the 10^{th} percentile, 2) schools between the 11^{th} and 25^{th} percentile, 2) schools between the 26^{th} and 75^{th} percentile, 4) schools between the 76^{th} and 89th percentile, and 5) schools at or above the 90^{th} percentile. They classified schools below the 10^{th} percentile as low-performers in efficiency and those above the 90^{th} percentile as high-performers in efficiency. The researchers then calculated how many schools changed categories, or rankings, between the various measures. They found that schools labeled as high performers using one method were often labeled as low-

performers using another method. For example, comparing the EPF measure to the APM measure for reading, 76.3% of the schools moved up or down only one category. However, 11.7% of the schools moved up two categories, while 11.9% of the schools moved down more than one category. A negligible percentage, 0.1%, moved down two categories. Because Rubenstein et al. only reported the number of schools that changed categories and did not report the characteristics of the schools, it is uncertain which schools switched their standing, or why this may have occurred.

Rubenstein et al. (2004) concluded that while the methods they employed were unlikely to produce vastly different lists of the highest and lowest performing schools, their results were susceptible to the quantitative approaches employed. They further concluded that "… simplistic measures of school performance, which do not account for the complex environment of schools, risk identifying the wrong schools as being exemplars or those in need of interventions" (Rubenstein et. al, p. 20).

These studies demonstrate the effect that various approaches in measuring academic progress can have on interpretations of student progress and school performance. However, none of the studies provided information on how the use of varying approaches affects interpretations for the subgroup of students with disabilities. Further investigation is needed to determine whether the results observed with students with disabilities follow the same patterns as those conducted with students in general education.

Summary

The inclusion of students with disabilities in large-scale assessments and school accountability are recent practices. Due to these students' prior unsystematic participation in state assessments and the nacency of legislative mandates requiring their

participation, researchers and policymakers have little experience with analysis and interpretation of performance trends for students with disabilities. The bulk of current knowledge regarding the effects of different approaches on school accountability is based on evidence using total school populations, leaving a void in information on the effects of various accountability approaches for the subgroup of students with disabilities (Koretz & Barton, 2003)

The overall goal of accountability systems is the improvement of instruction and student learning (O'Day, 2004). Moreover, recent educational reforms such as the NCLBA have made the performance of the subgroup of students with disabilities one of the components upon which the quality of schools will be rated. States are required to assess students with disabilities and include the results of these assessments in accountability systems. States and schools are also required to use the results of assessments to judge the overall performance of schools and to sanction and reward schools. Results of assessments for students with disabilities are also being used to make program adjustments and determine where scarce resources are most needed (Kane et al., 2001). As such, it is imperative that administrators and policymakers have confidence that the interpretations of assessments provide the information necessary to increase the academic achievement of all students.

Although substantial research exists examining the technical adequacy and pragmatic use of assessments within accountability systems, little attention has been paid to the approaches and methods employed to measure and interpret assessment results (Rubenstein et al., 2004). This is especially true for subgroups such as students with disabilities who must now be analyzed separately under the NCLBA. While theoretical

discussion indicates that longitudinal value-added models provide more accurate information and information that can be useful to educators, policymakers and parents, there is scant empirical evidence (Goldhabar, 2001) and the area remains largely unexplored (Hanushek & Raymond, 2003).

The different methodological approaches available to interpret assessment data can lead to divergent inferences about the performance of schools. Given that the goals of an accountability system are to identify schools in need of assistance and reward those that that are 'doing well," correctly rating schools is paramount to the success of accountability reforms. This study addressed these issues by applying five different accountability approaches to the subgroup of students with disabilities. In doing so, the study examined not only the classifications of schools under the different approaches, but also examined the results concerning the reliability, inclusiveness, and usefulness of the different approaches for this subgroup.

Chapter III

Data and Methodology

The purpose of this study was to compare five approaches used to evaluate schools based on the performance of students with disabilities. Using four years of extant assessment data from a large school district, each approach was applied to the data to classify schools as high-performing or low-performing. Results for each approach were examined to determine the characteristics of schools classified in each category and the reliability of each approach in rating schools.

This chapter presents the data and methodology for the current study. Descriptive information on the school district, schools, and students with disabilities in the district are described in the first section. Following this, the dependent and independent variables used in the current study are presented as they are defined and operationalized in the state used in the study. In the final section, the methods of data management and explanation of the analyses used in the study are described.

Data Sources

The data used in this study are from a large suburban school district in the mid-Atlantic region of the United States. Overall demographic data, state and district assessment and accountability policies, and special education accommodation and placement policies were obtained from the state and district websites. To protect the identity of the SEA and LEA, these sources will not be cited in this paper.

The district in this study was selected for its size, diversity, and available dataset. The district is the 18th largest school system in the United States and enrolled approximately 140,000 students from 161 nations in 2004. Of the district's total school-

age population, 22% were African American, 19% were Hispanic, and 45% were White. One-fifth of the students received free and reduced meals (FARMs) and the overall mobility rate in the district averages 15%. The district has 125 elementary, 36 middle, and 23 high schools averaging 504, 800, and over 1,000 students respectively. In addition, the Department of Special Education operates seven special schools serving approximately 925 students who have severe emotional, cognitive, or physical disabilities.

The demographic and assessment data for the current study were originally obtained from the district as part of a special education program evaluation conducted by a district advisory group. The original task of the advisory group was to develop a set of indicators to evaluate and improve special education programs in the district. To achieve this, four years of demographic and performance data were collected from the Special Education Accountability Office and the Office of Shared Accountability: SY99-00, SY00-01, SY01-02, and SY02-03 (hereafter 2000, 2001, 2002, and 2003).

The Special Education Office of Shared Accountability provided demographic and service delivery data for all students receiving special education services in the district, including: grade, race, gender, socioeconomic status (free and reduced meals), attendance, English language status, primary disability, hours of special education service, service providers, least restrictive environment (LRE), and school (home school, school of enrollment, and servicing school). The Office of Shared Accountability provided assessment data for students with disabilities in the district: CTBS and CAT/5 assessment scores for students in grades 2, 4, and 6; state assessment scores for students in grades 3, 5, and 8; and high school assessment scores for students in grades 9-12.

The demographic datasets contained information on: 16,110 students in 2000; 16,162 students in 2001; 16,251 students in 2002; and 16,779 students in 2003. Each year, the demographic data were merged with performance data using the unique student identification numbers. Several filters were then applied to each dataset to obtain the samples used in the current study (Table 3). First, all students enrolled in grades two, four, and six as of the December child count were selected.[8] Next, only students whose home school was a public school on the date of testing were selected.[9] Finally, all students who withdrew prior to the date of testing were excluded.

These samples were then further reduced by the exclusion of students with missing assessment data or students using non-standard accommodations. In each target year, only those students with disabilities who had valid reading and mathematics scores were included in the final samples. The demographic characteristics of students in the final subsamples are presented in Tables 4-7. Table 4 summarizes the racial characteristics of the samples, while Table 5 presents the gender and socioeconomic status of the samples. The disability categories and least restrictive environment (LRE) of the samples are presented in Tables 6 and 7.

The demographic characteristics of schools based on students with disabilities with valid assessment scores are presented in Tables 8-11. Table 8 provides information on the average number, minimum number, and maximum number of students with disabilities in the three types of schools (elementary, middle, and special schools). Tables 9-11 present the average percentage, minimum percentage, and maximum percentage of students in the schools by racial category (Table 9), socioeconomic status (Table 10), and disability category (Table 11).

Table 3

Number of Students Excluded from Sample

	2000	2001	2002	2003
Total Students with Disabilities	16,110	16,182	16,251	16,779
Not in grades 2, 4, or 6	12,459	12,505	12,598	13,042
Not in public school	62	70	111	99
Withdrew before test	46	45	51	50
Total	3,543	3,562	3,491	3,588

Table 4

Race of Subsamples by Year and Subject Area

	2002						2003					
	Sample		Reading Subsample		Mathematics Subsample		Sample		Reading Subsample		Mathematics Subsample	
	N	%	*n*	%	*n*	%	*N*	%	*n*	%	*n*	%
American Indian	6	0.2	5	0.2	3	0.2	11	0.3	11	0.3	11	0.3
Asian American	225	6.4	191	6.0	136	7.7	46	6.9	217	6.6	217	6.6
African American	952	27.3	870	27.3	389	22.0	991	27.6	899	27.4	896	27.4
Hispanic	655	18.8	610	19.2	289	16.3	44	20.7	685	20.9	686	21.0
White	1,653	47.4	1,508	47.4	955	53.9	1,596	44.5	1,466	44.7	1,462	44.7
Total	3,491	100.0	3,184	100.0	1,772	100.0	3,588	100.0	3,278	100.0	3,272	100.0

Table 5

Gender and Socioeconomic Status of Subsamples by Year and Subject Area

	2002						2003					
	Sample		Reading Subsample		Mathematics Subsample		Sample		Reading Subsample		Mathematics Subsample	
	N	*%*	*n*	*%*	*n*	*%*	*N*	*%*	*n*	*%*	*n*	*%*
Gender												
Female	1,120	32.1	1,013	31.8	496	28.0	1,159	32.3	1,060	32.3	1,056	32.3
Male	2,371	67.9	2,171	67.7	1,276	72.0	2,429	67.7	2,218	67.7	2,216	67.7
Total	3,491	100.0	3,184	100.0	1,772	100.0	3,588	100.0	3,278	100.0	3,272	100.0
SES												
High SES	1,856	53.2	1,671	52.8	1,087	61.3	1,896	52.8	1,734	52.9	1,731	52.9
Low SES	1,635	46.8	1,513	47.5	685	38.7	1,692	47.2	1,544	47.1	1,541	47.1
Total	3,491	100.0	3,184	100.0	1,772	100.0	3,588	100.0	3,278	100.0	3,272	100.0

Note. High SES = students who have never received FARMS; Low SES = students who currently or previously received FARMS.

Table 6

Disability Classification of Subsamples by Year and Subject Area

	2002						2003					
	Sample		Reading Subsample		Mathematics Subsample		Sample		Reading Subsample		Mathematics Subsample	
	N	%	*n*	%	*n*	%	*N*	%	*n*	%	n	%
Speech/Language	1,371	39.3	1,436	42.3	964	54.4	1,466	40.8	1,346	43.8	1,437	43.9
Learning Disabled	1,206	34.5	1,183	37.1	502	28.3	1,211	33.8	1,180	36.1	1,180	36.1
Other Health Impaired	245	7.0	262	7.4	137	7.7	278	7.7	235	8.0	262	8.0
Multi-handicapped	345	9.9	128	6.3	51	2.9	234	6.5	200	3.9	125	3.8
Emotionally Disturbed	75	2.1	93	3.3	55	3.1	108	3.0	104	2.8	92	2.8
Autistic	112	3.2	69	1.1	19	1.1	100	2.8	34	2.1	69	2.1
Mentally Retarded	57	1.6	12	0.3	0	0.0	90	2.5	11	0.4	12	0.4
Hard of Hearing	24	0.7	35	0.7	18	1.0	37	1.0	23	1.1	35	1.1
Orthopedically Impaired	14	0.4	17	0.3	5	0.3	18	0.5	11	0.5	17	0.5
Visually Impaired	12	0.3	16	0.4	10	0.6	16	0.4	12	0.5	16	0.5
Deaf	22	0.6	22	0.6	9	0.5	24	0.7	20	0.7	22	0.7
Traumatic Brain Injury	8	0.2	5	0.3	2	0.1	6	0.2	5	0.2	5	0.2
Total	3,491	100.0	3,278	100.0	1,172	100.0	3,588	100.0	3,184	100.0	3,272	100.0

Note. There were no students classified as deaf-blind either year.

Table 7

Least Restrictive Environment (LRE) of Subsamples by Year and Subject Area

	2002						2003					
	Sample		Reading Subsample		Mathematics Subsample		Sample		Reading Subsample		Mathematics Subsample	
	N	%	*n*	%	*n*	%	*N*	%	*n*	%	*n*	%
LRE A	1,590	45.5	1,554	48.8	1,213	68.5	1,851	51.6	1,819	55.5	1,818	55.6
LRE B	691	19.8	673	21.1	309	17.4	603	16.8	581	17.7	582	17.8
LRE C/SS	1,200	34.7	957	30.0	250	14.1	1,134	31.6	878	26.8	872	26.6
Total	3,491	100.0	3,184	100.0	1,772	100.0	3,588	100.0	3,278	100.0	3,272	100.0

Note. SS =special schools (public day and residential schools).

Table 8

Average, Minimum, and Maximum Number of Students with Disabilities in Elementary, Middle, and Special Schools

	2002						2003					
	Reading			Mathematics			Reading			Mathematics		
School Type	*M*	Min. *N*	Max *N*	*M*	Min. *N*	Max *N*	*M*	Min. *N*	Max *N*	*M*	Min. *N*	Max *N*
Elementary	16	2	40	15	2	40	16	3	35	16	3	34
Middle	35	16	57	35	16	57	35	17	60	35	17	60
Special	10	0	11	10	0	10	14	8	21	14	8	21

Table 9

Average, Minimum, and Maximum Percentage of Students with Disabilities per School by Race and Socioeconomic Status

	2002			2003		
	Mean %	Min. %	Max. %	Mean %	Min. %	Max. %
Race						
American Indian	0.3	0.0	12.5	0.3	0.0	12.5
Asian American	6.8	0.0	36.5	7.0	0.0	60.0
African American	26.4	0.0	90.0	27.0	0.0	100.0
Hispanic	17.6	0.0	72.7	21.2	0.0	100.0
White	48.9	0.0	100.0	44.5	0.0	100.0
SES						
High SES	55.7	0.0	100.0	53.9	0.0	100.0
Low SES	44.3	0.0	100.0	46.1	0.0	100.0

Note. High SES = students who have never received FARMS; Low SES = students who currently or previously received FARMS.

Table 10

Average, Minimum, and Maximum Percentage of Students with Disabilities in Elementary and Middle Schools by Disability Category

	2002			2003		
	Mean %	Min. %	Max. %	Mean %	Min. %	Max. %
Speech/Language	43.5	7.1	100.0	46.4	0.0	100.0
Learning Disabled	30.5	0.0	71.4	30.2	0.0	75.0
Other Health Impaired	6.4	0.0	29.6	7.6	0.0	37.5
Multi-handicapped	8.7	0.0	50.0	5.5	0.0	55.6
Autistic	2.2	0.0	42.8	2.9	0.0	50.0
Emotionally Disturbed	2.8	0.0	40.0	2.4	0.0	50.0
Mentally Retarded	1.7	0.0	40.0	2.2	0.0	60.0
Hard of Hearing	0.8	0.0	20.0	1.1	0.0	18.2
Orthopedically Impaired	0.3	0.0	21.1	0.4	0.0	29.2
Visually Impaired	0.4	0.0	20.0	0.5	0.0	14.3
Deaf	0.6	0.0	25.9	0.7	0.0	35.7
Traumatic Brain Injury	0.2	0.0	14.3	0.2	0.0	10.0

Note. There were no students identified as deaf-blind in any year.

Table 11

Average, Minimum, and Maximum Percentage of Students with Disabilities in Special Schools by Disability Category

	2002			2003		
	Mean %	Min. %	Max. %	Mean %	Min. %	Max. %
Speech/Language	1.4	0.0	5.7	5.3	0.0	21.2
Learning Disabled	1.4	0.0	5.7	-	-	-
Other Health Impaired	0.7	0.0	2.9	-	-	-
Multi-handicapped	71.3	17.0	100.0	56.0	10.0	100.0
Autistic	2.9	0.0	11.4	9.5	0.0	21.2
Emotionally Disturbed	20.8	0.0	83.3	23.9	0.0	90.0
Mentally Retarded	0.7	0.0	2.9	2.5	0.0	9.6
Orthopedically Impaired	-	-	-	2.4	0.0	9.6
Visually Impaired	-	-	-	-	-	-
Traumatic Brain Injury	0.7	0.0	2.9	-	-	-

Note. There were no students identified as deaf-blind, hard of hearing, or visually impaired enrolled in special schools in any year.

Dependent Variables

The dependent variables in all analyses were students' standardized mathematics and reading scale scores on standardized assessments. These subject areas were chosen because they are the subjects currently mandated under NCLBA for school accountability. For students in second and fourth grades, academic performance was measured by the CTBS. Performance for students in sixth grade was measured by the CAT/5. For all grades, reading performance was measured by the reading subtest. Mathematics performance was measured by the mathematics composite score, formed by averaging the mathematics and mathematics computation subtest scores.[10] Both the scale scores and gain scores were standardized within grade.

The CTBS and the CAT/5 are designed to measure "concepts, processes, and skills taught throughout the nation" and are based on content found in state, district, and private school curriculum guides, as well as major textbooks, programs and standards (CTB-McGraw Hill, 2001, p. 1). The reading assessment is designed to assess students' knowledge in comprehension, including main idea, inference, and drawing conclusions. The mathematics subtests are aligned with the National Council of Teachers of Mathematics (NCTM) standards and includes questions assessing students' skills in estimation, numeration, problem-solving, patterns, calculator use, and computation.

CTB McGraw-Hill reports that the CAT/5 and the CTBS are vertically equated, meaning that the scale scores of the two tests can be incorporated into longitudinal studies with a high degree of confidence (CTB McGraw-Hill, 2004).[11] Because the two assessments are vertically equated, scores can be added, subtracted, and averaged across test levels. In addition, the year-to-year growth of individual students or groups can also

be tracked using scale scores. The test developer reports that both assessments are appropriate to analyze strengths and weaknesses of programs, report progress, and as measures of school effectiveness.

Independent Variables

In this study certain demographic and service delivery data were chosen as independent variables. The selected variables are described in the following section. The independent variables are presented below, along with the definition as operationalized in the district. To aid in interpretation, the variable names were changed to a more "user-friendly" notation. The renamed variable appears first, followed by the variable names as they appear in the original dataset inserted in brackets.

Grade [EGRAD]. The student's grade of enrollment as reported in the December child count.

Race [ERACE]. The race of the student is the official race as reported on the district enrollment form. The district uses the five racial/ethnic categories established by the U.S. Department of Education: African American, American Indian, Asian American (to include Pacific Islander), Hispanic, and White. There is no category for multi-racial and no option to choose more than one race. For students in grades PreK-6, race is reported by the parent/guardian(s) upon enrollment. If parents do not indicate the student's race, it is determined by school officials.

Socioeconomic Status [FRRD]. The district uses free and reduced meals (FARMs) as the proxy for socioeconomic status (SES). In the original datasets students were coded into one of three SES categories: never having received free-reduced meals, previously having received free-reduced meals, or currently receiving free-reduced

meals. Due to the cumulative effects of poverty and fluidity between the "currently" and "previously" receiving free/reduced meals groups, these two categories were collapsed into two categories and renamed. The district also uses this dichotomous classification of SES when evaluating student and school performance. The two categories used in the current study were: 0= high SES (never received FARMS) and 1 = low SES (currently or previously received FARMs).

Calculator [CALC]. Calculator indicates whether the student received the accommodation of a calculator in completing the mathematics or mathematics computation assessments and is coded: 0 = no calculator used and 1 = calculator used.

Disability Category [DISABL]. The disability category is the official disability of the student as determined by the IEP team and reported on the student's IEP. The district classifies students according to the 13 disability categories defined in IDEA: mentally retarded, deaf, hearing impaired, autistic, learning disabled, speech and language impaired, emotionally disturbed, autistic, other health impaired, orthopedically impaired, traumatic brain injury, blind, and visually impaired. The characteristics and criteria for each disability category are defined in the state code of regulations (COMAR §13A.05.01.03) (Appendix B).

Due to the small number of students with valid assessment scores in certain disability categories, disability was collapsed into four groups in the current study and renamed: Disability 1 = "sensory/emotional impaired" (hard of hearing, deaf, visually impaired, orthopedically impaired, deaf/blind, emotionally disturbed, other health impaired; Disability 2 = "intellectually impaired" (mentally retarded, traumatic brain injury, autistic, multi-handicapped; Disability 3=learning disabled, and:

Disability 4 =speech and language impaired. While it is recognized that the academic characteristics of students within each of these groups may vary, the groups were collapsed based on similarities in assessment performance, state criteria for placement, and the number of students within each disability category.

Extra Time [EXTIM]. Extra time indicates whether the student received the accommodation of extra time in completing the assessments. This variable is coded: 0 = no extra time provided and 1 = extra time provided.

Least Restrictive Environment [LRE]. The least restrictive environment (LRE) is the setting where the IEP is implemented and represents either the location of services or the percentage of time the student is educated outside of the general education classroom. In the original dataset LRE was entered as an alpha code ranging from A to O: LRE A = outside of the general education classroom less than 21% of the time; LRE B = outside of the general education classroom between and 21% and 80% of the time; LRE C = outside of the general education classroom more than 80% of the time; LRE D = home instruction; LRE E = hospital services; LRE F = public day school; LRE G = private day school; LRE H = public residential school; and LRE I = private residential school. LRE categories J-O apply only to students in preschool. In the current study only LRE A, B, C, F, and H were used because they apply to students enrolled in and receiving special education services in public schools. Due to the small number of students with valid assessment scores in LRE C, LRE F, and LRE H, these categories were collapsed into one category and recoded "LRE C/special schools."

Last Date of Withdrawal [LWITH]. The date the student officially withdrew from the district. This variable was used to identify those students who officially withdrew from the school before the date of testing.

School. Three variables provided information on the student's school: last school of enrollment [LSCH], current school of enrollment or "home school" [RSCHL], and servicing school [SSCHL]. For each variable the school was identified by the official numeric school code used in state and local reporting. The current school of enrollment was used to identify students who were enrolled in a private school, home-schooled, or receiving services in a hospital. The servicing school was used to identify where the student was receiving special education services and where the students' assessment scores are "counted" in the state's accountability system.[12]

Using a publicly available list of schools obtained from the state department of education website (www.msde.md.us) the official school codes were used to classify schools as public or private and to categorize schools as elementary, middle, or special schools. Once schools were categorized and individual student data were aggregated by school, a random school number [RANDSCH] was generated for each school using SPSS. The randomly generated school numbers were non-consecutive and ranged from 1-200. Once schools were assigned a random number, the official school code numbers were deleted from the database.

Data Management and Analyses

All data were maintained and stored on a Dell personal computer hard drive. The individual student data were organized by the five to seven digit student identification number, and the aggregated school level data were organized by the official school code

number or randomly generated school code. All analyses were conducted using Statistical Program for Social Sciences (SPSS) Version 13.5 or Hierarchical Linear Modeling (HLM) version 5. SPSS is appropriate for generating descriptive statistics and conducting inferential analyses, while HLM is suitable for the analysis of school effects using multi-level and nested data (Raudenbush & Willms, 1995).

Prior to merging and completing the analyses the data were "cleaned." Missing values were properly coded, and value labels for alpha or numeric codes were entered for those undefined variables. Value labels for demographic and assessment data were obtained from the district's data system manual and the Office of School Accountability. Dependent variables were inspected for independence and normal distribution.

Cases missing data on both the mathematics and mathematics computation subtests or the reading subtest were coded as missing. For cases missing scores on only one of the mathematics subtests, the missing subtest score was replaced with the conditional mean based on SES, LRE, and performance on the remaining mathematics subtest. Less than 3% of mathematics scores each year were replaced using mean substitution.

Independent t-tests

Independent *t*-tests were conducted to determine if there were differences in students' reading and mathematics performance based on SES. The dependent variables were the standardized reading and mathematics scale scores and gain scores. The independent variable was the students' SES. Alpha in all analyses was set at .05. For groups with significant differences, effect sizes were calculated as Cohen's *d*, defined as

the standardized difference between groups and defined by the following formula: $d = M_1 - M_2$ /pooled SD.

ANOVAs

A series of one-way analysis of variance (ANOVAs) were conducted to determine if there were differences in students' reading and mathematics performance based on disability category and LRE. Students' standardized reading and mathematics scale scores and gain scores were set as the dependent variables. Disability category and LRE were set as the independent variables. Alpha in all analyses was set at .05.

Post hoc multiple comparisons were performed on groups with significant main effects to determine which groups were significantly different from the others. In analyses with equal variances, post hoc multiple comparisons were conducted using the Tukey HSD (honestly significant difference) procedure. Although the Tukey HSD procedure is conservative in that it reduces Type I error at the expense of power, it is a recommended procedure when comparisons between all groups is desired (Huck, 2000). When comparison groups were of unequal variances and unequal cell sizes, post hoc multiple comparisons were performed using the Tukey-Kramer procedure, an appropriate follow-up procedure when groups have unequal variance as well as unequal cell sizes (Toothacker, 1993).

Cross-Sectional Approach

To answer research question 2(a), each school was classified as high-performing or low-performing for students with disabilities using the cross-sectional approach. First, students' reading and mathematics scale scores were recoded as proficient or nonproficient using the cut scores delineated by CTB McGraw Hill based on the subject

area and student's grade level. The cut scores for proficiency in reading are 616 for second grade, 671 for fourth grade, and 696 for sixth grade. The cut scores for mathematics are 581 for second grade, 669 for fourth grade, and 709 for sixth grade.

Student achievement data were then aggregated to the school level to obtain the number of students with disabilities scoring at or above proficiency and the total number of students with disabilities in each school. From these data the percentage of students in each school scoring at or above proficient was calculated. Based on the percentage of students scoring proficient, each school was recoded as high-performing or low-performing based on whether the percentage of students scoring proficient met the target AMO established by the state. The target AMOs for each subject and grade configuration in the state are presented in Figure 2.

Cross-Sectional with Confidence Interval Approach

To answer research question 2(b), schools were classified as high-performing or low-performing for students with disabilities using the cross-sectional with 95% confidence interval approach. A one sample proportional *z* score was calculated for each school. This *z*-score, currently used by the state, measures the difference between the observed percent proficient in each school and the target AMO, adjusting for the number of students in the subgroup: *z* score = $p - P/\text{square root } P * (1-P)/n$. Within this formula, *P* = Target AMO, *p* = observed percent proficient in the school, and *n* = number of students. At the .05 level the calculated *z* score was compared to $z_{crit} = -1.645$. Schools with a *z* score greater than -1.645 were classified as high-performing, while schools with a *z*-score below -1.645 were classified as low-performing.

Figure 2

State Reading and Mathematics Target AMOs

School Configuration	School Type	Reading	Mathematics
Pre-K – 4	Elementary	43.75%	41.40%
Grades 5-8	Middle	43.00%	19.00%
PreK-12	Special	43.35%	30.68%
Grades 4-12	Special	44.47%	25.10%
Grades 6-12	Alternative	42.95%	19.95%

Three-year Averaging Approach

To answer research question 2(c), each school was classified as high-performing or low-performing for students with disabilities based on whether the average percentage of students with disabilities scoring proficient across three years met the target AMO for the subject and grade established by the state. First, students' reading and mathematics scores in each of the four years was recoded as proficient or nonproficient using the cut scores delineated by CTB McGraw Hill. Data were then aggregated by school to obtain the total number of students with disabilities in each school with valid assessment scores and the number of students with disabilities in each school scoring at or above proficiency.

These data were then used to calculate the average percentage of students scoring at or above proficiency across the three-year period. The three-year average for 2002 was obtained by averaging the percentage of students scoring proficient in 2000, 2001, and 2002. The three-year average for 2003 was obtained by averaging the percentage of students scoring proficient in 2001, 2002, and 2003. Each school was then recoded as high-performing or low-performing for students with disabilities based on whether the percentage of students scoring proficient over the three-year period met the target AMO established by the state (Figure 2).

Value-added Approaches

To classify schools as high-performing or low-performing based on students with disabilities using the value-added approaches, a multi-level model was employed. The individual student data were used at Level 1, and the aggregated data by school were used at Level 2.

First, the fully unconditional model was run to partition the total variance in reading and mathematics gain scores into that attributable to within-school differences and that due to between-school differences. This step was necessary to determine if there was sufficient variance around the mean to warrant a multi-level analysis. Students' reading and mathematics gain scores were set as the dependent variables and no predictors were entered at either the student or school-level.

For the unadjusted value-added approach, students' reading and mathematics gain scores were set as the dependent variables. The servicing school was added at the school level. No covariates were added at the student or school level. The unadjusted value-added approach was defined by the following model:

Level-1 Model: Y = B0 (gainscore) + R

Level-2 Model: B0 = G00 (School) + U0.

For the value-added approach adjusted for student demographics, students' reading and mathematics gain scores were set as the dependent variables. To "adjust" expected outcomes and decrease the variance in outcomes associated with student characteristics, LRE and SES were added as covariates at the student level. Both LRE and SES were grand mean centered and fixed. To examine effects among schools, the percentages of students receiving FARMS and the percentage of students in LRE A were added grand mean centered and fixed at the school level.

SES was entered as a dichotomous variable: 0=never received free/reduced meals (FARMS) and 1=currently/previously received FARMS. SES was chosen because it is accepted as a strong predictor of academic performance and was widely used by states as

a student variable in accountability systems prior to the NCLBA (e.g. Dallas, Minneapolis, Pennsylvania, and North Carolina).

LRE was included as a student level covariate to equate schools based on the severity of students' disabilities. LRE was entered as a dichotomous variable: 0=not LRE A and 1=LRE A. LRE A was chosen because it is an indicator of the academic support needed by students with disabilities. The IDEA states that children are to be educated in the regular education classroom only when "... the nature or severity of the disability of a child is such that education in regular classes with the use of supplementary aids and services cannot be achieved satisfactorily" [34 C.F.R § 300.550(b)(1). The US Department of Education further instructed states to consider the level and extent of services that students with disabilities needed in determining LRE (US Department of Education, 1994). The adjusted value-added approach was defined by the following formula:

Level-1 Model: Y = B0 (Gain Score) + B1*(FARMS) + B2*(LREA) + R

Level-2 Model: B0 = G00 + G01*(PCTFARMS) + G02*(PCTLREA) + U0.

In both the unadjusted and adjusted value-added approaches, the residuals generated at the school level were used to measure the effects or "value-added" of each school (Goldstein, 1991; Raudenbush & Willms, 1995). Although prior research used residuals to numerically rank individual schools (e.g., Clotfelter & Ladd, 1996), current research has questioned the precision of this method and recommends that residuals be used to only to distinguish schools in the top from those in the bottom (Crane, 2002; Raudenbush, 2004; Rubenstein et al., 2004). In the current study, the residuals were used to recode schools into three tercile groups. Schools in the top tercile were classified as

high-performing, schools in the bottom tercile were classified as low-performing, while schools in the middle tercile were classified as average performing.

School Classification Matrix

To display answers to research questions two through five, a matrix developed by Rubenstein et al. (2004) was adapted (Appendix C). The original matrix by Rubenstein et al. displayed only the number of schools that changed rating categories with each approach, but did not allow visual inspection and comparisons of individual schools. The matrix developed by Rubenstein et al. was adapted to allow visual comparisons of school classifications across approaches and years. For each approach, subject area, and year, the randomly generated school numbers were placed on the matrix in the cell for the corresponding classification. This allows for the visual inspection of school ratings across approaches, subjects, and years.

Chapter IV

Analyses and Findings

This purpose of this study was to examine the effects of five different accountability approaches used to evaluate schools based on the performance of students with disabilities in grades, two, four, and six. To accomplish this four years of extant assessment data from the CTBS and CAT/5 mathematics and reading assessments were used to classify schools as high-performing or low-performing based on the performance of students with disabilities.

This chapter presents the findings related to each of the research questions. First, one-way ANOVAs and independent *t*-tests were conducted to determine if the academic performance of students with disabilities differed by socioeconomic status (SES), least restrictive environment (LRE), and disability group. Following this, the reading and mathematics assessment data for students with disabilities were used to classify each school as high performing or low-performing using three status approaches (cross-sectional, cross-sectional with confidence interval, three-year averaging) and two value-added approaches (unadjusted and adjusted for SES and LRE). The characteristics of schools classified as high-performing and low-performing are presented for each approach and subject. In the final section, the reliability of the five approaches in classifying schools was examined within approaches and subject areas, as well as across approaches, subject areas, and years. It should be noted that the ratings assigned to schools using all five approaches are based only on the performance of students with disabilities in the school.

In this chapter results are discussed as follows: performance of students with disabilities, classification of schools using each approach, and the reliability of the five approaches in classifying schools. Within each section the research question is restated followed by the findings. The chapter concludes with a summary of the analyses.

Performance of Students with Disabilities

Research Question 1: Does the mathematics and reading performance of students with disabilities in grades two, four, and six differ by: (a) socioeconomic status, (b) disability group, and (c) LRE?

Using SPSS 13.0 the reading and mathematics composite scale scores and gain scores of students with disabilities, standardized within grade, were analyzed to determine if the performance of students with disabilities differed by socioeconomic status, disability group, and LRE. Independent *t*-tests were conducted to examine differences based on socioeconomic status. Effect sizes, reported as Cohen's *d*, were calculated for groups with significant differences. One-way ANOVAs were conducted to determine if there were differences between groups based on disability and LRE, and post hoc comparisons were conducted on groups with significant overall differences at the .05 level. To examine the consistency of results across years and grades, all analyses were conducted for the total subsamples and separately for each grade.

Socioeconomic Status

In the first analysis the standardized reading and mathematics composite scale scores of students with disabilities were set as the dependent variables and two levels of socioeconomic status (never received FARMS and currently/previously received FARMS) were set as the independent variables. In the second analysis, the standardized

reading and mathematics gain scores of students were the dependent variables and the two levels of socioeconomic status were set as the independent variables. Alpha in all analyses was set at .05.

Reading. The mean reading score of students who had never received FARMS was higher than the mean reading score of students who currently/previously received FARMS in 2002 $t(3,182) = 20.37$, $p = .000$, $d = .72$ and 2003 $t(3,276) = 21.47$, $p = .000$, $d = .74$. Significant differences were found at all grade levels across the target years (Table 12).

No significant differences in reading gain scores based on socioeconomic status were found in 2002 $t(331) = .35$, $p = .727$ (Table 12). In 2003 the mean reading gain score of students who had never received FARMS was higher than the mean gain score of students who currently/previously received FARMS $t(1,329) = 4.24$ $p = .000$, $d = .22$. Across grades in 2003, significant differences in reading gain scores were observed at grade six ($p = .000$), but not at grade four ($p = .070$).

Mathematics. The mean mathematics score of students who had never received FARMS was higher than the mean score of students who currently/previously received FARMs in 2002 $t\ (1,770) = 17.87$, $p = .000$, $d = .86$ and 2003 $t\ (3,270) = 20.93$, $p = .000$, $d = .73$. Significant differences were found at all grade levels across the target years (Table 13).

In 2002 the overall mean mathematics gain score of students who had never received FARMS was higher than the mean gain score of students who currently/previously received FARMS $t(623) = 2.23$, $p = .032$, $d = .18$ (Table 13). Across grades, differences were observed at grade six ($p = .018$), but no differences were

Table 12

Mean Standardized Reading Scores and Gain Scores by Socioeconomic Status and Grade

	2002		2003	
	High SES	Low SES	High SES	Low SES
Composite Scores				
Grade 2	.29**	-.42	.24**	-.38
Grade 4	.28**	-.31	.29**	-.34
Grade 6	.28**	-.36	.45**	-.38
Overall	.32**	-.35	.33**	-.37
Gain Scores				
Grade 4	.00	-.00	.07	-.09
Grade 6	.06	-.05	.14**	-.13
Overall	.02	-.02	.11**	-.12

Note. High SES = students who have never received FARMS; Low SES = students who have previously or currently received FARMS.

* $p<.05$, ** $p<.01$

Table 13

Mean Standardized Mathematics Scores and Gain Scores by Socioeconomic Status and Grade

	2002		2003	
	High SES	Low SES	High SES	Low SES
Composite Scores				
Grade 2	.29**	-.52	.23**	-.37
Grade 4	.28**	-.41	.31**	-.37
Grade 6	.37**	-.56	.42**	-.36
Overall	.31**	-.49	.32**	-.36
Gain Scores				
Grade 4	.03	-.05	.36	-.06
Grade 6	.11*	-.16	.01	-.01
Overall	.07*	-.11	.02	-.03

Note. High SES = students who have never received FARMS; Low SES = students who have previously or currently received FARMS.

* $p<.05$, ** $p<.01$

observed at grade four (p = .464). In 2003 the effect of socioeconomic status on mathematics gain scores was not statistically significant t (803) = 0.73, p = .466.

Disability Group

One-way ANOVAs were conducted to determine if there were differences in the reading and mathematics performance of students with disabilities based on disability group. The standardized reading and mathematics composite scale scores and gain scores were set as the dependent variables, and disability category was set as the independent variable. An alpha level of .05 was used for all analyses. The disability category originally consisted of twelve levels (mentally retarded, hard of hearing, deaf, speech and language impaired, visually impaired, emotionally disturbed, orthopedically impaired, specific learning disabled, multi-handicapped traumatic brain injury, autistic, and other health impaired).[13] The disability categories of speech/language impairment and specific learning disability were not altered. However, the disparate numbers of students in the remaining categories was problematic. In order to have cell sizes that were more balanced, the remaining disability categories were collapsed into two groups.

The disability categories of mentally retarded, multi-handicapped, traumatic brain injury, and autistic were combined into one group and labeled "intellectual impairment." The disability categories of hard of hearing, deaf, visually impaired, emotionally disturbed, orthopedically impaired, and other health impaired were collapsed into one group and labeled "sensory/emotional impairment." The disability groups were collapsed based on based on the intellectual and academic placement criteria for each disability category and the performance of students in each category on the CTBS and CAT/5. It is recognized that the academic and intellectual abilities of students within the collapsed

groups varies, particularly for students with emotional impairments. However, the extremely small number of students in the individual disability categories prevented them from being analyzed separately in this study.

Reading. The effect of disability group on reading performance was statistically significant in 2002 F (3, 3,180) = 36.20, p =.000 and 2003 F (3, 3,274) = 14.62, p =.000. Across target years and grade levels, students with speech/language impairments scored higher in reading than students with intellectual impairments (Table 14). Although students with specific learning disabilities and sensory/emotional impairments scored higher overall than students with intellectual impairments across target years, significant differences were not observed at all grade levels.

The effect of disability group on reading gain scores was statistically significant in 2002 F (3, 329) = 5.61, p=.001 and 2003 F (3, 1,327) = 7.14, p=.000. However, no significant differences in reading gain scores between disability groups were found across years and grade levels (Table 15). In both years the average reading gain score of students with speech/language impairments was higher than the average gain score of students with intellectual impairments, but significant differences were not found at all grade levels.

Mathematics. The effect of disability group on mathematics achievement was statistically significant in 2002 F (3, 1,768) = 21.04, p = .000 and 2003 F (3, 3,268) =17.02, p = .001. However, no significant differences were consistently found across years and grade levels (Table 16). In both years the overall mean mathematics scores of students with speech/language impairments, specific learning

Table 14

Mean Standardized Reading Scores by Disability Category

	2002				2003			
	Overall	Grade 2	Grade 4	Grade 6	Overall	Grade 2	Grade 4	Grade 6
Specific Learning Disabled	0.62[a]	-0.14	0.10 [a]	0.101 [a]	0.03[a]	0.08 [a]	0.05[a]	0.00 [a]
Speech/Language Impairment	0.02 [a]	0.05[a]	0.11 [a]	-0.041 [a]	0.04[a]	0.04 [a]	0.03[a]	0.03 [a]
Sensory/Emotional Impairment	0.25 [a, b, c]	0.12[a]	0.22 [a]	0.29 [a, b, c]	0.03[a]	-0.03	-0.03[a]	0.11 [a]
Intellectual Impairment	-0.56	-0.39	-0.74	-0.470	-0.44	-0.43	-0.42	-0.46

Note. Differences significant at the .05 level. a=higher than intellectual impairment, b = higher than speech/language impairment, c = higher than specific learning disability, d = higher than sensory/emotional impairment

Table 15

Mean Standardized Reading Gain Scores by Disability Category

	2002			2003		
	Overall	Grade 4	Grade 6	Overall	Grade 4	Grade 6
Specific Learning Disabled	0.07 [a]	0.10[a]	0.00[a]	-0.10	-0.04	-0.12
Speech/Language Impairment	0.04 [a]	0.01[a]	-0.04[a]	0.15 [a, c, d]	0.09	0.22 [c,d]
Sensory/Emotional Impairment	0.10 [a]	0.22[a]	0.29[a]	-0.08	-0.10	-0.06
Intellectual Impairment	-0.89	-0.74	-0.47	-0.13	-0.26	0.00

Note. Differences significant at the .05 level. a=higher than intellectual impairment, b = higher than speech/language impairment, c = higher than specific learning disability, d = higher than sensory/emotional impairment

Table 16

Mean Standardized Mathematics Scores by Disability Category

	2002				2003			
	Overall	Grade 2	Grade 4	Grade 6	Overall	Grade 2	Grade 4	Grade 6
Specific Learning Disabled	-0.093 [a]	-0.27	-0.06 [a]	-0.07	-0.05 [a]	-0.06 [a]	-.09	-.01 [a]
Speech/Language Impairment	0.077 [a, c]	0.08 [a, c]	0.09 [a]	0.04	0.09 [a, c]	0.08 [a]	.08 [c]	.12 [a]
Sensory/Emotional Impairment	0.129 [a, b]	0.01 [a]	0.17 [a]	0.18 [a]	0.02 [a]	-0.07 [a]	.06	.02 [a]
Intellectual Impairment	-0.808	-0.70	-0.91	-0.77	-0.41	-0.52	-.13	-.62

Note. Differences significant at the .05 level. a=higher than intellectual impairment, b = higher than speech/language impairment, c = higher than specific learning disability, d = higher than sensory/emotional impairment

Table 17

Mean Standardized Mathematics Gain Scores by Disability Category

	2002			2003		
	Overall	Grade 4	Grade 6	Overall	Grade 4	Grade 6
Specific Learning Disabled	-0.11	-0.22	-0.07	0.03	-0.02	0.06
Speech/Language Impairment	0.14 [a, c]	0.16[a]	0.12	-0.02	0.00	-0.03
Sensory/Emotional Impairment	-0.06	-0.11	-0.03	-0.08	-0.14	-0.05
Intellectual Impairment	-0.51	-0.48	-0.58	0.14	0.25	-0.11

Note. Differences significant at the .05 level. a=higher than intellectual impairment, b = higher than speech/language impairment, c = higher than specific learning disability, d = higher than sensory/emotional impairment

disabilities, and sensory/emotional impairments were higher than the mean scores of students with intellectual impairments; but significant differences were not observed at all grade levels. Students with speech/language impairments also demonstrated higher overall mathematics scores than students with specific learning disabilities across target years, but differences were not observed at all grade levels.

The effect of disability group on mathematics gain scores was statistically significant in 2002 F (3, 621) = 6.51, p = .000. The average gain score of students with speech/language impairments was higher than the average gain scores of students with intellectual impairments and specific learning disabilities, but significant differences were not found at all grade levels (Table 17). In 2003 the effect of disability group on mathematics gain scores was not statistically significant, F (3, 801) = .724, p = .538.

Least Restrictive Environment

To determine if there were differences in the reading and mathematics performance of students with disabilities based on the amount of time they spend in general education classrooms each day, one-way ANOVAs were conducted. The standardized reading and mathematics composite scores and gain scores were set as the dependent variables and three levels of LRE (LRE A, LRE B, LRE C/special schools) were set as the independent variables. The variable LRE originally consisted of five levels. However, the small number of students with valid assessment scores in LRE C, F, and G was problematic in the one-way ANOVAs. In order to have cell sizes that were more balanced, students in LRE C, LRE F, and LRE G, who spend more than 80% of their instructional time outside the general education classroom, were collapsed into one category and renamed "LRE C/special schools."

Reading. The effect of LRE on reading performance was statistically significant in 2002 F (2, 3,181) = 207.58, p = 000 and 2003 F (2, 3,275) = 176.57, p =.000. Students in LRE A had higher overall reading scores than students in LRE C/special schools and at all grade levels across target years (Table 18). Significant differences in overall reading scores were also found between students in LRE A and LRE B, and between students in LRE B and LRE C/special schools across years, but differences were not observed at all grade levels across the two years .

The effect of LRE on reading gain scores was statistically significant in 2002 F (2, 330) = 6.24, p =.002 and 2003 F (2, 1,328) = 19.58, p=.000. However, no significant differences in reading gain scores were found across the target years and grade levels (Table 19). The gain scores for students in LRE A and LRE B were higher than the gain scores of students in LRE C/special schools at all grade levels in 2003, but no significant differences were found between these groups in 2002.

Mathematics. The effect of LRE on mathematics scores was statistically significant in 2002 F (2, 1,170) = 118.62, p =.000 and 2003 F (2, 3,269) = 174.87, p =.000. Across the target years and grade levels, the mean mathematics scores of students in LRE A were higher than the mean scores of students in LRE C/special schools (Table 20). Significant overall differences were also observed between students in LRE A and LRE B and between students in LRE B and LRE C/special schools across the target years, but differences between these groups were not found across all grades.

The effect of LRE on mathematics gain scores was statistically significant in 2002 F (2, 622) = 7.98, p =.000, and 2003 F (2, 802) = 11.65, p = .000. However, no

Table 18

Mean Standardized Reading Scores by LRE

	2002				2003			
	Overall	Grade 2	Grade 4	Grade 6	Overall	Grade 2	Grade 4	Grade 6
LRE A	0.33[a,b]	0.22[b]	0.30[a, b]	0.50[a, b]	0.26[a, b]	0.20[a, b]	0.25[a, b]	0.33[a, b]
LRE B	-0.15[b]	0.02[b]	-0.33	-0.10[b]	-0.13[b]	-0.19	-0.12[b]	-0.13[b]
LRE C/Special Schools	-0.43	-0.52	-0.44	-0.39	-0.46	-0.48	-0.47	-0.43

Note. Differences significant at the .05 level. a= higher than LRE B, b = higher than LRE C/special schools

Table 19

Mean Standardized Reading Gain Scores by LRE

	2002			2003		
	Overall	Grade 4	Grade 6	Overall	Grade 4	Grade 6
LRE A	0.12[a]	0.09 [a]	0.19	0.16[b]	0.16[b]	0.15 [b]
LRE B	-0.40	-0.60	-0.25	0.02[b]	0.07[b]	0.01[b]
LRE C/Special Schools	-0.13	-0.06	-0.23	-0.23	-0.26	-0.21

Note. Differences significant at the .05 level. a= higher than LRE B, b = higher than LRE C/special schools

Table 20

Mean Standardized Mathematics Scores by LRE

	2002				2003			
	Overall	Grade 2	Grade 4	Grade 6	Overall	Grade 2	Grade 4	Grade 6
LRE A	.22[a,b,]	.19[b]	.19[a,b]	.30[a,b]	.26[a,b]	.24[a,b]	.22[a,b]	.33[a,b]
LRE B	-.26[b]	-.07[b]	-.43	-.26	-.13[b]	-.10[b]	-.14[b]	-.13[b]
LRE C/Special Schools	-.73	-.87	-.78	-.53	-.46	-.62	-.40	-.41

Note. Differences significant at the .05 level. a= higher than LRE B b = higher than LRE C/special schools

Table 21

Mean Standardized Mathematics Gain Scores by LRE

	2002			2003		
	Overall	Grade 4	Grade 6	Overall	Grade 4	Grade 6
LRE A	.11[b]	.11[b]	.11	-.14	-.16	-.12
LRE B	-.11	.01[b]	-.17	.26[c]	.27[c]	.25[c]
LRE C/Special Schools	-.31	-.57	-.10	.15[c]	.22[c]	.06

Note. Differences significant at the .05 level. a= higher than LRE B b = higher than LRE C/special schools c = higher than LRE A

significant differences were found across years and grade levels (Table 21). The mean mathematics gain scores for students in LRE B were higher than the mean gain scores for students in LRE A across all grade levels in 2003, but no differences between these groups were found in 2002.

School Classifications Using Status Approaches

Research Question 2: What are the characteristics of schools labeled high-performing and low-performing for students with disabilities using the three status approaches: (a) cross-sectional, (b) cross-sectional with 95% confidence interval, and (c) three-year rolling average? To answer research questions 2(a-c), students' reading and mathematics scores were first recoded as proficient or nonproficient using the cut scores delineated by the test developer (defined in Chapter 3). The total number of students with valid assessment scores and number of students scoring proficient in each subject area were then aggregated by school to calculate the percentage of students in each school scoring proficient. Student demographic information was also aggregated by school to obtain school characteristics including the number and percentage of: students who had received FARMs, students with valid assessment scores, minority students, and students being served in LRE A.

The aggregated reading and mathematics achievement data were then used to classify schools as high or low-performing based on the classification criteria for each approach. The demographic characteristics of schools classified as high-performing and low-performing in reading and mathematics using each of the status approaches is presented in Tables 22 and 23.

One hundred sixty-four schools had students that participated in the reading and mathematics assessments in 2002: 125 elementary; 35 middle; and 4 special schools. In 2003 a new middle school (#160) opened in the district, increasing the total number of schools to 165. When reporting the characteristics of schools classified using each approach, individual schools are referenced using the randomly generated school number. The classifications of each school by subject area and accountability approach are presented in Appendix C.

Cross-Sectional Approach.

To answer research question 2a, schools were classified as high-performing or low-performing based on whether the percentage of students with disabilities scoring proficient in each school was above or below the target AMO set by the state for the subject area and school configuration (Figure 2). Schools where the average percentage of students with disabilities scoring proficient fell at or above the target AMO were classified as high-performing, while schools where the average percentage scoring proficient was below the target AMO were classified as low-performing. The characteristics of schools rated high-performing and low-performing in reading and mathematics using the cross-sectional approach are presented in Tables 22 and 23.

Reading. Using the cross-sectional approach, 8% (*n*=13) of schools in 2002 and 9% (*n*=15) of schools in 2003 were classified as high-performing in reading for students with disabilities (Table 22). In 2002 all thirteen schools classified as high-performing were elementary schools, and 93% *(n*=14) of the 15 schools classified as high-performing in 2003 were elementary schools. In both target years two special schools did not have any students with valid reading scores and thus were not rated.

Across the target years the average percentage of students who had received FARMS was 30%-40% lower in the schools classified as high-performing schools than in the schools rated as low-performing. Only one school classified as high-performing each year fell in the top two quartiles based on the percentage of students with disabilities who had received FARMS, compared to 55% (n=81) of schools classified as low-performing.

The schools classified as high-performing using the cross-sectional approach averaged 10 and 18 students with valid reading scores in the target years, compared to an average of 22 and 29 students in the low-performing schools. Both the high-performing and low-performing schools averaged valid reading scores for approximately 90% of their students. However, one high-performing school in 2002 and two high-performing schools in 2003 had valid reading scores for 50% or less of their students with disabilities. In 2002 school #211 had valid reading scores for only 29% of their students. In 2003 schools #166 and #211 had valid reading scores for 40% and 50% of their students respectively. None of the schools classified as low-performing had valid reading scores for less than 50% of their students in either year.

Mathematics. Using the cross-sectional approach, 9% *(n*=15) of schools in 2002 and 5% (*n*=7) of schools in 2003 were classified as high-performing in mathematics for students with disabilities (Table 23). Fifty-three percent (*n*=8) of those schools rated high-performing in 2002 and all seven schools in 2003 were elementary schools. In each target year two special schools did not have students with valid mathematics assessment scores and thus were not rated.

Across the target years the high-performing schools averaged 20% more students who had received FARMS than the low-performing schools, and the majority of schools

classified as high-performing fell in the bottom two quartiles for the percentage of students with disabilities who had received FARMS. Of the 15 schools rated as high-performing in 2002, 13% (*n*=2) fell in the top two quartiles for the percentage of students who had received FARMs. In 2003 14% (*n*=1) of the seven schools fell in the top two quartiles for the percentage of students who had received FARMs. In contrast, 50%-55% of schools classified as low-performing in mathematics fell in the top two quartiles for the percentage of students with disabilities who had received FARMS.

The high-performing schools had valid mathematics scores for an average of 45% of students with disabilities in 2002 and 72% in 2003, compared to 62% and 93% in the low-performing schools in the two years respectively. Two elementary schools that were rated as high-performing in both years had valid mathematics scores for less than 50% of their students with disabilities. School #166 had valid mathematics scores for 50% and 44% of students in the two years respectively, and school #211 had valid mathematics test scores for 29% and 40% of students in the two years. No schools classified as low-performing in either year had valid mathematics scores for less than 50% of students.

Cross-Sectional with Confidence Interval Approach

To answer research question 2(b), a 95% confidence interval based on the number of students in each school was established around the percentage of students scoring proficient. Schools where the calculated confidence interval included the target AMO established by the state were classified as high-performing. Schools where the target AMO fell below the calculated confidence interval were classified as low-performing. The characteristics of schools classified as high-performing and low-performing for

Table 22

Characteristics of Schools Classified as High-performing and Low-performing in Reading Using Status Approaches

	# Schools		% Minority		% FARMs		% Valid Test Scores		% LRE A	
	2002 (N=164)	2003 (N=165)	2002	2003	2002	2003	2002	2003	2002	2003
Cross-sectional										
High-performing	13	15	16.1	26.2	14.6	19.6	92.6	88.9	91.0	70.4
Low-performing	149	148	46.5	58.4	54.5	48.8	91.7	93.0	53.1	60.3
Cross-sectional w/CI										
High-performing	57	53	29.9	51.7	29.4	27.5	90.5	95.3	69.0	69,.0
Low-performing	105	110	38.7	63.5	52.8	55.1	92.5	91.4	50.0	50.0
Three-year Averaging										
High-performing	12	10	13.9	12.9	9.2	9.2	74.4	91.0	83.2	84.7
Low-performing	150	153	46.5	45.8	47.0	47.0	70.2	93.3	52.1	52.7

Note. Two schools did not have valid assessment scores for any students either year and were not rated. CI = confidence interval.

Table 23

Characteristics of Schools Classified as High-performing and Low-performing in Mathematics Using Three Approaches

	# Schools		% Minority		% FARMS		% Valid Test Scores		% LRE A	
	2002 (N=164)	2003 (N=165)	2002	2003	2002	2003	2002	2003	2002	2003
Cross-sectional										
High-performing	15	7	24.7	44.9	24.0	27.0	45.0	72.0	71.0	85.0
Low-performing	147	155	46.0	55.9	47.0	46.0	62.0	93.2	71.0	60.0
Cross-sectional w/CI										
High-performing	98	53	37.6	42.4	38.5	30.0	53.1	91.3	72.0	72.0
Low-performing	66	110	54.0	61.7	53.9	53.8	60.0	93.2	72.0	56.0
Three-year Averaging										
High-performing	2	2	16.0	15.8	23.3	31.0	44.4	38.0	88.8	82.4
Low-performing	160	161	44.4	44.1	44.8	45.4	63.0	65.0	68.3	73.5

Note. Two schools did not have valid assessment scores for any students either year and were not rated. CI = confidence interval.

students with disabilities in reading and mathematics using the cross-sectional with confidence interval approach are presented in Tables 22 and 23.

Reading. Thirty-two percent *(n*=57) of schools in 2002 and 32% *(n*=53) in 2003 were classified as high-performing for students with disabilities in reading using the cross-sectional with 95% confidence interval approach (Table 22). All 57 of the schools classified as high-performing in 2002 were elementary schools, and 98% (*n*=52) of the 53 schools classified as high-performing in 2003 were elementary schools. Two special schools did not have any students with valid test scores and thus were not rated.

Across the target years the average percentage of students who had received FARMS was 20%-25% higher in the low-performing schools than the high-performing schools. Thirty percent (*n*=17) of the high-performing schools in 2002 and 19% (*n*=10) in 2003 fell in the top two quartiles for the percentage of students who had received FARMS. In comparison, 62% (*n*=65) and 65% *(n*=71) of the schools classified as low-performing in the two years respectively fell in the top two quartiles for the percentage of students with disabilities who had received FARMS.

The high-performing schools averaged 13 and 16 students with valid reading scores in the two years, compared to an average of 22 and 23 students in the low-performing schools. The average percentage of students with valid reading scores was 90% or greater in both the high-performing and low-performing schools. However, two schools classified as high-performing in 2002 and 2003 (#211, #166, discussed previously) had valid reading scores for less than 50% of their students with disabilities.

The range of students with disabilities scoring proficient in reading was 32%-100% in the high-performing schools, compared to 10%-36% in the low-performing

schools. Two schools classified as high-performing for students with disabilities in reading in 2002 (#166 and #7) and one school in 2003 (#217) did not have any students with disabilities scoring proficient in reading using the cross-sectional with confidence interval approach.

Mathematics. Using the cross-sectional with confidence interval approach, 60% (*n*=98) of schools in 2002 and 32% *(n*=53) in 2003 were classified as high-performing for students with disabilities in mathematics (Table 23). Sixty-eight percent (*n*=67) of the schools classified as high-performing in 2002 and 87% (*n*=46) in 2003 were elementary schools. Two special schools each year did not have students with valid test scores in mathematics and thus were not rated.

The average percentage of students with disabilities who had received FARMS was 15%-25% lower in the schools rated as high-performing than in the schools rated as low-performing. The majority of schools classified as high-performing in both years fell in the bottom two quartiles for the percentage of students who had received FARMS. Forty percent (*n*=40) of the schools classified as high-performing in 2002 and 23% (*n*=12) in 2003 fell in the top two quartiles for the percentage of students with disabilities who had received FARMS. In comparison, 62%-66% of schools classified as low-performing fell in the top two quartiles for the percentage of students with disabilities who had received FARMS using the cross-sectional with confidence interval approach.

Using the cross-sectional with confidence interval approach, the percentage of students with disabilities scoring proficient in mathematics ranged from 0%-100% in the schools rated as high-performing and 0%-24% in the schools rated as low-performing

schools. Of these, 40 schools classified as high-performing in mathematics in 2002 and ten schools in 2003 had a proficiency rate of 20% or less for students with disabilities.

Three-year Rolling Average Approach

To answer research question 2(c), assessment and demographic data for each school were averaged over three years. Data from 2000, 2001, and 2002 were aggregated to rate schools in 2002; and data from 2001, 2002, and 2003 were used to rate schools in 2003. Schools were classified as high performing or low-performing for students with disabilities based on whether the average percentage of students proficient over the three years met the target AMOs established by the state for the subject and grade. The characteristics of schools rated as high-performing and low-performing for students with disabilities in reading and mathematics using the three-year averaging approach are presented in Tables 22 and 23.

Reading. Using the three-year averaging approach, 7% (*n*=12) of schools in 2002 and 6% (*n*=10) of schools in 2003 were classified as high-performing for students with disabilities in reading (Table 22). Across target years all schools classified as high-performing were elementary schools. Two special schools did not have students with valid test scores either year and thus were not rated.

Across the target years the schools classified as high-performing averaged approximately 35% fewer students with disabilities who had received FARMS than the schools classified as low-performing. None of the schools classified as high-performing in reading using the three-year averaging approach fell in the top two quartiles for the percentage of students who had received FARMS. The average percentage of students with disabilities who were minority and were being served in LRE A was also

approximately 30% higher in the schools classified as high-performing than in the schools rated as low-performing.

The average percentage of students with valid reading scores in the high-performing and low-performing schools differed by 5% or less in each of the two years. The schools rated as high-performing averaged nine students with disabilities with valid reading scores each year, compared to 15 and 20 students in the low-performing schools.

Mathematics. Using the three year averaging approach, 1% (*n*=2) of schools in each target year were classified as high performing in mathematics for students with disabilities. In each target year both schools classified as high-performing were elementary schools.

The average percentage of students with disabilities who had received FARMS was 20%-25% lower in the high-performing schools than in the low-performing schools. Across target years both schools classified as high-performing fell in the bottom two quartiles for the percentage of students with disabilities who had received FARMS. The percentage of minority students was approximately 30% lower in the high-performing schools than in the low-performing schools, while the percentage of students being served in LRE A was approximately 20% higher in the high-performing schools.

The average percentage of students with disabilities with valid mathematics scores across the three-year period in the low-performing schools was 60%-65%, compared to 31% and 44% in the high-performing schools. In each target year, one of the two schools rated as high-performing had valid mathematics assessment scores for fewer than 50% of their students with disabilities. In 2002 school #211 had valid mathematics scores for 37% of students with disabilities, and in 2003 school #166 had

valid mathematics scores for 33% of students with disabilities. In comparison, approximately 15% of the schools classified as low-performing each year had valid mathematics scores for 50% or less of students with disabilities.

School Classifications Using Value-added Approaches

Research Question 3*:* Are value-added approaches practical for rating schools based on the subgroup of students with disabilities? If so, what are the characteristics of schools labeled high-performing and low-performing for students with disabilities using the (a) value-added approach unadjusted for student demographics, and (b) value-added approach adjusted for students' SES and LRE?

Only those students with disabilities who had valid reading and mathematics gain scores were included in the unadjusted and adjusted value-added approaches. The gain scores for 2002 were calculated as the difference in scale scores between 2002 and 2000, and the gain scores for 2003 were calculated as the difference in scale scores between 2003 and 2001. Gain scores were standardized within grade. Students who were in second grade and did not participate in the assessments the prior years; had invalid prior achievement scores due to accommodations; or were not in the special education database in the previous year were excluded from the value-added approaches (Table 25). In reading, the final matched sample included 11% (*n* =333) of the sample in 2002 and 41% (*n* =1,331) in 2003. In mathematics the final matched sample included 35% (*n*=625) of the sample in 2002 and 25% (*n*=805) in 2003.

At the school level, 52% (*n*=86) of schools in 2002 and 86% *(n* =142) in 2003 had a minimum of two students with disabilities with valid reading assessment scores and

Table 24

Students Excluded From Value-Added Analyses

	Reading				Mathematics			
	2002 (N = 3,184)		2003 (N=3,278)		2002 (N=1,773)		2003 (N=3,272)	
	n	%	*n*	%	*n*	%	*n*	%
Second grade	768	24.1	826	25.2	605	34.1	802	24.5
Prior Achievement Score Invalid due to Accommodations	1,071	33.6	59	1.8	25	1.4	681	20.8
Not in special education database	1,012	31.8	1,062	32.4	518	29.2	984	30.1
Total Included in Value-added Models	333	10.5	1,331	40.6	625	35.2	805	24.6

Note. N = total number of students with valid assessment scores.

were included in the value-added analyses. Sixty-six percent (*n*=109) of schools in 2002 and 78% (*n*=129) of schools in 2003 had a minimum of two students with valid mathematics assessment scores and were included in the value-added analyses. The intraclass correlation (ICC) for reading was .212 in 2002 and .269 in 2003. The ICC for mathematics was .318 in 2002 and .404 in 2003. This means that 21% to 27% of the variance in reading gain scores and 32% to 40% of the variance in mathematics gain scores may be attributed to the school a student attends and that there is sufficient variance around the mean to warrant multi-level analysis

Unadjusted Value-added Approach

Reading. Thirty-four percent (*n*= 29) of rated schools in 2002 and 33% (*n*=47) of rated schools in 2003 were classified as high-performing in reading for students with disabilities using the unadjusted value-added approach (Table 25). Ninety-three percent (*n*=27) of schools classified as high-performing in 2002 and all forty-seven schools classified as high-performing in 2003 were elementary schools.

The percentage of students with disabilities who were minorities or who had received FARMS in the high-performing and low-performing schools differed by ten percentage points or less in the target years. The high-performing and low-performing schools were also evenly distributed in regard to SES, with approximately 50% of the schools in each category in the bottom two quartiles and 50% in the top two quartiles for the percentage of students with disabilities who had received FARMS.

The percentage of students with disabilities being served in LRE A was 15%-29% higher in the high-performing schools than in the low-performing schools across target years. The number of students with disabilities who had valid reading scores in the high-

performing and low-performing schools differed by one in 2002, but differed by 18 in 2003.

Mathematics. Thirty-four percent (*n*=37) of rated schools in 2002 and 2003 (*n*=43) were classified as high-performing in mathematics for students with disabilities using the unadjusted value-added approach (Table 26). Ninety-seven percent (*n*=36) of the 37 schools rated as high-performing in 2002 and all 43 schools in 2003 were elementary schools.

The percentage of students with disabilities who were minority students and or who had received FARMS in the high-performing and low-performing schools differed by ten percentage points or less in the target years. The high-performing and low-performing schools were also evenly distributed in regard to SES, with approximately half of the students in each category in the bottom two quartiles and half in the top two quartiles for the percentage of students who had received FARMS.

The percentage of students being served in LRE A averaged 10%-20% higher in the high-performing schools than in the low-performing schools. The high-performing schools also averaged fewer students with disabilities with valid mathematics scores than the low-performing students, with an average of four students in each of the target years, compared to eight and ten students in the low-performing schools.

Adjusted Value-added Approach

To equate or adjust schools based on the characteristics of students, the LRE and SES of individual students were added as student level covariates in the adjusted value-added approach. The aggregated percentage of students in LRE A and the percentage of students who had received FARMS were also added as covariates at the school level.

Table 25

Characteristics of Schools Classified High-performing and Low-performing in Reading Using Value-Added Approaches

	# Schools		% Minority		% Low SES		% LRE A		Average # Students with Valid Scores	
	2002 N=86	2003 N=142	2002	2003	2002	2003	2002	2003	2002	2003
Unadjusted Value-added										
High-performing	29	47	44.1	50.4	47.7	41.5	74.4	56.8	3	5
Average Performing	29	48	44.7	46.1	40.2	41.2	69.4	54.3	3	5
Low-performing	28	47	41.1	41.3	45.1	52.8	54.7	40.8	4	18
Adjusted Value-added										
High-performing	28	47	41.6	53.9	45.1	49.5	66.7	49.2	3	5
Average-performing	30	49	44.2	54.0	38.6	42.5	73.3	54.2	3	6
Low-performing	28	46	44.5	54.0	49.5	43.6	58.4	48.3	4	18

Note. N = total number of schools rated in reading using the value-added approaches.

Table 26

Characteristics of Schools Classified High-performing and Low-performing in Mathematics Using Value-Added Approaches

	# Schools		% Minority		% Low SES		% in LRE A		Average# Students	
	2002 N=109	2003 N=128	2002	2003	2002	2003	2002	2003	2002	2003
Unadjusted Value-added										
High-performing	37	43	38.2	57.5	36.8	34.6	78.8	55.2	4	4
Average-performing	36	43	44.6	53.4	38.9	41.7	72.2	64.2	5	4
Low-performing	36	42	46.3	52.0	43.2	45.0	55.8	64.9	8	10
Adjusted Value-added										
High-performing	36	43	41.0	55.1	43.7	40.6	65.4	61.0	4	4
Average-performing	37	43	42.4	51.9	36.8	37.8	79.6	61.4	5	5
Low-performing	36	42	45.6	55.8	38.3	40.6	61.8	61.8	8	10

Note. N = total number of schools rated in mathematics using the value-added approaches.

The characteristics of schools rated as high-performing and low-performing for students with disabilities in reading and mathematics using the adjusted value-added approach are presented in Tables 25 and 26.

Reading. Thirty-three percent (*n*=28) of rated schools in 2002 and 2003 (*n*=47) were classified as high-performing in reading for students with disabilities using the adjusted value-added approach (Table 25). Ninety-six percent (*n*=27) of the 28 schools rated as high-performing in 2002 and all 47 schools in 2003 were elementary schools. In reading, the covariates of LRE and socioeconomic status reduced the variance between schools from 21% to 19% in 2002 and from 27% to 26% in 2003. In both years LRE made a significant ($p<.01$) contribution to reading gain scores, but SES was significant only in 2003 ($p=.020$). Reliability for reading gain scores was .431 in 2002 and .661 in 2003.

Using the adjusted value-added approach, the high-performing and low-performing schools differed by less than ten percentage points in regard to the average percentage of students with disabilities who were minorities or who minority students, students who had received FARMS, and students being served in LRE A.

Mathematics. Thirty-three percent (*n*=36) of rated schools in 2002 and 34% (*n*=43) of rated schools in 2003 were classified as high-performing in mathematics for students with disabilities using the adjusted value-added approach (Table 26). Of the schools classified as high-performing, thirty-six in 2002, 97% (*n*=35) and 100% (*n*=43) were elementary schools.

The addition of LRE and SES reduced the variance in mathematics from 32% to 28% in 2002. In 2003 the addition of LRE and socioeconomic status increased the

variance between schools from 40% to 43%. LRE was a significant contributor to mathematics gain scores ($p<.01$) in both years, but socioeconomic status was not significant in either year. Reliability was .621 in 2002 and .760 in 2003.

Across target years, the high-performing and low-performing schools differed by 6% or less in the average percentage of students with disabilities who were minorities, who were being served in LRE A, and had received FARMS.

Reliability of School Classifications

Research Question 4*:* Across target years what is the reliability of the five approaches in classifying schools (a) within approaches and subject areas, and (b) across approaches and subject areas? To examine the reliability of the five approaches in classifying schools, the number and percentage of schools consistently classified as high-performing or low-performing and the number and percentage of schools that changed classifications in reading and in mathematics across target years was calculated. To allow for visual inspection of school ratings across approaches, subject areas, and years, the random school numbers were placed on the classification matrix in the corresponding category (Appendix C).

Across target years, three-year averaging was the most reliable of the approaches in classifying schools, with less than 3% of schools changing classifications in reading and mathematics. In reading the unadjusted value-added approach was the most unreliable approach in classifying schools, with 55% of schools changing ratings between the two years using this approach. In mathematics the cross-sectional with confidence interval approach was the most unreliable of the approaches, with approximately 46% of schools changing classifications between the target years using this approach.

No schools were rated as high-performing in reading for students with disabilities across the target years using all five approaches (Table 27). Two schools (#211 and #269) were rated as high-performing both years using the cross-sectional, cross-sectional with confidence interval, and three-year averaging approaches. However, these schools were not rated using the two value-added approaches. School #114 was rated as high-performing using the cross-sectional, cross-sectional with confidence interval, and three-year averaging approaches in both years, and the value-added approaches in 2002. However, school #114 was not rated using the value-added approaches in 2003.

Thirty-five percent (*n*=56) of schools were rated low-performing in reading for students with disabilities using all five approaches across target years. Sixty-one percent *(n*=34) of those consistently classified as low-performing in reading were middle schools, 36% (n=20) were elementary, and 4% (*n*=2) were special schools.

In mathematics no schools were rated as high-performing for students with disabilities across the target years using all five approaches. School #211 was classified as high-performing in mathematics across the target years using the cross-sectional, cross-sectional with confidence interval, and three-year averaging approaches. However, this school was not rated using the two value-added approaches.

Sixteen percent (*n*=27) of schools were classified as low-performing in mathematics for students with disabilities across the target years using all five approaches. Of those consistently classified as low-performing in mathematics, 75% (*n*=20) were elementary schools, 22% (*n*=6) were middle schools, and the remaining 4% (*n*=1) were special schools.

Table 27

Schools Most Frequently Classified as High-performing in Reading and Mathematics

			% FARMS		Reading			Mathematics		
Random School Number	School Type	Total Rated High-performing[a]	2002	2003	Total Rated High-performing	Total Rated[b]	% Rated High-performing	Total Rated High-performing	Total Rated[b]	% Rated High-performing
45	Elementary	12	0.0	0.0	6	8	75.0	6	8	75.0
74	Elementary	12	8.3	0.0	7	10	70.0	5	8	63.0
77	Elementary	12	47.8	38.1	6	10	60.0	6	10	60.0
211	Elementary	12	28.6	40.0	6	6	100.0	6	6	100.0
38	Elementary	10	15.8	9.1	4	10	40.0	6	10	60.0
114	Elementary	10	14.3	22.2	8	8	100.0	2	8	25.0
81	Elementary	9	36.8	44.0	4	10	40.0	5	10	50.0
92	Elementary	9	0.0	0.0	4	10	40.0	5	10	50.0
112	Elementary	9	12.5	10.5	6	8	75.0	3	10	30.0
135	Elementary	9	10.7	0.0	7	8	88.0	2	10	20.0
182	Elementary	9	0.0	20.0	5	8	63.0	4	8	50.0
269	Elementary	9	0.0	22.2	6	6	100.0	3	6	50.0

a =Total number of times classified as high-performing in reading and mathematics (Total possible = 20), b=Total number of times rated (Total possible = 10)

Chapter V

Discussion

The purpose of this study was to examine the mathematics and reading performance of students with disabilities and to compare the effects of five different accountability approaches in evaluating schools based on this subgroup. The five accountability approaches included three status approaches that measured the level at which students were performing and two value-added approaches that measured the change in students' academic achievement. To accomplish this four years of large-scale reading and mathematics assessment data for students with disabilities in grades two, four, and six from a large school district in the mid-Atlantic were examined.

This chapter includes a summary of the major findings, discussion and implications of the findings for policy, and recommendations for future research. This chapter is divided into the following sections: (a) discussion of the performance of students with disabilities in reading and mathematics, (b) discussion of the classifications of schools using each of the accountability approaches, (c) discussion of the implications of findings for policy and future research.

Performance of Students with Disabilities

Significant differences were found in the overall reading and mathematics performance of students with disabilities based on socioeconomic status. In both years and at all grade levels, students who had received FARMS demonstrated significantly lower reading and mathematics achievement than students who had not received FARMs. However, no significant differences in reading and mathematics gain scores based on socioeconomic status were consistently found across subjects and years. Students who

had received FARMS overall had higher reading and mathematics gain scores than students who had not received FARMS in only one of the two years, with small to insignificant differences evident in the remaining year. Differences in gain scores based on socioeconomic status were also not consistent across grade levels.

These findings parallel research in the general population which has consistently found a negative relationship between poverty and overall academic achievement (Education Commission of the States, 2003; Raudenbush, 2004), but no relationship between socioeconomic status and students' academic gains (Raudenbush). These findings suggest that although there is a negative effect for poverty on all students, students can make equal gains regardless of socioeconomic status with proper academic instruction. The findings also support Sanders' theory that it is not necessary to control for socioeconomic status in approaches that measure change (Sanders et al., 1997).

No significant differences in the reading and mathematics performance of students with disabilities based on disability category were consistently found across target years and subject areas. In reading students with speech/language impairments scored significantly higher than students with intellectual impairments across target years and grade levels; but significant differences between these groups were not consistently found across target years and grade levels in mathematics. Although students with specific learning disabilities and sensory/emotional impairments scored significantly higher overall in reading than students with intellectual impairments, differences between these groups were not observed in second grade. In mathematics students with speech/language impairments, specific learning disabilities, and sensory/emotional

impairments scored higher overall than students with intellectual impairments, but differences were not observed across all grade levels.

In terms of gain scores, students with speech/language impairments, specific learning disabilities, and sensory impairments had significantly higher reading gain scores than students with intellectual impairments in 2002, but no significant differences between these groups were evident across all grade levels in 2003. In mathematics, no significant differences in gain scores based on disability group were consistently observed across target years or grade levels.

Significant differences in reading and mathematics achievement based on the amount of time students spend outside the general education classroom were found between students in LRE A and LRE C/special schools. Across subject areas and grades students in LRE A demonstrated higher overall reading and mathematics achievement than students in LRE C/special schools. Students in LRE B also had significantly higher overall reading and mathematics scores than students in LRE C/special schools, but significant differences were not observed across all grade levels. In terms of reading and mathematics gain scores, no significant differences based on LRE were consistently found across the target years and grade levels.

The vague nature of achievement based on disability category may have been influenced by several factors. Comparisons based on disability category were grounded in the assumption that students' intellectual and academic functioning closely matched the criteria for identification and placement established by the state. However, this assumption may have been flawed. Students frequently meet the criteria for more than one disability, leaving the IEP team to decide which disability will be labeled as the

primary, or most disabling condition. Researchers have also found that students are commonly labeled with more "desirable" disabilities, such as learning disabled or speech/language impaired, due to pressure from parents (McCaul & Schutz, 1991). In addition, students frequently exhibit academic difficulties in only one subject area, with average to above average achievement in other subject areas (Boudah & Weiss, 2002). These observations could serve to explain why no significant differences were consistently found based on disability group, or why students with disabilities considered less severe, such as speech/language impairments, did not consistently perform higher than students with more severe disabilities.

Differences in the reading and mathematics performance of students with disabilities based on LRE are not unexpected given that LRE placement is to be based on the intensity of students' needs and their ability to receive educational benefits in the general education classroom ("IDEA," 1997). Students who are not experiencing educational success in the general education classroom are frequently "pulled-out" and placed in more restrictive settings in the special education classroom, while the amount of time students receive instruction in the general education classroom is increased for students who are experiencing academic success.

Conclusions on the performance of students with disabilities based on disability group and LRE were also likely influenced by the small number of students in comparison groups and characteristics of students with valid assessment scores in each comparison group across years. The disability groups of intellectual impairment and sensory/emotional impairment, and LRE C were especially vulnerable to small group sizes. In addition, student participation in the assessments varied across disability groups

and LRE. For example, only 28% of students with an intellectual impairment had a valid mathematics score in 2002, and 77% of students in LRE C/special schools had valid reading scores in 2003. It is presumptive that those students with valid assessment scores represent the highest performing students in each category and are not representative of the group as a whole. Future analyses of academic achievement with larger numbers of students with disabilities in comparison groups and with equal participation across groups may produce different results.

One unexpected finding of the current study was the lack of gains made by students with disabilities between the fourth and sixth grade. The scale scores of students with disabilities changed minimally over the two-year period with smaller gains observed in reading than in mathematics. Minimal gains by students with disabilities in reading and mathematics were observed across socioeconomic levels, disability group, and LRE.

There are several possible explanations for this lack of process. One might be a function of incorrect scaling between the CTBS and CAT/5. Even though CTB McGraw-Hill asserts that the two tests are vertically equated, researchers question the accuracy with which assessments at different levels can be compared (Lissitz & Huynh, 2003). Vertically equating assessments is a difficult process complicated further when completed across two or more test levels as in the current study (Bielinski, Thurlow, Minnema, & Scott, 2000).

Assessments that are vertically equated are assumed to contain comparable content, have the same mix of test item types, and place the same demands on the test taker (Lissitz & Huynh, 2003). However, the processes and types of tasks used to measure academic achievement often change across grades, with the span of test

difficulty broadening as the level of the assessment increases (Feuer, Holland, Bertenthal, Cadelle-Hemphill, & Green, 1998; Lissitz & Huynh, 2003). It is plausible that the sixth grade assessment, in addition to containing more difficult subject matter, requires students to perform tasks or is formatted in ways that are more difficult for students with disabilities. For example, research has suggested that longer passages, fewer visual cues, and even smaller print can suppress the performance of lower-performing students (Thompson & Thurlow, 2002).

It is also possible that the gap between the students' abilities and the level of the sixth grade assessment is so broad that the test is no longer an appropriate performance measure for some students. One of the principle assumptions of current accountability models is that students' mastery of academic content can be accurately and validly measured (Fuhrman & Elmore, 2004; Linn, 2004). Large-scale assessments are designed to be appropriate measures for individuals functioning within specific ability ranges. Moderately difficult questions make up the bulk of test items at each level, however there is a paucity of simpler and more difficult test items (Bracey, 2000b). Therefore, an assessment designed for students at the sixth grade level is comprised mainly of questions that are moderately difficult for the average sixth grade student. This method of test development results in a "highly precise measurement for those examinees who can correctly answer between 40% and 80% of items" (Bielinski et al., 2000, p.3). However, the test may be a less precise measure of academic achievement for students who are functioning three or more years below the test level, and the scores obtained for these students may not be valid representations of students' true academic abilities (Bielinski et al, 2003).

A third explanation for the lack of performance gains might be that sixth grade students are less motivated and put forth less effort than students in the lower grades. The CTBS and CAT/5 are not high-stakes assessments in the state, as students receive no rewards or sanctions based on their individual performance. By sixth grade students are aware if a test "counts" and may adjust their effort accordingly (Hambleton, Impara, Mehrens, & Plake, 2000). Researchers have found that on low-stakes assessments a lack of student effort can lead to substantial underestimation of students' academic proficiency (Wise, Kingsbury, Thomason, & Xiaojing, 2004).

Finally, it is possible that the instruction of students with disabilities in middle school is less rigorous or misaligned with the standards or the assessment. In 2004 the district in this study commissioned an external audit to evaluate middle school programs. Evaluators identified several deficiencies: system-wide inconsistency in the implementation of curriculum and programs, a lack of focused supports and services to students who are not experiencing academic success, inconsistent course offerings, and unequal course access to all students. These deficiencies could contribute significantly to the poor performance of students with disabilities in middle schools.

This study did not examine the validity of the CTBS or CAT/5 for students with disabilities nor the academic programs available for these students in middle school. Further research regarding the assessment of students in middle schools is needed to determine if current assessments are valid for these students. Further examination of the performance of middle school students with disabilities in other districts and states is also necessary to determine if the observations in this study are prevalent across other settings and assessments. Further examination of the curriculum and support services could

determine if the observed trends in performance can be reversed through more rigorous instruction or support services.

Classification of Schools

In this study five accountability approaches were used to rate schools as high-performing or low-performing based on the assessment scores of students with disabilities. Schools were rated in both reading and mathematics across two consecutive years. The school ratings were examined to determine the reliability of the approaches in classifying schools within the same subject area and years, as well as across subjects and years. The ratings of schools were also analyzed to determine the level at which each approach met the standards of fairness, reliability, usefulness, and inclusiveness proposed by Baker et al. (2002).

Results indicate that school ratings and the characteristics of schools identified as high-performing for students with disabilities differed based on the methodological accountability approach employed. The approaches also differed in regard to their bias toward high poverty schools, the number of students and schools included, and in the usefulness of the ratings for the major stakeholders. However, school ratings using all five approaches were influenced by the small number of students with disabilities in each school and the small number of students with valid assessment scores.

Fairness

Of the five approaches, the cross-sectional and three-year averaging approaches were most subject to bias in favor of White and more affluent schools. Using the cross-sectional and three-year averaging approaches, schools classified as high-performing averaged fewer students who had received FARMS and fewer minority students than

schools classified as low-performing. This bias was more pronounced in reading than in mathematics. In mathematics the low-performing schools averaged 10%-30% more minority students and 10%-20% more students with disabilities who had received FARMS than high-performing schools. In reading the low-performing schools averaged 30%-40% more minority students and students who had received FARMS than high-performing schools.

The two value-added approaches were least biased toward schools in regard to the socioeconomic status of students and the percentage of minority students. Using the value-added approaches, less than ten percentage points separated the schools classified as high-performing from those classified as low-performing schools in terms of the percentage of minority students and students who had received FARMS.

These findings support those of Raudenbush (2004) and Clotfelter and Ladd (1996) who found that approaches based on average proficiency, such as cross-sectional and three-year averaging approaches, disproportionately identify high-poverty schools and schools with large minority populations as low-performing. However, while Clotfelter and Ladd found that school ratings based on the unadjusted value-added approach also showed bias toward more affluent and White schools, this was not observed in the current study.

Differences between the findings in this study and those reported by Clotfelter and Ladd (1996) could be a function of differences in the samples studied or in the number of years included. Clotfelter and Ladd analyzed school ratings based on all students, not specific subgroups. In addition, Clotfelter and Ladd analyzed school ratings across one year, while school ratings in this study were analyzed across two years.

The cross-sectional with confidence interval approach demonstrated some bias in favor of smaller schools. Because this approach adjusts the width of the confidence interval based on the number of students, schools with equal percentages of students with disabilities scoring proficient but with different numbers of test takers were often classified divergently. For example, 25% of students in schools #61 and #1 scored proficient in reading. However, school #61, with eight students, was classified as high-performing while school #1, with 20 students, was classified as low-performing. A statewide analysis of AYP results in Maryland revealed similar findings (Nelson, Rosenburg, & Kubic, 2004). A positive relationship was found between the number of students in the subgroup and the AYP failure rate.

Coladarci (2005) argues that confidence intervals are more fair because they reduce the chances of schools with small subgroups being identified as low-performing. However, Fuhrman (2003) contends that to be considered fair, an approach must treat all schools equally. Thus, the fairness of the cross-sectional with confidence interval approach is currently debated, with the determination left to each individual and his or her perceptions of equitableness.

The expectation that schools housing cluster programs for students with moderate to severe disabilities were more likely to be labeled low-performing was not borne out in the current study. Although special schools were consistently labeled low-performing, several schools with cluster programs for students with mental retardation, autism and other severe disabilities were frequently labeled as high-performing. One school (#200), which houses a district-wide program for students with autism was labeled high-performing in reading and mathematics using the adjusted value-added approach. A

second school (#211) which houses a cluster program for students with mental retardation was labeled as high-performing in reading and mathematics using the cross-sectional, cross-sectional with confidence interval, and three-year averaging approaches.

Closer inspection of these and other schools with cluster programs revealed that the majority of students with moderate to severe disabilities in this study participated in the alternate assessment, not the regular state assessment. In many of the schools with cluster programs, over fifty percent of the students took the alternate assessment. These findings suggest that because the 1% cap on the percentage of scores counting as proficient on the alternate assessment is not applied at the school level, schools with cluster programs are not at a disadvantage or treated unfairly in current accountability systems. The impact of cluster programs on school ratings is likely to further diminish as more students are removed from the regular state assessments and are administered modified assessments.

Inclusiveness

The cross-sectional, cross-sectional with confidence interval, and three-year averaging approaches were the most inclusive of the five approaches. All three of these approaches included the scores of all students with valid assessment scores, regardless of their school enrollment in previous years. In contrast, the two value-added approaches were the least inclusive of the five approaches, counting only those students with valid current and prior achievement scores. Using the value-added approaches, only 10%-40% of students in the current study were included. Only 40%-60% of schools in the two years had two or more students and were rated using the value-added approaches.

Meyer (1996) proposes that states assess students twice yearly to lessen the effects of student mobility and impute scores for students who are missing assessment data. While these proposals would most certainly increase the inclusiveness of value-added models, it is unlikely they will be adopted by school systems due to financial costs and the difficulties in justifying the use of derived scores. The inclusiveness of value-added approaches could also be increased through statewide data systems that track students across districts, but in highly mobile districts and schools large numbers of students are still likely to be excluded from accountability systems using value-added approaches.

Usefulness

While I reviewed the literature regarding the usefulness of accountability approaches, gathering this information directly from educators, parents, or policymakers was beyond the scope of this study. Past research has indicated that educators will find the value-added approaches most useful because these approaches provide information on students' gains, which can then be used to evaluate classroom instruction and make instructional adjustments. Policymakers may also find the value-added approaches useful in evaluating the effects of new programs, curricula, or instructional strategies and identifying schools with exemplary policies and practices. However, in order to make decisions regarding consequences and monitor equity, policymakers may also find it useful to know the average proficiency of schools. In contrast, past research has shown that parents do not gather information about school performance in making choices regarding schools (Orfield, 2003; Schneider & Buckley, 2002). Rather, parents frequently make school choices based on factors that have little to do with quality of

education, instead relying on factors such as the availability of day care, convenience, social factors, and sports (Schneider & Buckley).

Thus, the usefulness of any given accountability approach is likely to depend on the needs of the stakeholder group as well as the knowledge and skills of each individual. Further research, especially of a qualitative nature, is needed to provide empirical evidence regarding the usefulness of accountability approaches. This research should examine not only the informational needs of the major stakeholders, but how accountability systems can best be designed to provide this information in a concise and comprehensible manner that is beneficial to all stakeholders.

Reliability

In terms of reliability, none of the five accountability approaches employed in the current study produced consistent lists of high-performing schools based on the performance of students with disabilities. When using the same approach, schools were often classified as high-performing in only one of the two subject areas or in only one of the two years. When examined across approaches, schools were frequently rated high-performing using one approach and low-performing using the remaining approaches.

Within the same subject area, the three-year averaging and cross-sectional approaches were the most reliable in rating schools. Using the three-year averaging approach less than 3% of schools changed classifications from year to year, while less than 13% of schools changed classifications using the cross-sectional approach. The cross-sectional approach with a confidence interval consistently classified the largest percentage of schools as high-performing, with approximately one-fourth of schools rated high-performing using this approach. The value-added approaches were the most

unreliable of the approaches, with 40%-50% of schools changing classifications between the two years using the adjusted and unadjusted value-added approaches.

Across subject areas and years, no schools were labeled high-performing in reading and mathematics using all five approaches. One school (#211) was classified as high-performing using the cross-sectional, cross-sectional with confidence interval, and three-year averaging approaches. However, this school was not rated using the value-added approaches. A second school (#77) was rated as high-performing in reading and mathematics using both of the value-added approaches and the cross-sectional with confidence interval approach, but was classified as low-performing under the cross-sectional and three-year averaging approaches.

The reliability of approaches in the current study is lower than that found by other researchers. When comparing school ratings across four approaches, Rubenstein et al. (2004) found that 60%-80% of schools were consistently rated, while Raudenbush (2004) found that the correlation between value-added ratings was approximately .90. The lower reliability of school ratings in the current study is most likely attributable to the small number of students in the subgroup of students with disabilities. Rubenstein et al. and Raudenbush rated schools based on the total school population and did not examine school ratings for specific subgroups. However, both groups of researchers found that schools with larger numbers of students were more consistently ranked and that the reliability of ratings decreased when the number of students included was reduced.

In this study, the inclusiveness, fairness, and reliability of all of the approaches was impacted by the use of non-standard accommodations. Forty-one percent of mathematics scores in 2002 were invalidated because students used a non-standard

accommodation. While excluding these students limited the inclusiveness of all the accountability approaches, the option of counting these scores as Basic or nonproficient is unfair to schools providing non-standard accommodations to large numbers of students and may serve as a disincentive to providing certain accommodations to students with disabilities.

Although the state has decreased the number of accommodations considered non-standard, it has done so in a manner that was not transparent to educational practitioners. The state changed the calculator accommodation from non-standard to standard after districts had already administered the state assessments. Besides failing to provide districts and schools with sufficient notice of the change, the state also failed to provide a rationale for the change. In explaining the accommodation change to the local school board, the district superintendent stated, "The manner in which the state changed the calculation methodology and informed school districts is a matter of dispute. At this time, the calculation methodology still is not entirely clear….There has been no explanation from (the state)."

Because accommodations have the potential to significantly affect school ratings, it is imperative that policymakers inform practitioners of accommodation policies in advance of assessments. It is also important that researchers and policymakers be cognizant of the number and percentages of students with invalid assessment scores when making judgments regarding programs and when rating schools.

The findings of this study suggest that the standards of reliability, inclusiveness, fairness, and usefulness of accountability approaches may vary greatly from approach to approach and from year to year. These findings also support earlier research which

concluded that both status and value-added approaches produce estimates with "considerable uncertainty and some unknown bias" (Raudenbush, 2004, p.36). Due to the small number of students with disabilities in some schools and policies specific to this subgroup, there is a need for researchers to examine the uncertainties and biases of various accountability approaches on subgroups and not solely with the overall student population.

Implications for Policy and Research

There are a number of implications for policy and research that can be derived from this study. This study revealed that the majority of middle schools were rated low-performing using all of the accountability approaches and that middle school students made minimal gains in reading and mathematics between the fourth and sixth grade. These findings support findings from a nationwide analysis of schools that found a concentration of middle schools listed as "needing improvement" (Center on Education Policy, 2005).

The poor performance of students with disabilities in middle school raises questions regarding not only the assessment of these students, but the academic instruction these students receive. Further research is needed to confirm if these patterns are evident in other states and when using different assessments. Researchers and policymakers need to examine more closely the use of state assessments for students across a broad range of abilities. This is especially true for students with disabilities in middle and secondary school who may be functioning three or more years below the level of the state assessment.

The recently approved modified assessments may provide states with an option to measure the performance of students who are functioning too high to participate in the alternate assessment, but for whom the regular assessment may not produce a valid score. Researchers will need to address not only the technical adequacy of the new modified assessments, but also the process by which students are assigned to either the alternate, modified, or regular assessments (Consortium for Citizens with Disabilities, 2005). Policymakers will need to monitor the use of the various assessments at the school level and establish guidance to help ensure that students are assigned to an assessment based on the appropriateness of the chosen test and not whether the assessment increases a school's chances of making AYP.

The absence of middle schools rated as high-performing using any of the approaches also raises questions about the instruction students with disabilities receive in middle school. While states have implemented early literacy reforms at the elementary level and established exit exams at the secondary level, the need for reforms at the middle school level is just being recognized. The district in this study has implemented several new middle school programs as a result external evaluations: expanding the extended day program and reducing class size; purchasing English, Science and Life Skills textbooks for students with disabilities; and providing reading intervention programs including Read 180 and Corrective Reading. Policymakers and educators will need to evaluate these new programs to determine if they are beneficial in raising the academic achievement of students with disabilities.

Number of Students in Subgroups and Minimum n

Policymakers and researchers also need to address the small number of students in subgroups and its effect on school ratings and the minimum number of students ("minimum *n*") that must be in a group before the school is held accountable. School ratings based on the subgroup of students with disabilities Many states have established very high minimum *n*s, leaving a large percentage of schools not responsible for the performance of certain subgroups. The Center for Education Policy (2005) found that 92% of schools in California were not held accountable for students with disabilities in 2005 because the subgroup did not contain at least 100 students.[14] Additionally, an analysis of five states indicated that 80% of schools that made AYP did so without being accountable for students with disabilities (Center for Education Policy, 2005). While high minimum *n*s mean that schools may not be held accountable for the performance of many students, the alternative use of confidence intervals may be unfair to larger schools and may limit the usefulness of accountability approaches.

One of the primary goals of NCLBA is to hold schools accountable for the academic achievement of all students. To achieve this goal it is important that policymakers, with input from researchers, establish some clear guidelines on minimum *n*. States have extremely varied policies governing minimum *n,* and the basis for most of these policies is unclear. Policymakers will need to work closely with researchers to develop guidelines that assure that school ratings are reliable and valid, but prevent schools from using policies such as minimum *n* as a loophole not to be held accountable for the performance of some subgroups.

Hybrid Accountability Approaches

Findings in this study illustrate the inherent weaknesses of both status and value-added approaches in accountability systems. The status approaches were biased against schools with large percentages of students in poverty and minority students. The status approaches were also unfair to schools that were successfully increasing the academic achievement of students by not recognizing their accomplishments. However, the value-added approaches excluded large numbers of students and schools and rated schools as high-performing even though very few students demonstrated proficiency.

In order to accurately rate schools and produce information that is useful to all major stakeholders, both the level of performance and changes in academic achievement appear necessary (Chief State School Officers, 2005). Researchers have recently begun developing "hybrid" accountability approaches that combine status and value-added accountability components. Researchers at Northwest Evaluation Association developed the "hybrid success model" which identifies a growth target for each student that will result in reaching or surpassing proficiency by a set date and calculates the effect of the school on a student by tracking the proportion of the growth that was obtained (McCall, Kingsbury, & Olsen, 2004). A second hybrid approach, the REACH (Rate of Expected Academic Change), was developed at the Pacific Institute and measures student academic growth against the goal of subject-matter proficiency (Doran & Izumi, 2004).

The US Department of Education recently announced that grants would be awarded to ten states to pilot the use of hybrid accountability approaches (US Department of Education, 2005b). States that are awarded the grants must continue to disaggregate results by subgroups and have an overall goal of proficiency for all students by 2014, but

will be allowed to incorporate growth models into AYP calculations. Policymakers and researchers will need to monitor the use of the new approaches to ensure that they do not lower the goal of proficiency for all students. Researchers will also need to examine these new approaches in respect to validity, reliability, usefulness, fairness, and inclusiveness.

Theory of Action.

Finally, further research is needed to verify the theory of action of accountability systems. State policies are principally driven by the response from teachers, parents, and the public. Missouri, for example, justified lowering their proficiency rates in mathematics from 31.1% to 17.5% for fear that teachers "would just throw up their hands and say there was no way to meet the targets that were initially set" (Sherry, 2005, p. A13).

The purpose of accountability systems as currently implemented is to motivate teachers and schools to work harder. However, there is no empirical evidence to support this assumption. In fact, little is known about the true effects of accountability systems on the actions of the major actors, such as teachers and administrators. Do accountability systems motivate teachers to work harder? Does student achievement increase when schools are held accountable?

Preliminary research has found that external accountability systems are limited in their ability to foster school improvement (O'Day, 2004), especially among schools with students of lower socioeconomic and academic achievement (DeBray, Parson, & Woodworth, 2001). Some research suggests that low-performing schools may actually

lose ground relative to higher-performing schools when an external accountability system is instituted (DeBray et al., 2001).

States are already facing severe shortages of qualified teachers, especially in the field of special education and in schools with high poverty rates. Ninety-eight percent of school districts in 2000 reported shortages of special education teachers (Fideler, Foster, & Schwartz, 2000). Researchers have found that in all disciplines experienced teachers tend to shift to schools serving fewer poor, minority, and low-achieving students (Hanushek, Kain, & Rivkin, 2001).

It is plausible that accountability systems as currently being implemented will have the opposite effect than that advanced in their theory of action. Instead of motivating teachers to work harder, accountability systems may in fact drive teachers from the profession or drive dedicated individuals to seek jobs in schools that are labeled as high-performing. Any approach, regardless of its reliability, validity, inclusiveness, or even its usefulness is not likely to withstand political pressure if it leads to further shortages of teachers or produces disparities in the quality of teachers in high poverty schools. Further research is needed to provide a more in-depth view of teachers' and administrators' beliefs on accountability systems and their effects on motivation and retention.

Summary

The findings of this study are promising in that students with disabilities are being increasingly included in state assessments, and accountability reforms are requiring that schools be held accountable for the performance of these students. However, the results

of this study suggest that there are many uncertainties regarding the assessment and use of accountability systems for students with disabilities.

Accountability systems are grounded in the assumptions that we can accurately measure students' academic performance through assessments, use the results to reliably and validly rate schools, and motivate educators through rewards and sanctions. Yet, these assumptions are largely unproven, especially for subgroups such as students with disabilities. There are uncertainties regarding the use of large-scale assessments to measure the performance of students with disabilities, and the degree to which schools can be reliably rated using the performance of this subgroup is largely unknown.

More importantly, there is disagreement on the definition of a high-performing or effective school. Policymakers have enacted policies mandating that schools be rated on the absolute performance of students. Research and practitioners, however, believe that schools should be judged on the whether they are increasing the academic achievement of students. Without agreement on this basic guiding principle of accountability systems, there is likely to be continued debate on the use and application of accountability approaches in rating the quality of schools.

As policymakers work to redefine accountability systems, it is important that they not totally abandon the desire to hold schools accountable for the academic achievement of all students. Rather, by searching for answers to the basic questions and assumptions that underlie accountability systems, policymakers can develop accountability systems that lead to increased academic achievement. A properly designed accountability system based on solid research and with input from policymakers, educators, parents, can be the

catalyst for change, and is the right step in ensuring that all students reach their highest academic potential.

Appendix A

Disabilities as Defined in State Code of Regulations

Autism: A developmental disability which does not include emotional disturbance that significantly affects verbal and nonverbal communication and social interaction, is generally evident before 3 years old, and adversely affects educational performance. Autism may be characterized by: Engagement in repetitive activities and stereotyped movements, resistance to environmental change or change in daily routines, and unusual responses to sensory experiences.

Deaf-blindness: Concomitant hearing and visual impairments, the combination of which causes such severe communication and educational problems that the student cannot be accommodated solely as a student with deafness or a student with blindness.

Orthopedic Impairment: A severe orthopedic impairment that adversely affects a child's educational performance. The term includes impairments caused by congenital anomaly (e.g., club foot, absence of some member, etc.), impairments caused by disease (e.g., poliomyelitis, bone tuberculosis), and impairments from other causes (e.g., cerebral palsy, amputations, and fractures or burns).

Emotional Disturbance: A condition exhibiting one or more of the following characteristics over a long period of time and to a marked degree, and that adversely affects a student's educational performance: An inability to learn that cannot be explained by intellectual, sensory, or health factors, an inability to build or maintain satisfactory interpersonal relationships with peers and teachers, inappropriate types of behavior or feelings under normal circumstances, a general, pervasive mood of unhappiness or depression, or a tendency to develop physical symptoms or fears associated with personal

or school problems. Emotional disturbance includes schizophrenia, but does not include those determined socially maladjusted, unless they also have a diagnosis of emotional disturbance.

Deafness: A hearing impairment that is so severe that the student is impaired in processing linguistic information through hearing, with or without amplification, and adversely affects the student's educational performance.

Hearing impairment: An impairment in hearing, whether permanent or fluctuating, that adversely affects a student's educational performance.

Mental retardation: General intellectual functioning, adversely affecting a student's educational performance, which: Is significantly subaverage, exists concurrently with deficits in adaptive behavior, and is manifested during the developmental period.

Multiple Disabilities: Concomitant impairments, such as mental retardation-blindness or mental retardation-orthopedic impairment, the combination of which causes such severe educational problems that the student cannot be accommodated in special education programs solely for one of the impairments. Does not include students with deaf-blindness.

Other health impairment: Having limited strength, vitality, or alertness, including a heightened alertness to environmental stimuli that results in limited alertness with respect to the educational environment, that is adversely affecting a student's educational performance, or due to chronic or acute health problems such as: Attention deficit disorder or attention deficit hyperactivity disorder, diabetes, epilepsy, a heart condition, hemophilia, lead poisoning, leukemia, nephritis, rheumatic fever, or sickle cell anemia.

Specific Learning Disability (SLD): A disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, that may manifest itself in an imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations. SLD includes conditions such as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. SLD does not include students who have learning problems which are primarily the result of visual, hearing, or motor impairments, mental retardation, emotional disturbance, or environmental, cultural, or economic disadvantage.

Speech or language impairment: A communication disorder such as stuttering, impaired articulation, voice impairment, or language impairment that adversely affects a student's educational performance.

Developmental delay: A student 3 years old through 9 years old with a 25 percent or greater delay in adaptive, cognitive, communicative, emotional, physical, or social development, atypical development or behavior, a diagnosed physical or mental condition, visual impairment, including blindness; and who, because of the impairment, needs special education and related services.

Traumatic Brain Injury: An acquired injury to the brain, caused by an external force, resulting in total or partial functional disability or psychosocial impairment, or both, that adversely affects a student's educational performance. Includes open or closed head injuries resulting in impairments in one or more areas such as: cognition, language, memory, attention, reasoning, problem solving, sensory, perceptual, and motor abilities, physical functions, or information processing. Traumatic brain injury does not include brain injuries that are congenital or degenerative, or induced by birth trauma.

Appendix B

Description and Location of District Cluster Programs

Elementary Learning Centers (ELC) serve multiple-needs children in grades K-5 and includes students with learning disabilities and language, emotional, visual, hearing, or orthopedic impairments. ELCs are located in eleven elementary schools.

Secondary Learning Centers (SLC) are designed for academically challenged learning disabled students in middle and high school. Students receive special education instruction for several class periods but are integrated into the general education program whenever possible. SLC programs are located in five middle schools in the district, and three high schools.

Autism programs serve students whose needs cannot be met in less restrictive environments, and provide a low teacher/student ratio and highly structured individual curriculum. Autism programs are offered in five elementary schools, two middle schools, and two high schools.

Learning for Independence (LFI) programs emphasize individualized student learning in school and community sites and serve students with mild to moderate mental retardation and/or multiple disabilities. Students in LFI learn functional life skills and basic academics, but are often included in general education classes with modifications. LFI programs are offered in three elementary schools, seven middle and eight high schools.

School/Community Based (SCB) programs serve students with mild/moderate to severe/profound disabilities and are located in fifteen elementary schools, eight middle, and nine high schools.

Emotional disabilities (ED) cluster programs serve students with severe emotional and behavioral disabilities who need comprehensive behavior intervention and alternative structure. ED cluster programs are offered in nine high schools, eight middle schools, and five elementary schools.

Gifted LD programs serve students who have been diagnosed with a learning disability but who score two or more standard deviations above the mean on tests of intellectual and cognitive abilities.

In addition to these cluster programs, there are also district-wide programs for students with aspergers, and for those with vision and hearing impairments.

Table C1

School Classifications in Reading Using Cross-Sectional Approach

	2002	2003
High-performing	38, 45, 63, 74, 82, **112**, **114**, 154, 182**, 211**, 258, **269**, 288	6, 53, 110, **112**, **114**, 135**,** 151, 161**,** 166, **211**, 224, 227, **269**, 270, 289
Low-performing	1, 2, 6, 7, 9, 10, 11, 12, 14, 17, 19, 22, 23, 25, 27, 28, 29, 31, 35, 36, 40, 41, 50, 52, 53, 55, 56, 57, 58, 59, 60, 61, 64, 68, 71, 72, 75, 76, 77, 80, 81, 87, 88, 90, 91, 92, 93, 95, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 115, 120, 122, 129, 134, 135, 136, 138, 139, 141, 143, 144, 145, 150, 151, 153, 156, 158, 159, 161, 162, 163, 164, 165, 166, 167, 170, 171, 173, 174, 176, 177, 178, 180, 181, 183, 195, 196, 198, 200, 203, 204, 205, 209, 210, 212, 213, 215, 217, 222, 224, 225, 227, 231, 232, 233, 234, 237, 245, 247, 250, 253, 254, 260, 261, 262, 263, 266, 267, 270, 272, 273, 275, 277, 279, 280, 281, 283, 284, 285, 286, 289, 291, 294, 297, 298, 299	1, 2, 7, 9, 10, 11, 12, 14, 17, 19, 22, 23, 25, 27, 28, 29, 31, 35, 36, 38, 40, 41, 45, 50, 52, 55, 56, 57, 58, 59, 60, 61, 63, 64, 68, 71, 72, 74, 75, 76, 77, 80, 81, 82, 87, 88, 90, 91, 92, 93, 95, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 111, 115, 120, 122, 129, 134, 136, 138, 139, 141, 143, 144, 145, 150, 153, 154, 156, 158, 159, 160, 162, 163, 164, 165, 167, 170, 171, 173, 174, 176, 177, 178, 180, 181, 182, 183, 195, 196, 198, 200, 203, 204, 205, 209, 210, 212, 213, 215, 217, 222, 225, 231, 232, 233, 234, 237, 245, 247, 250, 253, 254, 258, 260, 261, 262, 263, 266, 267, 272, 273, 275, 277, 279, 280, 281, 283, 284, 285, 286, 288, 291, 294, 297, 298, 299
Not rated	62, 201	62, 201

Note. Schools classified as high-performing in both years are indicated in bold.

Table C2

School Classifications in Mathematics Using Cross-Sectional Approach

	2002	2003
High-performing	9, 11**, 45**, 82, 88, 92, 93, 122, 139, 145, 161, **166**, **211,** 224, 254	**45**, 74, 141, 154**, 166, 211**, 269
Low-performing	1, 2, 6, 7, 10, 12, 14, 17, 19, 22, 23, 25, 27, 28, 29, 31, 35, 36, 38, 40, 41, 50, 52, 53, 55, 56, 57, 58, 59, 60, 61, 63, 64, 68, 71, 72, 74, 75, 76, 77, 80, 81, 87, 90, 91, 95, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 115, 120, 129, 134, 135, 136, 138, 141, 143, 144, 150, 151, 153, 154, 156, 158, 159, 162, 163, 164, 165, 167, 170, 171, 173, 174, 176, 177, 178, 180, 181, 182, 183, 195, 196, 198, 200, 201, 203, 204, 205, 209, 210, 212, 213, 215, 217, 222, 225, 227, 231, 232, 233, 234, 237, 245, 247, 250, 253, 258, 260, 261, 262, 263, 266, 267, 269, 270, 272, 273, 275, 277, 279, 280, 281, 283, 284, 285, 286, 288, 289, 291, 294, 297, 298, 299	1, 2, 6, 7, 9, 10, 11, 12, 14, 17, 19, 22, 23, 25, 27, 28, 29, 31, 35, 36, 38, 40, 41, 50, 52, 53, 55, 56, 57, 58, 59, 60, 61, 63, 64, 68, 71, 72, 75, 76, 77, 80, 81, 82, 87, 88, 90, 91, 92, 93, 95, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 114, 115, 120, 122, 129, 134, 135, 136, 138, 139, 143, 144, 145, 150, 151, 153, 156, 158, 159, 160, 161, 162, 163, 164, 165, 167, 170, 171, 173, 174, 176, 177, 178, 180, 181, 182,183, 195, 196, 198, 200, 203, 204, 205, 209, 210, 212, 213, 215, 217, 222, 224, 225, 227, 231, 232, 233, 234, 237, 245, 247, 250, 253, 254, 258, 260, 261, 262, 263, 266, 267, 270, 272, 273, 275, 277, 279, 280, 281, 283, 284, 285, 286, 288, 289, 291, 294, 297, 298, 299
Not rated	62, 201	62, 201

Note. Schools classified as high-performing in both years are indicated in bold.

Table C3

School Classifications in Reading Using Cross-Sectional with Confidence Interval Approach

	2002	2003
High-performing	1,**6**, 7, 1**0**, 22, 23**, 29**, 31, **35, 38, 45, 53**, **56**, 57, 60**, 61, 63**, **74, 77, 81**, 82, **91, 92**, 98, 105, 106, **109, 110, 111, 112, 114**, **120,** 134, **135**, 145, **151**, 153, **154, 165, 166**, **178, 182,** 183, **198**, 200, **211**, **227, 258, 269, 270**, 272, 273, 275, 284, **288, 289**, 299	**6**, **10**, 14, **29**, **35, 38, 45**, 52, **53, 56**, 59, **61, 63**, **74**, **77, 81, 91, 92,** 95, **109, 110, 111**, **112, 114,** 115, **120**, **135**, **151, 154**, 161, 162, **165, 166, 178**, 180, **182, 198,** 203, **211,** 217, 224, **227,** 233, 245, **258,** 260, **269**, **270,** 286, **288, 289**, 291, 297
Low-performing	2 ,9, 11, 12, 14, 17, 19, 25, 27, 28, 36, 40, 41, 50, 52, 55, 58, 59, 64, 68, 71, 72, 75, 76, 80, 87, 88, 90, 93, 95, 97, 100, 101, 102, 103, 104, 107, 108, 115, 122, 129, 136, 138, 139, 141, 143, 144, 150, 156, 158, 159, 161, 162, 163, 164, 167, 170, 171, 173, 174, 176, 177, 180, 181, 195, 196, 203, 204, 205, 209, 210, 212, 213, 215, 217, 222, 224, 225, 231, 232, 233, 234, 237, 245, 247, 250, 253, 254, 260, 261, 262, 263, 266, 267, 277, 279, 280, 281, 283, 285, 286, 291, 294, 297, 298	1, 2, 7, 9, 11, 12, 17, 19, 22, 23, 25, 27, 28, 31, 36, 40, 41, 50, 55, 57, 58, 60, 64, 68, 71, 72, 75, 76, 80, 82, 87, 88, 90, 93, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 122, 129, 134, 136, 138, 139, 141, 143, 144, 145, 150, 153, 156, 158, 159, 160, 163, 164, 167, 170, 171, 173, 174, 176, 177, 181, 183, 195, 196, 200, 204, 205, 209, 210, 212, 213, 215, 222, 225, 231, 232, 234, 237, 247, 250, 253, 254, 261, 262, 263, 266, 267, 272, 273, 275, 277, 279, 280, 281, 283, 284, 285, 294, 298, 299
Not rated	62, 201	62, 201

Note. Schools classified as high-performing in both years are indicated in bold.

Table C4

School Classifications in Mathematics Using Cross-Sectional with Confidence Interval Approach

	2002	2003
High-performing	**1**, 7, 9, 11, 12, 22, **23**, 28, **31**, 35, 36, **38**, 40, **45**, **50**, 52, **53**, **56,** 57, 58, 60, 61, **63**, 68, 72, **74**, 76, **77,** 80, **81**, 82, **88**, 90, 91, **92,** 93, 98, 100, 101, 103, 106, 109, **110, 114**, **120, 122**, 134, **135**, 136, **139**, **141**, 143, 145, 150, 153, **154, 161**, 163, 164, 165, **166,** 174, 177, **178,** 181, **182,** 183, 195, **198**, 200, **204,** 209, **211**, **224,** 227, 231, 232, **233**, 247, 250, 254, **258**, **260**, 261, **263**, **269, 270**, 272, 279, 280, 281, 284, **288**, **289**, 291, 294, 297, 298	**1**, 6, 10, 14**, 23, 31, 38, 45, 50, 53,** 55**, 56,** 59**, 63, 74, 77, 81, 88, 92,** 102, 105**, 110,** 112, **114, 120, 122, 135, 139**, **141,** 151, **154,** 158, 160, **161,** 162, **166, 178, 182, 198, 204**, **211**, 215, 217, **224, 233, 258, 260, 263, 269, 270**, 285, **288, 289**
Low-performing	2, 6, 10, 14, 17, 19, 25, 27, 29, 41, 59, 64, 71, 75, 95, 97, 102, 104, 105, 107, 108, 111, 112, 115, 129, 138, 144, 151, 156, 158, 159, 162, 167, 170, 171, 173, 176, 180, 196, 203, 205, 210, 212, 213, 215, 217, 222, 225, 234, 237, 245, 253, 262, 266, 267, 273, 275, 277, 283, 285, 286, 299	2, 7, 9, 11, 12, 17, 19, 22, 25, 27, 28, 29, 35, 36, 40, 41, 52, 57, 58, 60, 61, 64, 68, 71, 72, 75, 76, 80, 82, 87, 90, 91, 93, 95, 97, 98, 100, 101, 103, 104, 106, 107, 108, 109, 111, 115, 129, 134, 136, 138, 143, 144, 145, 150, 153, 156, 159, 163, 164, 165, 167, 170, 171, 173, 174, 176, 177, 180, 181, 183, 195, 196, 200, 203, 205, 209, 210, 212, 213, 222, 225, 227, 231, 232, 234, 237, 245, 247, 250, 253, 254, 261, 262, 266, 267, 272, 273, 275, 277, 279, 280, 281, 283, 284, 286, 291, 294, 297, 298, 299
Not rated	62, 201	62, 201

Note. Schools classified as high-performing in both years are indicated in bold.

Table C5

School Classifications in Reading Using Three Year Averaging Approach

	2002	2003
High-performing	38, **45, 53**, **74,** 111, **114, 135**, 154, **182, 211, 269, 288**	**45, 53, 74, 114,135**, **182, 211, 269,** 270, **288**
Low-performing	1, 2, 6, 7, 9, 10, 11, 12, 14, 17, 19, 22, 23, 25, 27, 28, 29, 31, 35, 36, 40, 41, 50, 52, 55, 56, 57, 58, 59, 60, 61, 63, 64, 68, 71, 72, 75, 76, 77, 80, 81, 82, 87, 88, 90, 91, 92, 93, 95, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 112, 115, 120, 122, 129, 134, 136, 138, 139, 141, 143, 144, 145, 150, 151, 153, 156, 158, 159, 161, 162, 163, 164, 165, 166, 167, 170, 171, 173, 174, 176, 177, 178, 180, 181, 183, 195, 196, 198, 200, 203, 204, 205, 209, 210, 212, 213, 215, 217, 222, 224, 225, 227, 231, 232, 233, 234, 237, 245, 247, 250, 253, 254, 258, 260, 261, 262, 263, 266, 267, 270, 272, 273, 275, 277, 279, 280, 281, 283, 284, 285, 286, 289, 291, 294, 297, 298, 299	1, 2 , 6, 7, 9, 10, 11, 12, 14, 17, 19, 22, 23, 25, 27, 28, 29, 31, 35, 36, 38, 40, 41, 50, 52, 55, 56, 57, 58, 59, 60, 61, 63, 64, 68, 71, 72, 75, 76, 77, 80, 81, 82, 87, 88, 90, 91, 92, 93, 95, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 115, 120, 122, 129, 134, 136, 138, 139, 141, 143, 144, 145, 150, 151, 153, 154, 156, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 170, 171, 173, 174, 176, 177, 178, 180, 181, 183, 195, 196, 198, 200, 203, 204, 205, 209, 210, 212, 213, 215, 217, 222, 224, 225, 227, 231, 232, 233, 234, 237, 245, 247, 250, 253, 254, 258, 260, 261, 262, 263, 266, 267, 273, 275, 277, 279, 280, 281, 283, 284, 285, 286, 289, 291, 294, 297, 298, 299
Not rated	62, 201	62, 201

Note. Schools classified as high-performing in both years are indicated in bold.

Table C6

School Classifications in Mathematics Using Three-Year Averaging Approach

	2002	2003
High-performing	**211**, 224	166, **211**
Low-performing	1, 2, 6, 7, 9, 10, 11, 12, 14, 17, 19, 22, 23, 25, 27, 28, 29, 31, 35, 36, 38, 40, 41, 45, 50, 52, 53, 55, 56, 57, 58, 59, 60, 61, 63, 64, 68, 71, 72, 74, 75, 76, 77, 80, 81, 82, 87, 88, 90, 91, 92, 93, 95, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 114. 115. 120. 122. 129. 134. 135. 136. 138. 139. 141. 143. 144. 145. 150. 151. 153. 154. 156. 158. 159. 161. 162. 163. 164. 165. 166. 167. 170. 171. 173. 174. 176. 177. 178. 180. 181. 182. 183. 195. 196. 198. 200. 203. 204. 205. 209. 210. 212. 213. 215. 217. 222. 225. 227. 231. 232. 233. 234. 237. 245. 247. 250. 253. 254. 258. 260. 261. 262. 263. 266. 267. 269. 270. 272. 273. 275. 277. 279. 280. 281. 283. 284. 285. 286. 288. 289. 291. 294. 297. 298. 299	1, 2, 6, 7, 9, 10, 11, 12, 14, 17, 19, 22, 23, 25, 27, 28, 29, 31, 35, 36, 38, 40, 41, 45, 50, 52, 53, 55, 56, 57, 58, 59, 60, 61, 63, 64, 68, 71, 72, 74, 75, 76, 77, 80, 81, 82, 87, 88, 90, 91, 92, 93, 95, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 114. 115. 120. 122. 129. 134. 135. 136. 138. 139. 141. 143. 144. 145. 150. 151. 153. 154. 156. 158. 159. 160, 161. 162. 163. 164. 165. 167. 170. 171. 173. 174. 176. 177. 178. 180. 181. 182. 183. 195. 196. 198. 200. 203. 204. 205. 209. 210. 212. 213. 215. 217. 222. 224, 225. 227. 231. 232. 233. 234. 237. 245. 247. 250. 253. 254. 258. 260. 261. 262. 263. 266. 267. 269. 270. 272. 273. 275. 277. 279. 280. 281. 283. 284. 285. 286. 288. 289. 291. 294. 297. 298. 299
Rating Not Available	62, 201	62, 201

Note. Schools classified as high-performing in both years are indicated in bold.

Table C7

School Classifications in Reading using Unadjusted Value-added Approach

	2002	2003
High-performing	6, 31, **40**, 57, 61, 74, 76, **77**, 81, 91, 97, 101, 106, 120, **134**, **156**, 158, 162, 176, 178, 200, 212, 234, **227**, 253, **260, 266**, **275**, 277	2, 10, 17, 25, 29, **40**, 45, 71, **77**, 87, 92, 95, 98, 103, 109, 110, 111, 112, 114, 115, **134**, 135, 139, 144, 150, **156**, 165, 180, 181, 183, 205, 210, 213, 224, **227,** 234, 258, **260,** 262, **266**, 267, 270, 272, **275**, 279, 286, 289
Average-performing	1, 2, 12, 17, 38, 58, 59, 71, 92, 95, 103, 139, 150, 171, 173, 181, 183, 195, 204, 209, 237, 261, 262, 281, 283, 288, 289, 297, 298	1, 12, 19, 31, 35, 38, 41, 53, 55, 56, 57, 60, 64, 74, 76, 80, 81, 91, 104, 105, 106, 107, 108, 129, 141, 151, 153, 158, 162, 171, 178, 198, 200, 203, 212, 215, 231, 232, 233, 245, 253, 277, 280, 281, 285, 291, 297, 299
Low-performing	11, 36, 50, 72, 80, 88, 90, 93, 100, 102, 104, 115, 122, 136, 138, 151, 159, 161, 164, 170, 174, 177, 222, 225, 232, 250 ,263, 285	9, 11, 14, 23, 28, 36, 50, 58, 68, 72, 75, 88, 90, 93, 97, 100, 101, 122, 136, 138, 143, 159, 161, 163, 164, 170, 173, 174, 176, 177, 182, 195, 196, 204, 209, 222, 225, 237, 247, 250, 254, 261, 263, 283, 284, 294, 298
Not rated	7, 9, 10, 14, 19,22, 23,25, 27, 28, 29, 35, 41, 45, 52, 53, 55, 56, 60, 62, 63, 64, 68, 75, 82, 87, 98, 105, 107, 108, 109, 110, 111, 112, 114, 129, 135, 141, 143, 144, 145, 153, 154, 163, 165, 166, 167, 180, 182, 196, 198, 201, 203, 205, 210, 211, 213, 215, 217, 224, 231, 233, 245, 247, 254, 258, 267, 269, 270, 272, 273, 279, 280, 284, 286, 291, 294, 299	6, 7, 22, 27, 52, 59, 61, 62, 63, 82, 102, 120, 145, 154, 166, 167, 168, 201, 211, 217, 269, 273, 288

Note. Schools classified as high-performing in both years are indicated in bold.

Table C8

School Classifications in Mathematics Using Unadjusted Value-added Approach

	2002	2003
High-performing	1, 6, **17**, 23, 31, **38**, **41**, 56, 57, 59, 64, 71, 74, 75, **77**, 81, 95, **98, 101**, 102, **103**, **108**, 112, 141, 150, 162, **178**, 181, **200**, 205, 231, 254, 262, 266, 267, **277**, 291,	2, 10, **17**, 19, 25, **38**, 40, **41,** 45, 55, **77**, 91, 92, **98**, **101**, **103**, 104, 107, **108**, 111, 144, 151, 153, 158, 165, **178**, 182, 198, **200,** 203, 215, 224, 234, 245, 253, 260, 272, 275, **277,** 281, 284,285, 286
Average-performing	2, 19, 55, 60, 61, 87, 93, 97, 106, 107, 109, 111, 114, 120, 134, 135, 153, 158, 165, 171, 174, 180, 196, 209, 215, 217, 227, 234, 253, 263, 275, 281, 283, 288, 289, 297	1, 12, 14, 23, 29, 31, 35, 53, 56, 60, 64, 68, 71, 81, 95, 97, 105, 106, 109, 110, 112, 129, 134, 135, 141, 156, 170, 180, 181, 204, 205, 210, 213, 227, 231, 237, 263, 267, 270, 279, 283, 289, 291
Low-performing	9, 11, 12, 28, 36, 50, 58, 68, 72, 88, 90, 92, 100, 115, 122, 136, 138, 151, 159, 161, 164, 173, 176, 177, 195, 204, 222, 225, 232, 237, 247, 250, 261, 284, 294, 298	9, 11, 28, 36, 50, 58, 72, 76, 80, 87, 88, 90, 93, 100, 115, 122, 136, 138, 143, 159, 161, 164, 171, 173, 174, 176, 177, 195, 209, 222, 225, 232, 233, 247, 250, 254, 261, 262, 266, 294, 298, 299
Not rated	7, 10, 14, 22, 25, 27, 29, 35, 40, 45, 52, 53, 62, 63, 76, 80, 82, 91, 104, 105, 110, 129, 139, 143, 144, 145, 154, 156, 163, 166, 167, 170, 182, 183, 198, 201, 203, 210, 211, 212, 213, 224, 233, 245, 258, 260, 269, 270, 272, 273, 279, 280, 285, 286, 299	6, 7, 22, 27, 52, 57, 59, 61, 62, 63, 74, 75, 82, 102, 114, 120, 139, 145, 150, 154, 160, 162, 163, 166, 167, 183, 196, 201, 211, 212, 217, 258, 269, 273, 280, 288, 297

Note. Schools classified as high-performing in both years are indicated in bold.

Table C9

School Classifications in Reading Using Adjusted Value-added Approach

	2002	2003
High-performing	6, 31, **40**, 57, 61, 74, 77, 81, 91, 97, 101, 106, 120, **134**, **156**, 158, 162, 176, 178, **183, 200,** 212, **227, 234**, 253, 275, 277, 281,	2, 10, 17, 25, 29, **40**, 77, 87, 92, 95, 98, 103, 107, 109, 110, 111, 112, 114, 115, **134**, 135, 139, 141, 144, 150, **156**, 165, 180, 181, **183, 200,** 205, 210, 213, 224, **227, 234,** 245, 258, 266, 267, 270, 279, 280, 285, 286, 289
Average-performing	1, 2, 12, 17, 38, 58, 71, 76, 92, 95, 103, 115, 139, 150, 170, 171, 173, 181, 195, 204, 237, 260, 261, 262, 266, 283, 288, 289, 297, 298	12, 19, 35, 38, 45, 53, 55, 56, 57, 60, 64, 71, 74, 76, 80, 81, 91, 100, 101, 104, 108, 129, 151, 153, 158, 162, 170, 171, 178, 196, 198, 203, 212, 215, 231, 232, 237, 253, 260, 262, 272, 275, 277, 281, 283, 284, 291, 297, 299
Low-performing	11, 36, 50, 59, 72, 80, 88, 90, 93, 100, 102, 104, 122, 136, 138, 151, 159, 161, 164, 174, 177, 209, 222, 225, 232, 250, 263, 285	1, 5, 9, 11, 14, 23, 28, 31, 36, 41, 50, 58, 68, 72,, 75, 88, 90, 93, 97, 100, 105, 106, 122, 136, 138, 143, 159, 161, 163, 164, 173, 174, 176, 177, 182, 195, 204, 209, 222, 225, 247, 250, 254, 261, 263, 288, 294, 298
Not rated	7, 9, 10, 14, 19, 22, 23, 25, 27, 28, 29, 35, 41, 45, 52, 53, 55, 56, 60, 62, 63, 64, 68, 75, 82, 87, 98, 105, 107, 108, 109, 110, 111, 112, 114, 129, 135, 141, 143, 144, 145, 153, 154, 160, 163, 165, 166, 167, 180, 182, 196, 198, 201, 203, 205, 210, 211, 213, 215, 217, 224, 231, 233, 245, 247, 254, 258, 267, 269, 270, 272, 273, 279, 280, 284, 286, 291, 294, 299	6, 7, 22, 27, 52, 59, 61, 62, 63, 82, 102, 120, 145, 154, 160, 166, 167, 201, 211, 217, 269, 273, 288

Note. Schools classified as high-performing in both years are indicated in bold.

C10

School Classifications in Mathematics Using Adjusted Value-added Approach

	2002	2003
High-performing	1, 19, 31, **38**, **41**, 56, 57, 59, 64, 71, 74, 75, **77**, **81, 101,** 102**, 103**, 106, **107, 108**, 112, 141, 150, **158,** 162, **165, 200**, 231, **234**, 254, 262, 266, 267, **277,** 291, 297	2, 10, 17, 19, 25, 35, **38,** 40, **41**, 45, **77**, 81, 92, 98, 101, 103, 104, 107, 108, 111, 134, 144, 151, 153, **158**, **165,** 182, 198, **200,** 215, **234,** 245, 253, 260, 272, 275, **277**, 279, 281, 283, 284, 285, 286
Average-performing	6, 9, 17, 23, 55, 60, 61, 87, 93, 95, 97, 98, 109, 111, 115, 120, 134, 135, 136, 153, 171, 174, 178, 180, 181, 196, 205, 209, 215, 217, 227, 253, 263, 275, 281, 283, 289,	1, 12, 14, 23, 29, 31, 53, 55, 56, 60, 64, 71, 87, 91, 95, 97, 105, 106, 109, 112, 129, 135, 141, 156, 170, 171, 178, 180, 181, 203, 204, 205, 210, 213, 224, 227, 231, 237, 263, 267, 270, 289, 291
Low-performing	2, 11, 12, 28, 36, 50, 58, 68, 72, 88, 90, 92, 100, 114, 122, 138, 151, 159, 161, 164, 173, 176, 177, 195, 204, 222, 225, 232, 237, 247, 250, 261, 284, 288, 294, 298	9, 11, 28, 36, 50, 58, 68, 72, 76, 80, 88, 90, 93, 100, 110, 115, 122, 136, 138, 143, 159, 161, 164, 173, 174, 176, 177, 195, 209, 222, 225, 232, 233, 247, 250, 254, 261, 262, 266, 294, 298, 299
Not rated	7, 10, 14, 22, 25, 27, 29, 35, 40, 45, 52, 53, 62, 63, 76, 80, 82, 91, 104, 105, 110, 129, 139, 143, 144, 145, 154, 156, 160, 163, 166, 167, 170, 182, 183, 198, 201, 203, 210, 211, 212, 213, 224, 233, 245, 258, 260, 269, 270, 272, 273, 279, 280, 285, 286, 299	6, 7, 22, 27, 52, 57, 59, 61, 62, 63, 74, 75, 82, 102, 114, 120, 139, 145, 150, 154, 160, 162, 163, 166, 167, 183, 196, 201, 211, 212, 217, 258, 269, 273, 280, 288, 297

Note. Schools classified as high-performing in both years are indicated in bold.

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Almond, P. J., Lehr, C., Thurlow, M. L., & Quenemoen, R. (2002). Participation in large-scale state assessment and accountability systems. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 341-370). Mahwah,NJ: Lawrence Erlbaum and Associates.

Baker, E. L., & Linn, R. L. (2004). Validity issues for accountability systems. In R. Elmore & S. Fuhrman (Eds.), *Redesigning accountability systems*. New York, NY: Teachers College Press.

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems. Policy Brief 5*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved March 3, 2004, from http://www.cse.ucla.edu/products/newsletters/polbrf54.pdf.

Ballou, D. (2002). Sizing up test scores. *Education Next, 2*, 10-15.

Barton, P. E., & Coley, R. J. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade. Policy Information Report*. Educational Testing Service. Retrieved October 11, 2003, from www.//ftp.ets.org/pub/res/growsch.pdf.

Betebenner, D. (2004). *An analysis of school district data using value-added methodology. CSE Report 622*. National Center for Research on Evaluation,

Standards, and Student Testing. Retrieved June 15, 2004, from www.cse.ucla.edu/products/reports_set.htm.

Bielinski, J., Thurlow, M. L., Minnema, J., & Scott, J. (2000). *How out-of-level testing affects the psychometric quality of test scores*. National Center on Educational Outcomes. Retrieved April 1, 2005, from www. education.umn.edu/NCEO/OnlinePubs/OOLT2.html.

Boudah, D. J., & Weiss, M. P. (2002). *Learning disabilities overview: Update 2002*. Retrieved January 3, 2005, from www.ericdigests.org/2002-4/learning-disabilities.html.

Bracey, G. W. (2000a). *Bail me out! Handling difficult data and tough questions about public schools*. Thousand Oaks,CA: Corwin Press, Inc.

Bracey, G. W. (2000b). *Thinking about tests and testing: A short primer in "Assessment Literacy"* American Youth Policy Forum. Retrieved February 11, 2005, from http://www.aypf.org/publicatons/BraceyRep.pdf.

Brualdi, A. (1999). *Traditional and modern concepts of validity.* ERIC/AE Digest. Retrieved January 18, 2004, from www.ed.gov/databases/ERIC_Digests/ed435714.html.

Bryk, A. (2003). No Child Left Behind Chicago-style. In P. E. Peterson & M. R. West (Eds.), *No Child Left Behind? The politics and practice of school accountability* (pp. 242-268). Washington, DC: Brookings Institution Press.

Carlson, D. E. (2000). *All students or the ones we taught?* Paper presented at the Council of Chief State School Officers Annual National Conference on Large-scale Assesment, Snowbird, UT.

Carlson, D. E., & Parshall, L. (1996). Academic, social, and behavioral adjustments for students declassified from special education. *Exceptional Children, 63*, 89-100.

Carnoy, E., & Loeb, S. (2004). Does external accountability affect student outcomes? A cross-state analysis. In R. Elmore & S. Fuhrman (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Center on Education Policy. (2005). *From the capitol to the classroom: Year 3 of the No Child Left Behind Act*. Retrieved March 19, 2005, from www.ctredpol.org/pubs/ncly3/press/cep-nclby3_21March2005.pdf.

Chief State School Officers. (2005). *Memorandum to Chief State Schools Officers*. CCSSO. Retrieved May 28, 2005, from http://www.ccsso.org/content/pdfs/Growthmemo.pdf.

Choi, K., Seltzer, M., Herman, J., & Yamachiro, K. (2004). *Children left behind and AYP in schools: Validation of AYP focusing on student progress and the distribution of student gains.* Paper presented at the American Educational Research Association, San Diego, CA.

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding Schools Accountable*. Washington, DC: Brookings Institution Press.

Coladarci, T. (2005). Adequate yearly progress, small schools, and students with disabilities: The importance of confidence intervals when making judgments about AYP. *Rural Special Education Quarterly, 24*(1), 40-47.

Coleman, J. (1966). *The Coleman report*. Harvard University. Briticannica Online. Retrieved March 4, 2005, from http://www.eb.com:180/cgi-bin/g?DocF=micro/702/16.html.

Consortium for Citizens with Disabilities. (2005). *Letter to US Department of Education*. Retrieved May 4, 2005, from http://www.c-c-d.org/EduTFLetter2Spellings.pdf.

Cowan, K. T. (2003). *The new Title I: The changing landscape of accountability*. Tampa, FL: Thompson Publishing Group, Inc.

Crane, J. (2002). *The promise of value-added testing*. Progressive Policy Institute. Retrieved February 17, 2004, from www.ppionline.org/documents/Value_Added_Testing.pdf.

CTB-McGraw Hill. (2001). *Terra Nova Technical Report*. Monterey, California: CTB-McGraw Hill.

CTB McGraw-Hill. (2004). *Technical quality*. CTB McGraw-Hill. Retrieved January 15, 2004, from http://www.ctb.com/.

DeBray, E., Parson, G., & Woodworth, K. (2001). Patterns of response in four high schools under state accountability policies in Vermont and New York.

. In S. Fuhrman (Ed.), *From Capitol to the Classroom: Standards-Based Reform in the States* (pp. 170-192). Chicago, IL: University of Chicago Press.

Doran, H. C. (2003). Adding value to accountability. *Educational Leadership, 61*(3), 55-59.

Doran, H. C., & Izumi, L. T. (2004). *Putting education to the test: A value-added model for California*. Pacific Research Institute. Retrieved June 8, from http://www.pacificresearch.org/pub/sab/educat/2004/Value_Added.pdf.

Education Commission of the States. (2003). The progress of education reform 2003: Closing the achievement gap (Vol. 4).

Education Commission of the States. (2004). *Report to the nation: State implementation of the NCLBA*. Education Commission of the States. Retrieved May 18, 2004, from www.ecs.org.html.Special/NCLB/.

Education for All Handicapped Children Act, U.S.C. § (1975).

Elliott, S. N., Ysseldyke, J. E., & Thurlow, M. L. (1998). What about assessment and accountability? *Teaching Exceptional Children, 31*(2), 3-24.

Elmore, R. (2004). Conclusion: The problem of stakes in performance-based accountability systems. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 320). New York: Teacher's College Press.

Feuer, M. J., Holland, P. W., Bertenthal, M. W., Cadelle-Hemphill, F., & Green, B. F. (Eds.). (1998). *Equivalency and linkages of educational tests*. Washington, D.C.: National Academies Press.

Fideler, E. F., Foster, E. D., & Schwartz, C. (2000). *The urban teacher challenge: Teacher demand and supply in the great city schools*. The Urban Teacher Collaborative. Retrieved March 25, 2005, from www.rnt.org/quick/utc.pdf.

Figlio, D. (2002). *Aggregation and accountability In No Child Left Behind: What will It take?* Paper presented at the conference sponsored by Thomas B. Fordham Foundation, Washington, DC.

Fuhrman, S. (2003). *Redesigning accountability systems for education* (No. RB-38). Philadelphia, PA: Consortium for Policy Research in Education (CPRE).

Fuhrman, S., & Elmore, R. (Eds.). (2004). *Redesigning accountability systems for education*. New York: New York: Teacher's College Press.

Gaddy, B., McNulty, B., & Waters, T. (2002). *The reauthorization of the Individuals with Disabilities Education Act: Moving toward a more unified system*. Mid-Continent Research for Education and Learning. Retrieved March 24, 2003, 2003, from http://www.mcrel.org/PDF/PolicyBriefs/5022PI_PBReauthorizationIDEA.pdf.

Goldhabar, D. (2001). Ways that states use assessment data. *Basic Education, 45*(6), 1-4.

Goldschmidt, P., & Choi, K. (2004). *Evaluation, validity assessments, and accountability.* Paper presented at the CCSSO ASR State Collaborative on Assessment and Student Standards (SCASS), Minneapolis, MN.

Goldstein, H. (1991). Better ways to compare schools. *Journal of Educational Statistics, 16*(2), 27-43.

Gong, B. (2002). *Designing school accountability systems: Towards a framework and process*. State Collaborative on Assessment and Student Standards, Council of Chief States Schools Officers. Retrieved February 6, 2005, from http://www.ccsso.org/content/pdfs/designing_school_acct_syst.pdf.

Gong, B. (2004). *Accountability Validity: Frameworks and examples.* Paper presented at the CCSSO ASR State Collaborative on Assessment and Student Standards (SCASS),, Minneapolis, MN.

Greene, J. P. (2002). The business model. *Education Next*, 20-23.

Hambleton, R. K., Impara, J., Mehrens, W., & Plake, B. S. (2000). *Psychometric review of the Maryland School Performance Assessment Program (MSPAP)*

Abell Foundation. Retrieved February 16, 2005, from http://www.marylandpublicschools.org/NR/rdonlyres/517D465A-F0B5-40DD-BB4B-E98EA716EC46/1581/Hambleton.pdf.

Hamilton, L. S., & Koretz, D. (2002). Tests and their use in test-based accountability systems. In S. P. Klein, B. Stecher & L. Hamilton (Eds.), *Making sense of test-based accountability in education*. Santa Monica, CA: RAND Publishing.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). *Does special education raise academic achievement for students with disabilities?* National Bureau of Economic Research. Retrieved February 20, 2002, from www.nber.org/papers/w6690.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2001). *Why public schools lose teachers*. Cambridge, MA:National Bureau of Economic Research. Retrieved August 15, 2005, from www.nber.org/ papers/w8599.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2002). Does special education raise academic achievement for students with disabilities? *Review of Economics and Statistics, 84*(4), 584-599.

Hanushek, E. A., & Raymond, M. E. (2002). The confusing world of educational accountability. *National Tax Journal, LIV*(2), 365-384.

Hanushek, E. A., & Raymond, M. E. (2003). Lessons about the design of state accountability systems. In M. R. West & P. E. Peterson (Eds.), *No Child Left Behind? The politics and practice of school accountability* (pp. 127-151). Washington, DC: Brookings Institution Press.

Heistad, D., & Spicuzza, R. (2000). *Measuring school performance to improve student achievement and to reward effective programs.* Paper presented at the American Educational Research Association, New Orleans, LA.

Herman, J. L. (2004). The effects of testing on instruction. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 142-166). New York, NY: Teacher's College Press.

Hill, R. (2001). *Issues related to the reliability of school accountability scores.* Retrieved March 12, 2003, from www.nciea.org.

Hill, R., & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. *Educational measurement: Issues and Practice, 22*(2), 12-21.

Huck, S. W. (2000). *Reading statistics and research*. New York, NY: Addison Wesley Longman, Inc.

Improving America's Schools Act, U.S.C. 20 § 8001 (1994).

Individuals with Disabilities Education Act, U.S.C. 20 § 1400 et seq. (1997).

Individuals with Disabilities Education Improvement Act of 2004, U.S.C. § 612 et seq. (2004).

Kane, T. J., & Staiger, D. O. (2002). Volatility in test scores: Implications for test based accountability systems. In D. Ravitch (Ed.), *Educational Policy 2002*. Washington, D.C.: Brookings Institution Press.

Kane, T. J., Staiger, D. O., & Geppert, J. (2001). *Assessing the definition of "Adequate Yearly Progress" in the House and Senate Education bills*. Retrieved August 15, 2003, from http://www.dartmouth.edu/~dstaiger/Papers/housesenate6.pdf.

Kane, T. J., Staiger, D. O., & Geppert, J. (2002). Randomly accountable. *Education Next*, 57-61.

Kean, M. H. (2004). Educational assessment in a reform context. In J. E. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 325-334). Greensboro, NC: CAPS Press.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas, and policy*. Thousand Oaks, CA: Corwin Press, Inc.

Koenig, J. A., & Bachman, L. F. (2004). *Keeping score for all: The effects of inclusion and accommodations on large-scale educational assessments*. Washington, DC: National Academy of Sciences.

Koretz, D. (1996). Using student assessments for educational accountability. In E. Hanushek & D. W. Jorgenson (Eds.), *Improving America's Schools*. Washington, D.C.: National Academy Press.

Koretz, D., & Barton, K. (2003). *Assessing students with disabilities: Issues and evidence (Technical Report 587)*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved April 15, 2004, from www.cse.ucla.edu.

Ladd, H. F. (1996). *Holding schools accountable: Performance based reform in education*. Washington, DC: Brookings Institution Press.

Ladd, H. F. (1999). The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review, 18*(1), 1-16.

Ladd, H. F. (2002). School based accountability systems: The promise and the pitfalls. *New Tax Journal, 54*(2), 385-400.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 1-15.

Linn, R. L. (2001, April). *The design and evaluation of educational assessment and accountability systems*. National Center for Research on Evaluation, Standards and Student Testing (CRESST). Retrieved September 15, 2003, from http://www.cse.ucla.edu/products/reports_set.htm.

Linn, R. L. (2004). Validity issues for accountability sytems. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47). New York, NY: Teachers College Press.

Linn, R. L., & Baker, E. L. (2002). Accountability systems: Implications of requirements of the "No Child Left Behind Act of 2001". *Educational Researcher, 31*(6), 3-16.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*(1), 29-36.

Lissitz, R. W., & Huynh, H. (2003). *Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school*

accountability. Retrieved March 28, 2005, from http://PAREonline.net/getvn.asp?v=8%n=10.

Mazzeo, C. (2002). Frameworks of state assessment policy in historical perspective. *Teacher's College Record, 103*(3), 367-397.

Mazzeo, C., Carlson, J. E., Voeklk, K., & Lurkus, A. D. (2000). Increasing the participation of special needs students in NAEP. *Education Statistics Quarterly, 2*(1), 157-161.

McCall, M., Kingsbury, G. G., & Olsen, A. (2004). *Individual growth and school success*. Northwest Evaluation Association. Retrieved December 28, 2004, from http://www.nwea.org/assets/research/national/individual%20growth%20and%20school%20success%20-%20complete%20report.pdf.

McCaul, E. J., & Schutz, P. N. (1991). Learning disability identification criteria: Impact of a state discrepancy in a rural state. *Teacher Education and Special Education, 14*, 116-120.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (1997). *Educating One and All: Students with Disabilities and Standards-Based Reform*. Washington, D.C.: National Academy Press.

McLaughlin, M. J., Embler, S., Hernandez, G., & Caron, E. (2005). No Child Left Behind and students with disabilities in rural and small schools. *Rural Special Education Quarterly, 24*(1), 32-39.

McLaughlin, M. J., Schofield, K., & Rhim, L. M. (1998). Educational reform: Issues for the inclusion of students with disabilities. In M. Coutinho & A. Repp (Eds.),

Inclusion: The intergration of students with disabilities. (pp. 37-58). Belmont, CA: Wadsworth.

McLaughlin, M. J., & Thurlow, M. L. (2003). Educational accountability and students with disabilities: Issues and challenges. *Educational Policy, 17*(4), 431-451.

Meyer, R. (1996). Value-added indicators of school performance. In E. Hanushek & D. W. Jorgenson (Eds.), *Improving America's Schools* (pp. 197-224). Washington, D.C.: National Academy Press.

Meyer, R. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16*(3), 283-301.

Moe, T. M. (2003). Politics, control, and the future of school accountability. In M. R. West & P. E. Peterson (Eds.), *No Child Left Behind? The politics and practice of accountability* (pp. 80-106). Washington, DC: Brookings Institution Press.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC.

Nelson, F. H., Rosenburg, B., & Kubic, M. (2004, October 15). *AYP status and school ratings in Maryland*. Retrieved October 15, 2005, from http://www.aft.org/topics/nclb/downloads/MDabstract.pdf.

No Child Left Behind, U.S.C. 20 § 6301 (2001).

O'Day, J. A. (2004). Complexity, accountability, and school improvement. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 293-329). New York: Teachers College Press.

Olson, J. F., & Goldstein, A. A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of*

recent progress (No. NCES 97-482). Washington, DC: National Center for Education Statistics.

Orfield, K. (2003, August 17, 2005). *Critical issue: NCLB option—choosing to change schools*. Retrieved November 9, 2005, from http://www.ncrel.org/sdrs/areas/issues/envrnmnt/famncomm/pa600.htm.

Paige, R. (2002). *Dear Colleague Letter*. U.S. Department of Education. Retrieved August 18, from http://www.ed.gov/News/Letters/020724.html.

Pizzuro, S. (2001). *The Individuals with Disabilities Education Act and nature of American politics: A handbook on public policy*. Retrieved August 5, 2002, from http://www.edrs.com/Members/ebsco.cfm?ED=ED455633.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Educational Testing Service. Retrieved 14 March, 2005, from www.ets.org/research/pic/angoff9.pdf.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307-335.

Ravitch, D. (2002). Testing and accountability, historically considered. In M. E. Williamson & H. J. Walberg (Eds.), *School Accountability*. Stanford, CA: Hoover Press.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teacher's College Record, 104*(8), 1525-1567.

Rubenstein, R., Stiefel, L., Schwartz, A. E., & Hadj Amor, H. B. (2004). Measuring school efficiency: What have we learned? In *Developments in School Finance*. Washington, DC: National Center for Educational Statistics (NCES).

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to student assessment. In J. Millman (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Sanela, M. (2002). *Value added assessment*. Evergreen Freedom Foundation. Retrieved December 18, 2004, from www.effwa.org/edcation_briefs.php.

Schneider, M., & Buckley, J. (2002). What do parents want from schools? Evidence from the Internet. . *Educational Evaluation and Policy Analysis, 24*(2), 133-144.

Sherry, M. (2005). *Missouri lowers testing goals*. The Kansas City Star. Retrieved May 13, 2005, from http://www.kansascity.com/mld/kansascity/news/10704934.htm

Shriner, J. (2000). Legal perspectives on school outcomes assessment for students with disabilities. *The Journal of Special Education, 33*, 232-239.

Stecher, B., Hamilton, L., & Gonzalez, G. (2003). *Working smarter to leave no child behind: Practical insights for school leaders*. RAND Corporation. Retrieved August 18, 2004, from www.rand.org.

Stevens, J., Estrada, S., & Parkes, J. (2000). *Measurement issues in the design of state accountability systems*. Retrieved April 21, 2003, from http://www.unm.edu/~jstevens/papers/AERA00.pdf.

Stone, J. E. (1999). Value-added assessment: An accountability revolution. In M. Kanstoroom & C. E. Finn (Eds.), *Better Teachers, Better Schools*. Washington, DC: Thomas B. Fordham Foundation.

Taylor, W. L., & Piche, D. M. (2002, 9 January). Will new school law really help? *USA Today,* p. A13.

Thompson, S., & Thurlow, M. L. (2002). *Universally designed assessments: Better tests for everyone*. National Center on Educational Outcomes. Retrieved 15 April 2005, from http://education.umn.edu/NCEO/OnlinePubs/Policy14.htm

Thurlow, M. L. (2000). Standards-based reform and students with disabilities: Reflections on a decade of change. *Focus on Exceptional Children, 33*(3), 1-15.

Thurlow, M. L. (2004). Biting the bullet: Including special-needs students in accountability systems. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Columbia Teacher's College Press.

Thurlow, M. L., McGrew, K. S., Tindal, G., Thompson, S. L., Ysseldyke, J. E., & Elliott, J. L. (2000). *Assessment accommodations research: considerations for design and analysis*. National Center for Educational Outcomes. Retrieved September 9, 2002, from http://education.umn.edu/nceo/OnlinePubs/Technical26.htm.

Thurlow, M. L., Quenemoen, R., Thompson, S., & Lehr, C. (2001). *Principles and characteristics of inclusive assessment and accountability systems*. National Center on Educational Outcomes. Retrieved April 4, 2004, from http://education.umn.edu/nceo/OnlinePubs/Synthesis40.html.

Tindal, G. (2002). Large-scale assessments for all students: Issues and options. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale asessment programs for all students*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Toothacker, L. E. (1993). *Multiple comparisons procedures: Quantitative applications in the social sciences*. Thousand Oaks, CA: Sage Publications.

US Department of Education. (1994). *OSEP Memorandum 95-9, 21 IDELR 1152 2 ECLPR 81*. Office of Special Education. Retrieved July 25, 2004, from http://www.ibwebs.com/oseplre.htm.

US Department of Education. (2001). *Twenty-third annual report to Congress*. Retrieved August 8, 2003, from http://www.ed.gov/about/reports/annual/osep/2001/index.html.

US Department of Education. (2005a). *Flexibility for states raising achievement for students with disabilities*. Retrieved May 14, 2005, from www.ed gov/policy/elec/guid/raising/disab-factsheet.html.

US Department of Education. (2005b). *Letter to Chief State School Officers*. Retrieved December 4, 2005, from http://www.ed.gov/policy/elsec/guid/secletter/051121.html.

Walberg, H. J. (2003). Real accountability. In P. E. Peterson (Ed.), *Our schools and our future: Are we still at risk?* Stanford, CA: Hoover Institution Press.

Walker, D. K., Singer, J. D., Palfrey, J. S., Orza, M., Wenger, M., & Butler, J. A. (1988). Who leaves and who stays in special education: A 2-year follow-up study. *Exceptional Children, 54*, 393-402.

West, M. R., & Peterson, P. E. (2003). The politics and practice of accountability. In M. R. West & P. E. Peterson (Eds.), *No Child Left Behind? The policits and practice of school accountability* (pp. 1-22). Washington, DC: Brookings Institution Press.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Xiaojing, K. (2004). *An investigation of motivation filtering in a statewide achievement testing program*. Xiaojing Kong, James Madison University

Retrieved April 8, 2005, from http://www.jmu.edu/assessment/wm_library/Wise_K12_motivation_filtering_(1).pdf.

Ysseldyke, J., & Bielinski, J. (2000). *Interpreting trends in the performance of special education students*. National Center on Education Outcomes. Retrieved July 7, 2003, from www.umn.edu/nceo/onlinepubs/techreport27.htm.

Ysseldyke, J., & Bielinski, J. (2001). *Critical questions to ask when interpreting or reporting trends in the large-scale test performance of students with disabilities*. Washington, DC: Council of Chief State School Officers.

Ysseldyke, J., & Bielinski, J. (2002). Effect of different methods of reporting and reclassification on trends in test scores for students with disabilities. *Exceptional Children, 68*(2), 189-200.

Ysseldyke, J., & Nelson, J. R. (2002). Reporting results of student performance on large-scale assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale Assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 467-483). Mahwah, NJ: Lawrence Erlbaum Associates.

Footnotes

[1] While missing scores can be imputed for students without prior achievement, the more common practice is to omit these students from the analyses.

[2] While federal laws as well as the Joint Standards on Testing do not make a distinction between accommodations and modifications, researchers differentiate between the two, basing their distinction on whether the alteration is believed to change the construct the assessment is measuring. For purposes of this paper the term accommodation is used to refer to any change in the standard administration of an assessment

[3] Although NCLBA contains a small market-based component offered through parental school choice, the model is based primarily on theories of standards and performance-based reforms.

[4] States may use also additional categories, such as below basic to further define academic performance.

[5] States were required under NCLBA to decide if mandatory sanctions applied only to schools receiving Title I funds, or if the sanctions applied to all schools. The state in this study chose to have the sanctions apply to all schools, regardless of Title I status.

[6] In a letter to the US Department of Education Dated April 11, 2004, MSDE requested changes to this policy. To date, these changes have neither been approved nor denied.

[7] Students with mental retardation were excluded from analyses because only five students identified with mental retardation had an assessment score

[8] Under the IDEA states are required to provide the federal government with a count of all students receiving special education services on December 1 of each year. This is considered the official enrollment data for schools. .

[9] Students whose home school was a hospital, private school, or who were home-school were excluded

[10] A reading composite could not be calculated because the vocabulary subtest score was not available.

[11] While the scale scores of the CTBS and CAT/5 are comparable, the percentile ranks of the two tests are not on the same scale and should not be used interchangeably or in longitudinal studies without equating.

[12] According to the state's accountability plan, students' scores are attributed to the school where they receive services and not the students' home school.

[13] According to the IDEA, there are thirteen disability categories. In this study, there were no students labeled as deaf-blind in any year, resulting in only twelve disability categories.

[14] California's policy states that the subgroups must equal at least 100 students, or 50 if this represents at least 15% of valid test scores.