

Statistical Properties of Accountability Measures Based on ACT's Educational Planning and Assessment System

Jeff Allen
Dina Bassiri
Julie Noble

Table of Contents

Abstract	vi
Introduction	1
<i>Terminology</i>	2
<i>Organization of Report</i>	3
Overview of Types of Accountability Models	4
Sample and Data	7
Status Models	12
<i>Reliability of Status Measures</i>	16
<i>Status Measures: Relationships with Prior Mean Academic Achievement and School Contextual Factors</i>	17
<i>Summary: Status Models</i>	20
Improvement Models	20
<i>Improvement Measures: Relationships with Prior Mean Academic Achievement and School Contextual Factors</i>	22
<i>Summary: Improvement Models</i>	23
Growth Models	24
<i>Growth Models: The Wright/Sanders/Rivers (WSR) Method</i>	25
<i>Growth Models: Simple Extrapolation Based on Vertical Scaling</i>	28
<i>Reliability of Growth Measures</i>	32
<i>Growth Measures: Relationships with Prior Mean Academic Achievement and School Contextual Factors</i>	33
<i>Aggregating Growth Measures</i>	34
<i>Summary: Growth Models</i>	34
Value-Added Models	35
<i>Value-Added Models: Estimating School Effects on ACT Scores</i>	37
<i>Value-Added Models: Estimating School Effects on EPAS Growth Trajectories</i>	41
<i>Uncertainty of Estimated School Effects</i>	46
<i>Reliability of Value-Added Measures</i>	50
<i>Value-Added Measures: Relationships with Prior Mean Academic Achievement and School Contextual Factors</i>	51
<i>Summary: Value-Added Models</i>	53
Case Examples: EPAS-Based Accountability Measures for Two High School Cohorts	54
<i>A High Poverty, High Minority High School</i>	54
<i>A Low Poverty, Low Minority High School</i>	58
Relation of EPAS-based Accountability Measures and College Enrollment and Retention Rates	60
<i>College Enrollment and Retention Data</i>	60
<i>Analysis of Aggregated College Enrollment and Retention Rates</i>	62
Discussion	66
References	73
Appendix A States and Locales of High School Cohorts Studied	77
Appendix B Projection Parameters from WSR Method for Projecting ACT Scores	78

List of Tables

TABLE 1. Summary of Types of Accountability Models Studied..... 6

TABLE 2. Advanced Organization of Growth and Value-Added Accountability Models..... 7

TABLE 3. Summary Statistics for High School Cohorts in Study Sample and Population..... 10

TABLE 4. Sample and Population Race/Ethnicity and Gender Breakdown..... 11

TABLE 5. Summary Statistics of Students’ ACT Scores..... 11

TABLE 6. EPAS College Readiness Benchmarks 12

TABLE 7. Distributions of PLAN Status Measures..... 13

TABLE 8. Distributions of ACT Status Measures 15

TABLE 9. Intercorrelations of Status Measures and Prior Mean Academic Achievement 16

TABLE 10. Autocorrelations of Status Measures 17

TABLE 11. Beta Weights for Predicting Status Measures..... 19

TABLE 12. Distributions of Improvement Measures..... 21

TABLE 13. Beta Weights for Predicting Improvement Measures..... 23

TABLE 14. Distributions of Growth Measures Based on WSR Growth Method..... 27

TABLE 15. Comparison of Observed and WSR Growth ACT Scores 28

TABLE 16. Distributions of Growth Measures Based on VP-Growth Method..... 29

TABLE 17. Comparison of Observed and VP-Growth ACT Scores 30

TABLE 18. Approximate Standard Errors of Measurement of Projected Scores 31

TABLE 19. Autocorrelations of Growth Measures..... 32

TABLE 20. Beta Weights for Predicting Growth Measures 33

TABLE 21. Distributions of Estimated School Effects on ACT Scores 39

TABLE 22. Distributions of Context-Adjusted Estimated School Effects on ACT Scores..... 40

TABLE 23. Intercorrelations of School Effects on ACT Scores..... 40

TABLE 24. Distributions of Estimated School Effects on EPAS Growth Trajectories..... 44

TABLE 25. Distributions of Context-Adjusted Estimated School Effects on EPAS Growth Trajectories 45

TABLE 26. Intercorrelations of School Effects on EPAS Growth Trajectories 46

TABLE 27. Classifications of School Effects 48

TABLE 28. Autocorrelations of Value-Added Measures..... 50

TABLE 29. Beta Weights for Predicting Value-Added Measures..... 52

TABLE 30. Accountability Measures for a High Poverty, High Minority High School 55

TABLE 31. Accountability Measures for a Low Poverty, Low Minority High School..... 59

TABLE 32. Descriptive Statistics for Measures Related to College Enrollment Rates 63

TABLE 33. Statistical Relationships of Accountability Measures and College Enrollment Rates 65

TABLE 34. Intercorrelations of Composite Accountability Measures and School Contextual Factors..... 68

List of Figures

FIGURE 1. Number of High School Cohorts Sampled Per State..... 9
FIGURE 2. Conceptual Model for Validating Accountability Measures..... 60

Abstract

Educational accountability has grown substantially over the last decade, due in large part to the No Child Left Behind Act of 2001. Accordingly, educational researchers and policymakers are interested in the statistical properties of accountability models used for NCLB, such as status, improvement, and growth models; as well as others that are not currently used for NCLB, such as value-added models. This study examines the statistical properties of accountability measures that are based on ACT's Educational Planning and Assessment System (EPAS). Utilizing data on 1,019 high school cohorts and over 70,000 students with test scores from three time points (8th, 10th, and 11th /12th grades), different types of accountability measures are contrasted and key statistical properties are discussed - including reliability, associations with prior mean academic achievement and school contextual factors, and associations with college enrollment and retention rates. Our findings highlight how status, improvement, growth, and value-added models can lead to different conclusions about a school's effectiveness. Unlike status, improvement, and growth models, value-added models attempt to isolate and measure the school's effect on student's learning. Thus, value-added measures have smaller associations with prior mean academic achievement and, by extension, school contextual factors such as poverty level and proportion of racial/ethnic minority students. This study also highlights the need for reporting the statistical uncertainty about estimates of schools' effects so that results can be properly interpreted.

Statistical Properties of Accountability Measures Based on ACT's Educational Planning and Assessment System

Introduction

Educational accountability has gained considerable attention in the United States, especially with the passage of the No Child Left Behind (NCLB) Act of 2001. Under NCLB, states and school districts must implement assessments each year in grades 3 through 8; high schools must administer assessments in reading / language arts, mathematics, and science for at least one grade level between grades 10 and 12. Schools that receive Title I funding must demonstrate “adequate yearly progress” towards reaching 100% proficiency by the 2013-14 academic year. The assessments must be aligned with the state’s academic content standards, and students’ progress towards proficiency must be reported annually. The definitions of and standards for proficiency vary substantially from state to state (Linn, 2006; NCES, 2007). While NCLB provides a framework for each state’s accountability system, each state has developed its own specific plans for implementation and many of the details are left to state and local educators to fill in (U.S. Department of Education, 2002), though the federal government retains the right to review and accept or reject each state’s plans.

Accountability systems have utility beyond meeting NCLB’s requirements. Historically, test-based accountability systems have been used to help clarify expectations for teaching and learning, monitor educational progress of schools and students, identify schools and programs that need improvement, and provide a basis for the distribution of rewards and sanctions to schools and students (Linn, 2006).

ACT’s Educational Planning and Assessment System (EPAS) is designed to guide and support schools, districts, and states in their efforts to improve students’ readiness for life after high school through a longitudinal approach to educational and career planning, assessment,

instructional support, and evaluation. EPAS assessment results are reported on a single score scale designed to inform students, parents, teachers, counselors, administrators, and policymakers about students' strengths and weaknesses. EPAS consists of EXPLORE (for eighth graders), PLAN (for tenth graders), and the ACT (for eleventh and twelfth graders). All three components of EPAS measure academic achievement, respective to the curriculum of the grade level for which it is intended.

In this study, we examine the statistical properties of different types of EPAS-based accountability measures and the implications of their use in evaluating schools. Different types of accountability measures are contrasted, including a discussion of each measure's reliability, relationships with prior mean academic achievement and school contextual factors, and validity for measuring the academic "effects" of schools. The analyses are based on a large sample of high school cohorts with students who took the EXPLORE, PLAN, and ACT tests in grade 8, 10, and 11/12, respectively.

Terminology

The terminology used to describe an accountability system varies considerably across entities (researchers, policymakers, and educators), causing confusion when policies are developed and results are communicated. So, in this section, we describe several terms that are used throughout this report. We use a fictitious school, "Lincoln High School," to give usage examples of each term.

First, we describe the terms *accountability system*, *accountability model*, and *accountability measure*. The term *accountability system* is used to refer to the overarching system of student assessment, implementation of accountability models, reporting and dissemination of accountability measures, and uses of these measures for decision making. Such

decisions could be considered *high-stakes* (e.g., school sanctions or rewards, teacher evaluation for promotion) or *low-stakes* (e.g., identification of areas in need of improvement, formative evaluation). A school's accountability system may be mandated by NCLB and its state, or may be unique to the school. For example, Lincoln High School's accountability system involves assessing 9th and 11th grade students in mathematics and reading and publicly reporting the proportion of students who are proficient in each subject and grade. The term *accountability model* is used to describe specific approaches for aggregating achievement (i.e., proportion proficient) or for measuring school effectiveness. The term *accountability measure* is used to describe the numeric descriptors produced by the *accountability model*. For example, the proportion of students who are proficient in mathematics and reading in grades 9 and 11 are two of the accountability measures used in Lincoln's accountability system. The mean gains in mathematics and reading scores from grade 9 to grade 11 are other examples of accountability measures that could be produced with Lincoln's accountability system.

Organization of Report

This report begins with an overview of the types of accountability models considered in this study, followed by a description of the sample of high school cohorts and students used in this study. Next, we describe how EPAS data can be used to generate status measures, improvement measures, growth measures, and value-added measures. For each of these types of accountability measures, we discuss their reliabilities and relationships with prior mean academic achievement and school contextual factors. Because improvement measures require data across cohorts, reliability cannot be easily assessed and so we only discuss relationships of these measures with prior mean academic achievement and school contextual factors. Then, examples of accountability models are given at two actual high schools in the sample – a high poverty,

high minority school and a low poverty, low minority school. Next, we examine how the EPAS-based accountability measures are related to high schools' aggregated college enrollment rates. The report concludes with a summary of findings and recommendations.

Overview of Types of Accountability Models

We now provide brief descriptions of the types of accountability models that are considered in this study. These include models that are commonly referenced under NCLB, including *status*, *improvement*, and *growth models*; as well as *value-added* models, which are not currently used for NCLB.

A *status* model is a type of accountability model that uses a single year's assessment results as an indicator of school performance (Goldschmidt & Choi, 2007). For example, the proportions of 10th graders in a given year who are proficient in mathematics and reading are examples of *status* measures. If decision rules (such as whether to reward or sanction the school) are linked to these status measures, the accountability system would be called a *status system*.

An *improvement model* is an accountability model that uses multiple years' assessment results at the same grade level to obtain projections of a school's status. For example, a high school's year 2014 projected proportions of 10th graders who are proficient in mathematics (where the projected values are based on current and past years' status) is an example of an *improvement* measure. Again, if decision rules are attached to improvement measures, we call the accountability system an *improvement system*. Improvement systems are consistent with NCLB's "adequate yearly progress" provision. Under this provision, schools must show that their status trend lends itself to 100% proficiency by the year 2014.

A *growth model* is an accountability model that uses two or more years of individual students' assessment results to obtain projections of the school's status (Goldschmidt & Choi,

2007). For example, a *growth measure* could be defined as the proportion of students who are projected to reach grade-level proficiency by grade 12, based upon their mathematics scores from 9th and 10th grade. In November 2005, U.S. Secretary of Education Margaret Spellings announced a Growth Model Pilot program to which states might submit proposals for accountability models as alternatives to status and improvement models (U.S. Department of Education, 2007). As of July 2008, eleven states (Tennessee, North Carolina, Delaware, Arkansas, Iowa, Florida, Ohio, Alaska, Arizona, Michigan, and Missouri) had their growth models approved (U.S. Department of Education, 2008).

A *value-added* accountability model is an accountability model that uses two or more years of individual students' assessment results to estimate how much a particular school has "added value" to their students' test scores (Rubin, Stuart, & Zanutto, 2004). Typically, value-added models relate students' test scores to background factors and, in some cases, school-level characteristics. An example of a value-added measure is a *mean school growth estimate* that is produced by a value-added model. As we will discuss later in this report, hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002) is one class of statistical models that can be used to implement value-added models. The overarching principle of a value-added accountability system is that schools should not be held accountable for students' levels of academic proficiency and background upon entry, but should be held accountable for adding "value," ensuring that students receive at least one year of growth for one year of schooling (Callender, 2004). Value-added models have been used to estimate teacher effects (Ballou, Sanders, & Wright, 2004), but in this report we only use them to estimate effects of high schools (i.e., school effectiveness). Value-added accountability measures are not currently accepted under NCLB, but are used for other purposes. For example, the Tennessee Value-Added Assessment System

(Ballou, Sanders, & Wright, 2004) is used to measure teacher effectiveness and to provide information to teachers, parents, and the public on how well schools are helping students learn.

In Table 1, we list the different accountability models, an example measure emanating from the model, the minimum data requirements for implementing the model, and whether the model is currently in compliance with NCLB.

TABLE 1

Summary of Types of Accountability Models Studied

Type of model	Example accountability measure	Data requirements	Used for NCLB
Status	Proportion of 10 th graders proficient in mathematics.	Assessment results from a single year.	Yes
Improvement	Year 2014 projected proficient of 10 th graders in mathematics.	Assessment results from multiple years on different cohorts of students.	Yes
Growth	Proportion of 10 th graders projected to become proficient in mathematics by 12 th grade.	Assessment results from multiple years on the same cohort of students.	Yes (Growth Model Pilot)
Value-Added	Number of mathematics score points attributed to a school, above or below what can be attributed to schools on average.	Assessment results from multiple years on the same cohort of students.	No

There are several variants of growth and value-added models that differ methodologically in how projected scores are obtained (growth models) and how school effects are estimated (value-added models). In this study, we examine two subtypes of growth models and four variants of value-added models. In Table 2, we list the different variants growth and value-added models, examples of the resulting accountability measure based on EPAS data, and an abbreviation that is later used when referring to the model. This table serves as a quick reference

guide for the reader if the terminology becomes cumbersome and it is difficult to distinguish between different types of growth and value-added models.

TABLE 2

Advanced Organization of Growth and Value-Added Accountability Models

Type of model	EPAS subtype	EPAS example	Abbreviation
Growth	11/12 th grade projected status based on Wright-Sanders-Rivers (WSR) method	Proportion projected to meet ACT College Readiness Benchmarks	WSR-growth
	11/12 th grade projected status based on vertical projection of 8 th and 10 th grade scores	Proportion projected to meet ACT College Readiness Benchmarks	VP-growth
Value-Added	School effect on ACT scores	Number of ACT score points attributed to a school, above or below what can be attributed to schools on average	ACT-VAM
	Context-adjusted school effect on ACT scores	Number of ACT score points attributed to a school, above or below what can be attributed to schools serving similar students, on average	ACT-CAVAM
	School effect on EPAS growth trajectory	Amount of students' level of growth attributed to a school, above or below what can be attributed to schools on average	EPAS-VAM
	Context-adjusted school effect on EPAS growth trajectory	Amount of students' level of growth attributed to a school, above or below what can be attributed to schools serving similar students, on average	EPAS-CAVAM

Sample and Data

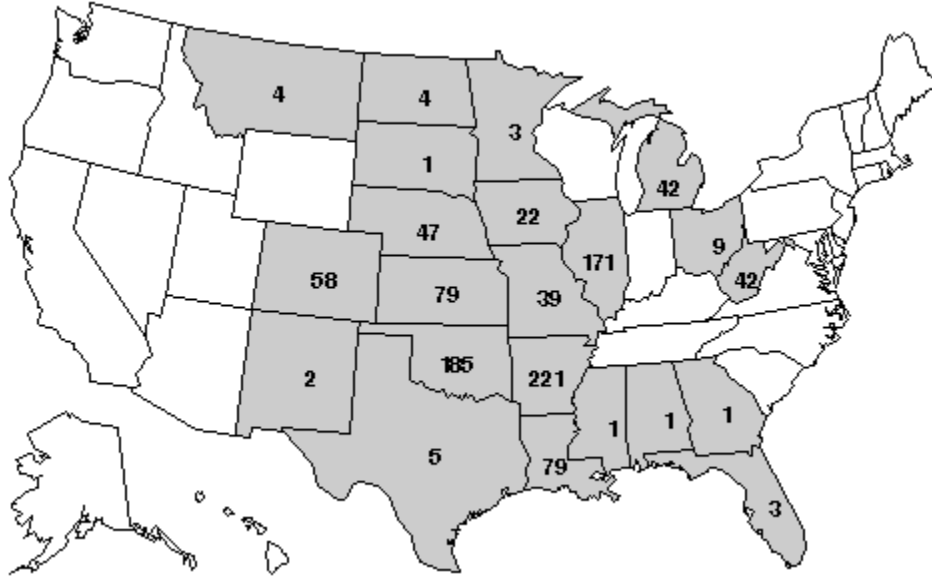
The data represent 485 high schools for which there were up to five cohorts of available data. In all, there were 1,019 cohort-by-high school combinations; on average, there were 2.1

cohorts per high school. To be included in our study sample, the proportion of students who took EXPLORE, PLAN, and the ACT must have been at least 0.50 for a given high school cohort, where PLAN and ACT were taken at the same high school. Here, *proportion tested* was defined as $N \div \text{Enroll}_{11}$, where N is the number of students who took all three assessments (EXPLORE, PLAN, ACT) and Enroll_{11} is the high school cohort's enrollment count as of 11th grade. With this inclusion criterion, the sample was restricted to high school cohorts where the majority of students were represented. As we discuss later, maximizing student representation is a crucial element of any accountability system.

Of the 485 high schools, 213 had one cohort that met the inclusion criterion, 124 had two, 68 had three, 46 had four, and 34 had five. Among the 1,019 high school cohorts, the median proportion tested was 0.57; the 25th percentile was 0.53 and the 75th percentile was 0.64. The mean sample size was 72; the median sample size was 40 with 25th percentile 23 and 75th percentile 90.

Figure 1 displays the frequency of the 1,019 high school cohorts, by state. Much of the sample comes from the Midwestern and south-central U.S, with little representation from the eastern and western states. This is due to the fact that most schools that use all three EPAS tests are from Midwestern and south-central states. The states with the most high school cohorts represented include Arkansas (221), Oklahoma (185), and Illinois (171).

FIGURE 1. Number of High School Cohorts Sampled Per State



In Appendix A, the cross-tabulations of high school cohort locale (large city, mid-size city, urban fringe of city, large town, small town, or rural) and state are given. To assess how well the sample represents the population of public high schools, we compared the sample to all high schools in the NCES Common Core of Data for 2004 (Sable, Thomas, & Sietsema, 2006). Relative to the population, the sample has more high school cohorts from rural (60% vs. 40%) and small town locales (20% vs. 11%); relative to the population, the sample has fewer high school cohorts from the urban fringe of a city (13% vs. 28%), mid-size cities (5% vs. 10%), and large cities (2% vs. 10%).

In Table 3, the high school cohorts are described in terms of enrollment size (grade 11 enrollment), poverty level (school's proportion of students eligible for free or reduced lunch), and proportion minority (school's proportion of students who are Black, American Indian, or Hispanic). Again, the sample can be compared to the general population of public high schools. In the sample, the average grade 11 enrollment is 128.3 (standard deviation=142.2, median=70).

The high school cohorts in the sample are somewhat smaller than the typical school in the population, where the average grade 11 enrollment is 176.8, with median 108. In the sample, the average poverty level is 0.32, with median 0.30. These are similar to the population average of 0.35 and median of 0.31. The sample's average proportion minority is 0.15, with median 0.07. The sample of high school cohorts has relatively fewer high-minority schools than the population, where the mean proportion minority is 0.31, with median 0.17.

TABLE 3

Summary Statistics for High School Cohorts in Study Sample and Population

Variable	Group	Mean	SD	Min	P₂₅	Med	P₇₅	Max
Grade 11 enrollment	Sample	128.3	142.2	9	40	70	157	806
	Population	176.8	184.6	0	35	108	270	1,346
Poverty level	Sample	0.32	0.19	0.00	0.17	0.30	0.43	0.99
	Population	0.35	0.25	0.00	0.16	0.31	0.50	1.00
Proportion minority	Sample	0.15	0.19	0.00	0.02	0.07	0.23	0.99
	Population	0.31	0.32	0.00	0.04	0.17	0.52	1.00

Note: $n=1,019$ high school cohorts, population total derived from 2004 Common Core of Data (Sable et al., 2006), min=minimum, $P_{25}=25^{\text{th}}$ percentile, med=median, $P_{75}=75^{\text{th}}$ percentile, max=maximum

In summary, the sample of high school cohorts is similar to the population of public high schools with respect to poverty level, but has relatively fewer large and high-minority schools. Later, we will discuss how the study's findings could be impacted by these differences.

Nested within the 1,019 high school cohorts are 73,240 students. Table 4 compares the gender and racial/ethnic group breakdowns for the sample and population of 11th grade public high school students nationally. White students are over-represented in the sample (77% vs. 62%), while Hispanic (3% vs. 17%), African American (7% vs. 15%), and Asian American students (2% vs. 5%) are under-represented. A portion of the sample (7%) has unknown or missing race/ethnicity. Females are slightly overrepresented (53% vs. 50%).

TABLE 4**Sample and Population Race/Ethnicity and Gender Breakdown**

Race/ethnicity	Gender			Total	
	Female	Male	Missing	Sample	Population
African American	3,034 8%	2,014 6%	14 2%	7%	15%
American Indian	692 2%	621 2%	4 1%	2%	1%
Asian American	741 2%	724 2%	7 1%	2%	5%
Hispanic	1,245 3%	1,073 3%	12 2%	3%	17%
White	30,140 78%	26,257 78%	141 24%	77%	62%
Other	938 2%	732 2%	7 1%	2%	<1%
Missing	2,048 5%	2,393 7%	403 69%	7%	0%
Sample total	53%	46%	1%	100%	
Population total	50%	50%	0%		100%

Note: $n = 73,240$, population total derived from 11th grade totals in 2004
Common Core of Data (Sable et al., 2006)

In Table 5, the student sample is described with respect to ACT test scores. The average ACT scores range from 20.8 for Mathematics to 21.4 for Reading. Nationally, for 2008 ACT-tested high school graduates, the mean scores ranged from 20.6 for English to 21.4 for Reading (ACT, 2008); the student sample appears to be quite typical of ACT-tested populations in terms of academic achievement.

TABLE 5**Summary Statistics of Students' ACT Scores**

Test	Mean	SD	P _{Bench}
English	21.1	5.7	0.73
Mathematics	20.8	5.0	0.41
Reading	21.4	5.9	0.53
Science	21.1	4.5	0.28

Note: $n = 73,240$, P_{Bench}=proportion meeting College Readiness Benchmark

Status Models

Currently, NCLB provides sanctions for schools whose proficiency rate, or proportion of students who meet or exceed a certain proficiency cutoff score, is below a targeted level. Each student's test scores are dichotomized and the proficiency rate is the simple proportion of students at or above the specified proficiency cutoff score. A proficiency rate is an example of a status measure.

Researchers at ACT found the scores on the ACT that correspond to a 50% chance of obtaining a "B" or higher grade in four standard first-year college courses: English Composition, College Algebra, Social Science, and Biology (Allen & Sconing, 2005). These *College Readiness Benchmarks* provide a way to dichotomize EPAS test scores in a manner that is meaningful with respect to college readiness. The Benchmarks for EXPLORE and PLAN are the scores corresponding to a 50% chance of meeting the corresponding ACT Benchmarks. The College Readiness Benchmarks are given in Table 6.

TABLE 6

EPAS College Readiness Benchmarks

Subject	EXPLORE	PLAN	ACT
English	13	15	18
Mathematics	17	19	22
Reading	15	17	21
Science	20	21	24

In this report, EPAS test scores are dichotomized using the corresponding College Readiness Benchmarks. In practice, states often dichotomize scores into "proficient" and "not proficient" according to state standards and achievement-level descriptions (i.e., below basic, basic, proficient, beyond proficient). But, for the sake of constructing status measures for this report, we use the College Readiness Benchmarks as cutoffs for proficiency. Though state

proficiency standards vary considerably in difficulty (NCES, 2007), it is generally true that the Mathematics, Reading, and Science Benchmarks are more difficult to meet than state proficiency standards. Because all high school students in Illinois and Colorado take the ACT in 11th grade, data from these states provide a basis to compare the difficulty of states' proficiency standards to that of the College Readiness Benchmarks. For example, in 2003, approximately 88% and 68% of 8th graders in Colorado met the state proficiency standards in reading and mathematics, respectively; In Illinois, approximately 65% and 54% of 8th graders met the state's standards in reading and mathematics, respectively (NCES, 2007). Later, within the same Colorado cohort, 47% and 37% met the ACT College Readiness Benchmarks in Reading and Mathematics, respectively; within the same Illinois cohort, 47% and 38% met the ACT College Readiness Benchmarks in Reading and Mathematics, respectively (ACT, 2007a; ACT, 2007b). Thus, for these two states, we see that the ACT Benchmarks are more difficult to meet than the states' proficiency levels in reading and mathematics.

Table 7 summarizes the distributions of the 10th grade (PLAN) status measures for the 1,019 high school cohorts. The median proportion of students meeting the English Benchmark is 0.83; the other median proportions are 0.39 (Mathematics), 0.58 (Reading), and 0.23 (Science). Clearly, the Science Benchmark is the hardest to meet and the English Benchmark is the easiest.

TABLE 7

Distributions of PLAN Status Measures

Subject	Proportion of students meeting benchmark				
	Min	P ₂₅	Med	P ₇₅	Max
English	0.00	0.75	0.83	0.88	1.00
Mathematics	0.00	0.28	0.39	0.50	0.86
Reading	0.08	0.49	0.58	0.67	1.00
Science	0.00	0.15	0.23	0.31	0.67

Note: $n = 1,019$ high school cohorts

From Table 7 it is also apparent that there is great variation across high school cohorts in the proportion of students meeting the PLAN Benchmarks. For example, for the Mathematics Benchmark, the 25th percentile is 0.28 and the 75th percentile is 0.50. The minimum is 0.00 and the maximum is 0.86. High school cohorts with small sample sizes are more likely to have extreme proportions; this is simply due to the fact that proportions based on small sample sizes have greater sampling error. The standard error of a proportion is $\sqrt{\frac{p(1-p)}{n}}$, where p is the proportion and n is the sample size. So, for example, the standard error of a proportion of 0.50 with $n=20$ is 0.112, whereas the standard error with $n=200$ is .035. Because of the inverse relationship of sample size and standard errors, one must interpret status measures for small high school cohorts with great caution. Moreover, as we will discuss later, standard errors of accountability measures should be reported, especially when the measures are used as the basis for rewarding or leveling sanctions against a school.

Table 8 summarizes the distributions of the 11th/12th grade (ACT) status measures for the 1,019 high school cohorts in the sample. The median proportion of students meeting the English Benchmark is 0.71; the other median proportions are 0.36 (Mathematics), 0.51 (Reading), and 0.23 (Science). As is the case with the PLAN Benchmarks, the Science Benchmark is the hardest to meet and the English Benchmark is the easiest.

TABLE 8**Distributions of ACT Status Measures**

Subject	Proportion of students meeting benchmark				
	Min	P₂₅	Med	P₇₅	Max
English	0.00	0.62	0.71	0.79	1.00
Mathematics	0.00	0.25	0.36	0.46	0.89
Reading	0.00	0.41	0.51	0.59	0.96
Science	0.00	0.15	0.23	0.32	0.62

Note: n = 1,019 high school cohorts

Comparing Table 8 with Table 7, we see that the median proportions meeting the College Readiness Benchmarks in English and Reading dropped from grade 10 to grade 11/12 (0.83 to 0.71 and 0.58 to 0.51, respectively).

Table 9 contains the intercorrelations of the PLAN and ACT status measures, as well as measures of prior mean academic achievement (proportion meeting EXPLORE Benchmarks). It is apparent that same-subject correlations tend to be higher. For example, the correlation of the PLAN and ACT Mathematics status measures is 0.84, but the correlations of the PLAN Mathematics status measure with those for ACT English and Reading are 0.56 and 0.66, respectively. Also, we see that PLAN and ACT status measures are strongly correlated with prior mean academic achievement. This suggests that high school status measures are heavily influenced by prior mean academic achievement.

TABLE 9**Intercorrelations of Status Measures and Prior Mean Academic Achievement**

Proportion meeting Benchmark in...	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
EXPLORE											
1. English	1.00										
2. Math	0.63	1.00									
3. Reading	0.76	0.67	1.00								
4. Science	0.56	0.66	0.67	1.00							
PLAN											
5. English	0.72	0.53	0.67	0.45	1.00						
6. Math	0.59	0.79	0.64	0.67	0.56	1.00					
7. Reading	0.65	0.61	0.69	0.57	0.69	0.64	1.00				
8. Science	0.54	0.67	0.64	0.67	0.54	0.74	0.66	1.00			
ACT											
9. English	0.69	0.51	0.65	0.50	0.74	0.56	0.69	0.55	1.00		
10. Math	0.58	0.75	0.62	0.66	0.58	0.84	0.64	0.71	0.63	1.00	
11. Reading	0.65	0.60	0.71	0.61	0.65	0.66	0.73	0.66	0.74	0.68	1.00
12. Science	0.56	0.68	0.65	0.68	0.55	0.73	0.68	0.74	0.62	0.79	0.74

Note: $n = 1,019$ high school cohorts

Reliability of Status Measures

Because many of the high schools in the sample have multiple cohorts of data, the reliability of each status measure can be examined. Table 10 contains the autocorrelations of the status measures for adjacent cohorts (one year apart), as well as for cohorts that are 2 and 3 years apart. Earlier, we discussed how status measures are less reliable for cohorts with smaller sample sizes. In lieu of this problem, the correlations in Table 10 are weighted according to the average sample size (across cohorts) for each high school. As expected, the correlations are greater for adjacent cohorts and decrease as the time between cohort increases. For example, the correlation of grade 11/12 mathematics status (proportion meeting the ACT Mathematics Benchmark) is 0.83 for adjacent cohorts, 0.80 for cohorts that are two years apart, and 0.70 for cohorts that are three years apart. The correlations in Table 10 suggest that the status measures tend to be repeatable: Schools that score high one year will likely score high the next year.

TABLE 10**Autocorrelations of Status Measures**

Proportion meeting Benchmark in...	Years between cohorts		
	1	2	3
PLAN			
English	0.65	0.55	0.55
Mathematics	0.78	0.77	0.70
Reading	0.63	0.62	0.60
Science	0.64	0.62	0.56
ACT			
English	0.74	0.76	0.64
Mathematics	0.83	0.80	0.73
Reading	0.73	0.74	0.71
Science	0.75	0.74	0.68

Note: $n=422$ high schools for 1 year between cohorts, 279 for 2 years, 161 for 3 years.

Status Measures: Relationships with Prior Mean Academic Achievement and School Contextual Factors

One of the principles of NCLB is to “set expectations for annual achievement based on meeting grade-level proficiency, not on student background or school characteristics” (U.S. Department of Education, 2007). So, for example, a school in a high-poverty area with a high proportion of non-native English speaking students would be expected to perform as well as a school in an affluent area with 100% native English speaking students. Numerous studies have shown that aggregate school achievement is strongly related to school poverty level and minority concentration (e.g., Howley, Strange, & Bickel, 2000; Linn, 2001). Therefore, it is not surprising that accountability systems that are based on status measures are perceived as unfair. Critics of status systems argue that status measures reflect contextual factors that are beyond the school’s control, such as entering student achievement and poverty level. Rather than a sound measure of school effectiveness, they believe that status measures are more a reflection of the background of the students served by the school. Ballou, Sanders, and Wright (2004) write: “Holding teachers

and administrators accountable for student outcomes without regard for differences in student background is manifestly unfair and, in the long run, counter-productive. Such policies will alienate educators, making it more difficult to staff schools serving the neediest population.”

In this report, we examine the associations of accountability measures with factors that are outside the school’s control.

In Table 11, we show how school characteristics contribute to the prediction of status measures (proportion meeting Benchmarks in grades 10 and 11/12). We present beta weights (standardized regression coefficients) obtained from regressing each status measure on selected predictor variables using a multiple linear regression model. The beta weights tell us each characteristic’s association with the status measures, beyond that explained by other school characteristics and prior mean academic achievement (mean number of EXPLORE benchmarks met). We consider two sets of models: In the first, prior mean academic achievement is not used as a predictor variable; in the second, prior mean academic achievement is used.

TABLE 11**Beta Weights for Predicting Status Measures**

PLAN or ACT Benchmark	Grade 11 enrollment	Proportion tested	Poverty level	Proportion minority	Mean number of EXPLORE Benchmarks met
Model 1: School characteristics					
PLAN					
English	-0.01	-0.10	-0.30	-0.29	
Mathematics	0.05	0.03	-0.39	-0.27	
Reading	0.08	-0.02	-0.31	-0.25	
Science	0.12	0.04	-0.31	-0.21	
ACT					
English	0.08	-0.14	-0.35	-0.26	
Mathematics	0.10	0.02	-0.42	-0.24	
Reading	0.10	-0.07	-0.34	-0.28	
Science	0.14	-0.03	-0.40	-0.21	
Model 2: School characteristics + prior mean academic achievement					
PLAN					
English	-0.03	-0.08	-0.06	-0.10	0.62
Mathematics	0.02	0.05	-0.12	-0.05	0.69
Reading	0.06	0.00	-0.04	-0.04	0.68
Science	0.10	0.06	-0.03	0.01	0.69
ACT					
English	0.06	-0.12	-0.13	-0.08	0.56
Mathematics	0.07	0.03	-0.17	-0.04	0.63
Reading	0.08	-0.06	-0.09	-0.08	0.64
Science	0.11	-0.01	-0.14	-0.02	0.64

Note: $n=1,019$ high school cohorts. The status measure is the proportion of students meeting the PLAN or ACT College Readiness Benchmark.

From Table 11, we see that prior mean academic achievement is the strongest predictor of status measures. Poverty level and proportion minority do not appear to have much influence on these status measures, once prior mean academic achievement is accounted for. However, when prior mean academic achievement is not accounted for, we see that poverty level and proportion minority are inversely related to status measures. Thus, high-poverty and high-racial/ethnic minority schools would be more likely to be sanctioned under a status system if

mean entering student achievement level is not accounted for. These findings lend support to the argument that status measures reflect the entering achievement level of the students served by the school.

Summary: Status Models

- The status measures we studied are highly correlated across content areas and across years.
- Status measures are strongly associated with students' entering achievement levels. High-poverty and high-minority schools are more likely to be sanctioned in a status system.

Improvement Models

An improvement measure is derived using the trend of status measures over two or more years for the same grade level. For example, suppose Lincoln High School's proportion of proficient students in grade 10 mathematics was 0.25 in 2002, 0.30 in 2003, and 0.35 in 2004. Then, the improvement in the proportion of proficient students was 0.05 from 2002 to 2003 and 0.05 from 2003 to 2004. Improvement measures have gained popularity under NCLB because they can determine whether schools are making adequate yearly progress (AYP). To determine if Lincoln High School is making AYP, we must extrapolate Lincoln's proportion of proficient students in grade 10 mathematics to the year 2014, at which time all schools must have 100% of students at proficiency. Using a simple linear extrapolation (i.e., annual improvement of 0.05), Lincoln's expected status in 2014 is 0.85. Because this proportion is less than 1.00, Lincoln is not making AYP. In order to show AYP in 2005, Lincoln must show improvement of 0.065 over their 2004 status of 0.35 (e.g., proficiency rate of at least 41.5%). In this example, we assumed a

simplistic linear extrapolation of status; in practice, other extrapolation methods have been used to show AYP.

In this section, improvement models are illustrated for the high schools in the sample. Because improvement models require at least two cohorts of data at the same grade level, only the 272 high schools in the sample with multiple cohorts of data are considered. For each of these high schools, projected status for the year 2014 is calculated using a simple linear extrapolation based on the observed status measures (proportion meeting the College Readiness Benchmarks) from 2002 to 2006. In Table 12, we summarize the distributions of improvement measures for the 272 high schools.

TABLE 12

Distributions of Improvement Measures

PLAN or ACT benchmark	Proportion of students projected to meet benchmark							Prop. of schools making AYP
	Mean	SD	Min	P ₂₅	Med	P ₇₅	Max	
PLAN								
English	0.71	0.33	0.00	0.55	0.82	1.00	1.00	0.29
Mathematics	0.41	0.38	0.00	0.00	0.35	0.76	1.00	0.14
Reading	0.58	0.40	0.00	0.12	0.68	1.00	1.00	0.29
Science	0.36	0.37	0.00	0.00	0.27	0.65	1.00	0.14
ACT								
English	0.61	0.36	0.00	0.31	0.69	0.96	1.00	0.23
Mathematics	0.40	0.35	0.00	0.00	0.37	0.68	1.00	0.10
Reading	0.48	0.37	0.00	0.08	0.48	0.82	1.00	0.19
Science	0.28	0.31	0.00	0.00	0.18	0.47	1.00	0.07

Note: $n = 272$ high schools. The improvement measure is the projected proportion of students meeting the PLAN or ACT College Readiness Benchmark in 2014.

Because the median of the projected status measures ranges from 0.18 (ACT Science) to 0.82 (PLAN English), one could conclude that the typical school is not projected to be at 100% proficiency by 2014 in any subject area at any of the two grade levels examined. However, some schools are projected to be 100% proficient at certain time points in certain subject areas. For

example, 14% of the schools are projected to be 100% proficient in 10th grade science. Only 9 schools (3%) are projected to reach 100% proficiency in all four subject areas at grade 10; thus, the overwhelming majority of schools would not be viewed as making AYP in all subject areas. The proportions of schools making AYP in grades 11/12 are especially small, ranging from 0.07 in science to 0.23 in English.

Improvement Measures: Relationships with Prior Mean Academic Achievement and School Contextual Factors

As with status measures, improvement measures can be assessed by examining their associations with prior mean academic achievement and high school characteristics. The beta weights in Table 13 show how each of these characteristics contributes to the prediction of each improvement measure. The results indicate that projected 10th and 11th/12th grade status appears to be influenced most by prior mean academic achievement (mean number of EXPLORE Benchmarks met). Comparing the beta weights in Table 13 with those in Table 11, we see that projected status is related to prior mean academic achievement and school characteristics in much the same way as ordinary status measures. This is not surprising, because *projected* status is mostly determined by *current* status.

Because projected status in the year 2014 is consonant with NCLB's AYP provision, one could argue that an accountability system based on improvement models unfairly penalizes high schools that begin with lower-achieving students. Indeed, one of the most common criticisms leveled against NCLB is that the "adequate yearly progress" provision disproportionately identifies certain types of schools as failing (Choi, Goldschmidt, & Yamashiro, 2005). As we will demonstrate later, school poverty level and proportion minority have much smaller

associations with value-added accountability measures, which are designed to measure the effects that schools have on academic performance.

TABLE 13**Beta Weights for Predicting Improvement Measures**

PLAN or ACT Benchmark	Grade 11 enrollment	Proportion tested	Poverty level	Proportion minority	Mean number of EXPLORE Benchmarks met
PLAN					
English	0.05	-0.04	-0.12	0.05	0.28
Mathematics	-0.10	-0.12	0.03	-0.05	0.34
Reading	0.08	-0.03	0.02	0.04	0.35
Science	-0.08	-0.08	0.06	-0.07	0.40
ACT					
English	0.09	-0.17	-0.04	0.01	0.40
Mathematics	-0.05	-0.02	-0.08	0.06	0.44
Reading	0.03	-0.05	0.07	-0.01	0.40
Science	-0.06	-0.03	-0.09	-0.01	0.34

Note: $n=272$ high school cohorts. The improvement measure is the projected proportion of students meeting the PLAN or ACT College Readiness Benchmark in 2014.

Summary: Improvement Models

- Projected status in the year 2014 based on EPAS data is an improvement measure consistent with NCLB's AYP provision.
- Most of the high schools in this study are not projected to reach 100% proficiency by 2014 in all four subject areas, and are therefore not making AYP.
- Like status measures, improvement measures are influenced by entering student achievement levels. High-poverty and high-minority schools are more likely to be sanctioned in an improvement system that does not adjust for students' entering achievement level.

Growth Models

It is possible for a state to employ growth measures when it has multiple years of test data for individual students in each school. Under the U.S. Department of Education's Growth Model Pilot Program, growth measures can be used to demonstrate AYP. Individual students within a school are making adequate yearly progress if their scores are projected to be at or above the proficiency level within a set time frame (e.g., by 12th grade). Then, the school's AYP status is determined by the percentage of students making AYP: Schools are meeting AYP if the percentage of their students making AYP projects to 100% in 2014. Goldschmidt & Choi (2007) classify these growth models according to the way in which scores are projected and the way in which different levels of growth are awarded points. The first method uses students' current test scores and projects their scores three years into the future using the state's average growth (the mean three-year growth currently observed in the state). The second method also uses students' current scores and projects their scores three years into the future using their current estimated growth (based upon two or more years of data). The third method, referred to as "value tables" by Goldschmidt & Choi (2007), awards points according to student movement along proficiency levels from one year to the next. For example, movement from "basic" to "proficient" might be awarded 100 points and movement from "below basic" to "basic" might be awarded 75 points.

In our analysis of growth measures, we examine the first and second methods. We will use scores from 8th grade (EXPLORE) and 10th grade (PLAN) to obtain projected 11th grade ACT scores. The first method we examine is based on the projection methodology used by the state of Tennessee for the Growth Model Pilot Program (Wright, Sanders, & Rivers, 2005). This methodology does not require the assessment scores used for the projection to be *vertically scaled*. Vertical scaling is a process of placing scores from two or more tests on the same scale

when those tests differ in difficulty and content but are similar in the construct measured. The second method we examine assumes that the EXPLORE, PLAN, and ACT tests are vertically scaled, and projected ACT scores are obtained for individual students based on their PLAN scores and growth from EXPLORE to PLAN.

Growth Models: The Wright/Sanders/Rivers (WSR) Method

Wright, Sanders, and Rivers (2005) developed a methodology for obtaining students' projected scores using prior test scores. This methodology, which we refer to as "the WSR method," is used by Tennessee for NCLB's Growth Modeling Pilot Program and has some important features, including:

- 1) It does not assume vertical scaling, although vertically-scaled assessments could be used.
- 2) The projected score is obtained as a function of possibly several prior test scores.
- 3) Not all students need to have all prior test scores; hence the WSR method accommodates missing and fragmented data.
- 4) The projected scores are interpreted as the score that a student would be expected to make, assuming that the student has an *average* schooling experience in the future. Hence, the WSR method is the most consistent with the first method described above (student projections based on state average growth).

The basic formula for obtaining projected ACT scores under the WSR method is given in Equation 1. We refer to this model using the abbreviation "WSR-growth."

Equation 1: WSR-Projected ACT scores

$$Y = M_Y + b_1(X_1 - M_1) + b_2(X_2 - M_2) + \dots + b_p(X_p - M_p)$$

In Equation 1, Y represents the projected ACT score. The mean within-school average test scores are given by $M_Y, M_1, M_2, \dots, M_p$, where p represents the number of prior test scores used for

projecting. The *projection parameters* include these means as well as the coefficients b_1, b_2, \dots, b_p associated with the difference between a student's scores and the school means. Wright, Sanders, and Rivers (2005) describe how these projection parameters are obtained; for brevity, we omit these technical details.

We used the WSR method to obtain projected ACT scores based on all observed EXPLORE and PLAN scores: Each projected ACT score is a function of eight prior test scores (in our sample, we have no missing EXPLORE or PLAN scores). In actual practice, the projection parameters are obtained using prior years' data (using students who have both the response variable Y and the predictor variables X) and then applied to current data (students who have X , but not Y) to obtain projected scores. To simulate actual practice, we took a simple random sample of 25% from each high school cohort and used the sample to obtain the projection parameters. We then applied the projection parameters to obtain projected ACT scores for the remaining 75% of the students. In Appendix B, we report the projection parameters that were used for each of the four ACT scores.

Projected ACT scores can form the basis of determining AYP. A student has made AYP if their projected scores meet or exceed the proficiency cutoffs, which in our case are given by the ACT College Readiness Benchmarks. In Table 14, we summarize the distributions of the growth measures (projected ACT proficiency) for the 1,019 high school cohorts in the sample. Consistent with ACT status measures (Table 8), we see that the English Benchmark is much easier to attain than the Science Benchmark; the median proportion projected to meet the English Benchmark is 0.74; the median proportion projected to meet the Science Benchmark is 0.18. If the projected ACT scores based on the WSR method are used to determine AYP, none of the

high school cohorts in the sample would be considered 100% proficient in mathematics, reading or science.

TABLE 14

Distributions of Growth Measures Based on WSR Growth Method

Subject	Proportion of students projected to meet ACT Benchmark				
	Min	P ₂₅	Med	P ₇₅	Max
English	0.00	0.63	0.74	0.83	1.00
Mathematics	0.00	0.20	0.32	0.43	0.85
Reading	0.00	0.40	0.52	0.62	0.93
Science	0.00	0.10	0.18	0.27	0.67

Note: $n = 1,019$ high school cohorts, 55,500 students (approximately 75% of sample). The growth measure is the proportion of students projected to meet the ACT Benchmark in 2014.

It is interesting to examine how well the WSR-projected ACT scores match the actual observed scores. The projected and observed scores are compared in Table 15. ACT English scores are slightly under-predicted (projected mean of 20.9, observed mean of 21.1), Mathematics scores are slightly under-predicted (20.6 vs. 20.8), and Science scores are slightly under-predicted (21.0 vs. 21.1). In Table 15, we also present the proportion of students whose WSR-projected ACT score is within three score points of their observed ACT score (Pw_3). These proportions ranged from 0.69 for Reading to 0.82 for Mathematics. The WSR-projected ACT scores are highly correlated with actual scores (0.85 for English, 0.85 for Mathematics, 0.80 for Reading, and 0.79 for Science). It is also interesting to note that the standard deviations of the projected scores are smaller than the standard deviations of the observed scores. The projected scores are obtained by a regression equation (Equation 1) and the smaller standard deviations are partly a consequence of the inability of the regression model to explain 100% of the variance in ACT scores.

TABLE 15**Comparison of Observed and WSR Growth ACT Scores**

Subject	ACT scores				Projected ACT scores				Difference				
	Mean	SD	P ₂₅	P ₇₅	Mean	SD	P ₂₅	P ₇₅	Mean	SD	P ₂₅	P ₇₅	P _{w3}
English	21.1	5.7	17	25	20.9	4.9	17	24	0.2	3.0	-2	2	0.77
Mathematics	20.8	5.0	17	24	20.6	4.2	18	23	0.2	2.7	-2	2	0.82
Reading	21.4	5.9	17	26	21.4	4.8	18	25	0.1	3.5	-2	2	0.69
Science	21.1	4.5	18	24	21.0	3.6	18	23	0.1	2.8	-2	2	0.81

Note: $n = 55,500$ students, P_{w3} =proportion of projected ACT scores that are within 3 score points of the actual ACT score

Growth Models: Simple Extrapolation Based on Vertical Scaling

A growth model that utilizes the vertical scaling of the EXPLORE, PLAN, and ACT tests allows for projected ACT scores that are easy to compute and more transparent than those obtained using WSR-growth. Projected ACT scores can be calculated based on PLAN scores and growth between EXPLORE and PLAN. Because EXPLORE is usually given in the fall of 8th grade, PLAN in the fall of 10th grade, and the ACT in the spring of 11th grade or fall of 12th grade, we make the simplifying assumption that the assessments are equally spaced in time. Later, we discuss the implications for accountability models of the time spacing of the assessments. Using the assumptions of equal time spacing, we obtain projected ACT scores based on a straight-line trajectory of EXPLORE and PLAN scores, as given in Equation 2. The abbreviation “VP-growth” is used when referring to this method for generating projected ACT scores.

Equation 2: Vertically-Projected ACT scores

$$ACT_{projected} = PLAN + (PLAN - EXPLORE) = 2PLAN - EXPLORE$$

In rare cases, Equation 2 can result in a projected ACT score falling outside the ACT score range (1-36); when this occurs, the projected score is set to 1 or 36. Another assumption underlying Equation 2 is that score gains between EXPLORE and PLAN are expected to be the

same as those between PLAN and the ACT. If in fact score gains between EXPLORE and PLAN typically exceed those between PLAN and the ACT, the resulting projected ACT scores will be too large. Similarly, if the reverse is true, the projected ACT scores will be too small.

Table 16 summarizes the distributions of the growth measures generated using VP-growth for the 1,019 high school cohorts in the sample. Consistent with the ACT status measures (Table 8), we see that the English Benchmark is easier to attain than the Science Benchmark: The median proportion projected to meet the English Benchmark is 0.73 and the median proportion projected to meet the Science Benchmark is 0.21. If projected ACT scores are used to determine AYP, none of the high school cohorts in the sample would be considered 100% proficient in mathematics, reading, or science.

TABLE 16

Distributions of Growth Measures Based on VP-Growth Method

Subject	Proportion of students projected to meet ACT Benchmark				
	Min	P ₂₅	Med	P ₇₅	Max
English	0.00	0.65	0.73	0.80	1.00
Mathematics	0.00	0.25	0.35	0.44	0.76
Reading	0.08	0.36	0.44	0.52	0.85
Science	0.00	0.14	0.21	0.28	0.53

Note: $n = 1,019$ high school cohorts. The growth measure is the proportion of students projected to meet the ACT Benchmark in 2014.

In Table 17, the accuracy of the vertically-projected ACT scores is assessed by comparing them to the observed ACT scores. ACT English scores are slightly over-predicted (projected mean of 21.2, observed mean of 21.1), Mathematics scores are more over-predicted (21.2 versus 20.8), Reading scores are under-predicted (21.5 versus 20.6), and Science scores are slightly under-predicted (21.1 versus 21.0). The underlying assumption that students' scores

gains from EXPLORE to PLAN are expected to be the same as score gains from PLAN to ACT may not be true, which may explain the discrepancies in the means of the observed and vertically-projected ACT scores. We found that vertically-projected ACT scores are highly correlated with actual scores; the correlations are 0.68 for English, 0.71 for Mathematics, 0.56 for Reading, and 0.55 for Science. However, these correlations are noticeably smaller than those observed for the WSR-growth scores. It is also interesting to note that the standard deviations of the VP-growth scores are the same or larger than those of the observed ACT scores. As we later discuss, this is partly due to large standard errors of measurement of the VP-growth scores.

TABLE 17

Comparison of Observed and VP-Growth ACT Scores

Subject	ACT scores				Projected ACT scores				Difference				
	Mean	SD	P ₂₅	P ₇₅	Mean	SD	P ₂₅	P ₇₅	Mean	SD	P ₂₅	P ₇₅	P _{w3}
English	21.1	5.7	17	25	21.2	5.8	17	25	-0.2	4.6	-3	3	0.55
Mathematics	20.8	5.0	17	24	21.2	5.8	17	24	-0.4	4.2	-3	2	0.62
Reading	21.5	5.9	17	26	20.6	6.3	16	25	0.8	5.7	-3	5	0.45
Science	21.1	4.5	18	24	21.0	4.5	18	23	0.2	4.3	-3	3	0.60

Note: n = 73,240 students, P_{w3}=proportion of projected ACT scores that are within 3 score points of the actual ACT score

In Table 17, we also present the proportion of students whose VP-growth score is within three score points of their observed ACT score (P_{w3}). These proportions are significantly smaller than those observed for the WSR-growth scores, suggesting that the WSR-growth scores are more accurate. The proportions for the VP-growth (vs. WSR-growth) scores are 0.55 (0.77) for English, 0.62 (0.82) for Mathematics, 0.45 (0.69) for Reading, and 0.60 (0.81) for Science. One reason for these discrepancies is that the WSR-growth scores are based on eight prior test scores while the VP-growth scores are based on two prior test scores. Another reason for the discrepancies is that VP-growth scores have standard errors of measurement (SEM) that are

considerably larger than the corresponding SEMs of EXPLORE, PLAN, and ACT scores. In Equation 3 and Equation 4 the SEM of WSR-growth and VP-growth scores are derived as a function of the SEMs of PLAN and EXPLORE scores.

Equation 3: Standard Error of Measurement of WSR-growth Scores

$$SEM_{WSR-Projected\ ACT} = \sqrt{\sum_{i=1}^8 b_i^2 SEM^2(X_i)}$$

where X_1, X_2, \dots, X_8 represent the eight PLAN and EXPLORE scores and b_1, b_2, \dots, b_8 represent the corresponding regression coefficients of the WSR method.

Equation 4: Standard Error of Measurement of VP-growth Scores

$$SEM_{Vertically-Projected\ ACT} = \sqrt{4SEM_{PLAN}^2 + SEM_{EXPLORE}^2}$$

Table 18 contains the approximate SEMs for the projected ACT scores, as well as SEMs for the observed scores.

TABLE 18

Approximate Standard Errors of Measurement of Projected Scores

Subject	Observed			WSR-projected ACT	Vertically-projected ACT
	EXPLORE	PLAN	ACT		
English	1.61	1.67	1.71	1.05	3.71
Mathematics	1.63	1.83	1.47	1.17	4.01
Reading	1.51	2.18	2.18	1.03	4.61
Science	1.41	1.60	2.00	0.76	3.50

Note: EXPLORE SEM values are the mean SEM across two test forms administered to grade 8 students (ACT, 2007c). PLAN SEM values are the mean SEM across four test forms (ACT, 1999). ACT SEM values are the median SEM across six ACT administrations in 2005-2006 (ACT, 2006).

For English, Mathematics, and Reading, the SEM for the VP-growth score is more than double the SEM for ACT score; for Science it is 75% larger. The large SEMs of VP-growth scores are a

consequence of projecting scores from two other scores that are measured with error. This problem is present whenever projections are based on prior scores and prior growth; it is not unique to projections that are based on EPAS test scores. WSR-growth scores, on the other hand, have SEMs that are smaller than those for observed ACT scores.

Reliability of Growth Measures

Table 19 contains the autocorrelations of the growth measures (proportion projected to meet ACT Benchmark) for adjacent cohorts (one year apart), as well as for cohorts that are two and three years apart. The correlations are weighted according to the average sample size (across cohorts) for each high school.

TABLE 19

Autocorrelations of Growth Measures

Proportion projected to meet ACT Benchmark in	Years between cohorts		
	1	2	3
<i>WSR-growth</i>			
English	0.72	0.62	0.59
Mathematics	0.77	0.75	0.66
Reading	0.71	0.63	0.59
Science	0.74	0.70	0.62
<i>VP-growth</i>			
English	0.57	0.49	0.43
Mathematics	0.68	0.65	0.62
Reading	0.57	0.57	0.50
Science	0.53	0.42	0.35

Note: $n=422$ high schools for 1 year between cohorts, 279 for 2 years, 161 for 3 years.

As expected, the correlations are larger for adjacent cohorts and decrease as the years between cohort increases. Clearly, the growth measures generated from WSR-growth scores are more reliable than the growth measures based on VP-growth scores. For example, the correlation of

the English growth measure ranges from 0.72 (one year apart) to 0.59 (three years apart) when the WSR method is used to derive projected ACT scores, but ranges from 0.57 (one year apart) to 0.43 (three years apart) when VP-growth scores are used. Comparing these correlations to those for status measures (Table 10), we see that the autocorrelations for the growth measures based on the WSR method are smaller, but comparable to those for status measures.

Growth Measures: Relationships with Prior Mean Academic Achievement and School

Contextual Factors

Table 20 contains beta weights obtained from regressing each growth measure on prior mean academic achievement (mean number of EXPLORE benchmarks met) and school characteristics using a multiple linear regression model.

TABLE 20

Beta Weights for Predicting Growth Measures

Growth model and ACT Benchmark	Grade 11 enrollment	Proportion tested	Poverty level	Proportion minority	Mean number of EXPLORE Benchmarks met
<i>WSR-growth</i>					
English	-0.03	-0.06	-0.01	-0.08	0.77
Mathematics	0.05	0.04	-0.09	0.03	0.77
Reading	0.04	-0.03	0.00	-0.02	0.81
Science	0.11	0.05	-0.05	0.08	0.76
<i>VP-growth</i>					
English	0.02	-0.06	-0.08	-0.10	0.46
Mathematics	0.06	0.02	-0.14	-0.09	0.47
Reading	0.13	0.00	-0.06	0.00	0.49
Science	0.11	0.04	-0.04	-0.08	0.36

Note: $n = 1,019$ high school cohorts. The growth measure is the proportion of students projected to meet the ACT Benchmark in 2014.

The beta weights indicate that the growth measures are strongly related to prior mean academic achievement. The growth measures based on WSR-growth tend to have beta weights that are

larger than those based on VP-growth; this is probably because the growth measures based on VP-growth are measured with greater error, leading to greater attenuation of the beta weights. Because both growth measures appear to be influenced by prior mean academic achievement, high-poverty and high-minority schools that have lower entering achievement levels are more likely to be sanctioned under a growth system that does not adjust for students' entering achievement level.

Aggregating Growth Measures

Up to this point, our examination of growth measures has focused on using students' prior test scores to obtain projected scores, and then using the projected scores to determine the proportion of students making AYP (based on the ACT College Readiness Benchmarks). This use of prior information is consistent with the current requirements of NCLB under the Growth Model Pilot Program. As we will show, students' prior test scores can also be used to produce aggregated growth measures, or *mean growth estimates*. In this report, we categorize mean growth estimates as *value-added* accountability measures.

Summary: Growth Models

- EPAS scores can be used to calculate growth measures for high school cohorts with two or more test scores obtained longitudinally on individual students.
- AYP is determined based on the proportion of projected test scores that meet or exceed the proficiency cutoff (or ACT College Readiness Benchmarks).
- We examined two methods for calculating projected scores: the Wright/Sanders/Rivers (WSR) Method, which does not require vertical scaling; and the vertical-projection method, which does. The WSR method is currently used by the state of Tennessee for the

NCLB Growth Model Pilot Program and has some features that make it particularly useful for NCLB, including the ability to accommodate missing data on students' prior test scores and the ability to obtain projections based on the "average schooling experience." Because the WSR method accommodates missing data, it can be used to obtain projected ACT scores from EXPLORE and/or PLAN scores. The vertical-projection method, on the other hand, requires both EXPLORE and PLAN scores.

- Under both types of growth models, none of the cohorts would be considered 100% proficient in mathematics, reading or science. Hence, no cohorts are projected to reach 100% proficiency in all subject areas.
- VP-growth scores have much larger standard errors of measurement, relative to WSR-growth scores. Related to this, WSR-growth scores have greater consistency over time and larger correlations with observed ACT scores.
- Like status and improvement measures, measures obtained from both types of growth models are influenced by prior mean academic achievement. High-poverty and high-minority schools are more likely to be sanctioned in a growth system that does not adjust for students' entering achievement level.

Value-Added Models

The fundamental purpose of value-added models is to isolate and estimate the effects of teachers, schools, and/or academic programs. Because status, improvement, and growth (projection) models do not account for students' entering academic proficiency or contextual factors such as student and school-level poverty level, policymakers have expressed interest in value-added models as a means to measure school and teacher effectiveness for high-stakes (i.e., as the basis for rewards or sanctions) and low-stakes (i.e., to improve practice or identify

teachers' and schools' strengths and weaknesses) accountability. Experts disagree on the extent to which value-added measures truly measure a school's effectiveness. But, they agree that value-added models can at least be used to produce descriptors of school effectiveness that are more meaningful than those produced by status, improvement, and growth models (Amrein-Beardsley, 2008).

The current principles of NCLB are not compatible with using value-added models as a means of measuring school effectiveness. The "bright line" principles of NCLB conflict with the philosophy of value-added models, most notably the principle that expectations for annual achievement are based on meeting grade-level proficiency, not on student background or school characteristics (U.S. Department of Education, 2007). Even though value-added measures are not currently used for NCLB reporting, there is a growing interest in using them for other purposes, such as evaluating teacher performance (Ballou, 2002) and improving school practice (Hershberg, Simon, & Lea-Kruger, 2004).

In our examination of value-added modeling, we use two general methods: The first method estimates the effect of schools on ACT scores, explicitly controlling for EXPLORE scores as covariates in a regression model. This method only requires EXPLORE and ACT scores and does not utilize the vertical scaling of EPAS test scores. The second method requires EXPLORE, PLAN, and ACT scores and estimates the effect of schools on EPAS growth trajectories; that is, the degree to which attending a particular school affects students' growth from grade 8 to grade 10 to grades 11/12. The second method utilizes the vertical scaling of EPAS test scores. For each method, we examine two approaches: one estimates school effects irrespective of contextual factors, and the second estimates school effects, adjusted for student-level factors such as family income and race/ethnicity, and school-contextual factors such as

poverty level and proportion of racial/ethnic minority students in the school. Later, we discuss arguments for and against adjusting value-added measures for contextual factors.

Value-Added Models: Estimating School Effects on ACT Scores

The first model is given in Equation 5. In this model, the four EXPLORE subject area test scores are covariates (X_1, X_2, X_3, X_4) , the school effect is denoted τ , and ε is the residual error for the regression model. The school effects and residual errors are assumed to be normally distributed and independent with mean 0 and unknown variances. This model can be fit for each of the four ACT subject tests, resulting in estimated school effects on students' academic performance in English, mathematics, reading, and science. Later, this model is referred to as the "ACT-VAM" model.

Equation 5: Value-Added Model for Deriving Estimated School Effect on ACT Scores

$$ACT_{score} = \beta_0 + \sum_{p=1}^4 \beta_p X_p + \tau + \varepsilon$$

ACT-VAM is a special case of a hierarchical linear model (Raudenbush & Bryk, 2002) and can be fit using statistical software packages such as HLM[®] or SAS[®]. The estimated school effect can be interpreted as an estimate of a school's contribution to students' academic performance, adjusted for their incoming performance level (EXPLORE scores). The school effect can also be interpreted here as the number of ACT score points attributable to a school, above and beyond what can be attributed for the average school. For example, if $\tau = 0.8$ for school A, then the number of ACT score points attributed to school A is 0.8 *more* than what could be expected of an average school. If $\tau = -0.5$ for school B, then the number of ACT score points attributed to school B is 0.5 *less* than what could be expected of an average school. To adjust for contextual factors, Equation 5 is easily extended by including the contextual factors as

additional covariates; this context-adjusted model is given in Equation 6 and is referred to as the “ACT-CAVAM” model.

Equation 6: Value-Added Model for Deriving Context-Adjusted Estimated School Effect on ACT Scores

$$ACT_{score} = \beta_0 + \sum_{p=1}^4 \beta_p X_p + \sum_{q=1}^7 \lambda_q B_q + \sum_{r=1}^4 \theta_r S_r + \tau + \varepsilon$$

In this model, the terms X_1, X_2, X_3, X_4 , τ , and ε are described as in Equation 5. Race/ethnicity and family income are introduced as student-level covariates (B_1, B_2, \dots, B_7). Additionally, grade 11 enrollment, proportion of students tested, school poverty level, and proportion of racial/ethnic minority students are introduced as school-level covariates (S_1, S_2, S_3, S_4). The estimated school effect from this model can be interpreted as an estimate of the school’s contribution to students’ academic performance, adjusted for their incoming performance level (EXPLORE scores), students’ family income and race/ethnicity, and school size, proportion of students tested, poverty level, and proportion of racial/ethnic minority students. Note that race/ethnicity is a seven-category nominal variable (African American, American Indian, Asian/Pacific Islander, Caucasian/White, Hispanic, Other, Missing) and so requires six dummy-coded covariates (B_1, B_2, \dots, B_6). Family income (B_7) is a ten-category ordinal variable that is treated as continuous. For about 24% of the student sample, family income is missing. We imputed family income with a multiple linear regression model using EXPLORE scores, race/ethnicity, and school-level characteristics as predictor variables.

Table 21 summarizes the distributions of the value-added measures generated by ACT-VAM (Equation 5). Interestingly, there is greater variation in school effects on ACT English score than there is on ACT Science score (SD=0.91 for English and 0.48 for Science). This suggests that there is greater consistency across high schools in their influence on science

performance relative to English performance. Under the assumed value-added model, the “average” school effect is always 0. The 25th and 75th percentiles of the estimated school effects can give us a rule of thumb of what constitutes a “good” score for a high school cohort and what constitutes a “poor” score. For example, only 25% of the high school cohorts have an English effect larger than 0.61; 0.61 could be considered a good score for the number of ACT English score points that could be attributed to a high school, over and above what could be expected of an “average” high school. Similarly, -0.33 could be considered a poor score for the number of ACT Science score points that could be attributed to a high school.

TABLE 21

Distributions of Estimated School Effects on ACT Scores

Subject	Estimate of school effect on ACT score					
	Min	P ₂₅	Med	P ₇₅	Max	SD
English	-2.62	-0.61	-0.02	0.61	2.74	0.91
Mathematics	-2.47	-0.51	-0.01	0.50	2.22	0.75
Reading	-1.79	-0.40	0.01	0.37	2.24	0.59
Science	-1.57	-0.33	0.00	0.31	1.74	0.48

Note: n = 1,019 high school cohorts

Table 22 summarizes the distributions of the context-adjusted value-added measures generated by ACT-CAVAM (Equation 6). The distributions of these measures are very similar to those of the ACT-VAM measures, but there is slightly less variation in the context-adjusted measures. Because the contextual factors (student’s family income and race/ethnicity, school’s grade 11 enrollment, proportion of students tested, poverty level and proportion of racial/ethnic minority students) explain some of the variation in ACT scores across high school cohorts, there is less to be attributed to the school itself. Hence, the standard deviations of the context-adjusted school effects are smaller than the corresponding standard deviations of the unadjusted school effects.

TABLE 22**Distributions of Context-Adjusted Estimated School Effects on ACT Scores**

Subject	Estimate of school effect on ACT score					
	Min	P ₂₅	Med	P ₇₅	Max	SD
English	-2.59	-0.60	-0.03	0.57	2.76	0.86
Mathematics	-2.20	-0.48	-0.01	0.44	2.32	0.69
Reading	-1.79	-0.36	0.03	0.34	1.74	0.55
Science	-1.44	-0.28	0.01	0.28	1.34	0.43

Note: $n = 1,019$ high school cohorts

Table 23 contains the intercorrelations of the estimated school effects on ACT scores. Of particular interest are the large correlations between the context-adjusted and unadjusted school effects on ACT scores. The correlation of the ACT-VAM and ACT-CAVAM English measure is 0.93; the correlations for the other subject areas are 0.92 (mathematics), 0.93 (reading), and 0.91 (science).

TABLE 23**Intercorrelations of School Effects on ACT Scores**

Estimated school effect on ...	1.	2.	3.	4.	5.	6.	7.	8.
1. English	1.00							
2. Math	0.56	1.00						
3. Reading	0.70	0.56	1.00					
4. Science	0.62	0.67	0.76	1.00				
5. English (context-adjusted)	0.93	0.45	0.61	0.55	1.00			
6. Math (context-adjusted)	0.52	0.92	0.47	0.59	0.53	1.00		
7. Reading (context-adjusted)	0.65	0.42	0.93	0.67	0.69	0.47	1.00	
8. Science (context-adjusted)	0.58	0.52	0.67	0.91	0.63	0.60	0.73	1.00

Note: $n=1,019$ high school cohorts

These correlations suggest that a school's value-added scores (estimated effects on ACT scores) are not likely to be influenced much by whether or not contextual factors are statistically controlled. In other words, schools that are considered above average using the context-adjusted model will most likely be considered above average using the non-context-adjusted model.

Moreover, they suggest that value-added measures are less influenced by contextual factors, a proposition that we examine later in greater detail. The value-added measures are also correlated across subject areas, suggesting that high school cohorts that score well in one area will likely score well in other areas.

Value-Added Models: Estimating School Effects on EPAS Growth Trajectories

A three-level hierarchical linear model was used to model EPAS growth trajectories. This model utilizes all three EPAS components (EXPLORE, PLAN, and ACT) and is different from the first value-added model we discussed, which used only EXPLORE score as the measure of prior achievement (covariate) to predict ACT score. At the first level of this model, we assume that students' EPAS scores can be explained with an intercept (initial level of academic achievement) and a slope (rate of change in level of academic achievement); this model is given in Equation 7.

Equation 7: Level 1 of Hierarchical Model for EPAS Growth Trajectories

$$Y_{ij} = \gamma_{0ij} + TIME_{ij}\gamma_{1ij} + e_{ij}$$

In Equation 7, Y_{ij} represents the EPAS score at the t^{th} time ($t=1$ for EXPLORE, $t=2$ for PLAN, $t=3$ and $t=4$ for the ACT for students with scores from both 11th and 12th grades; for 99% of the students in our sample, one ACT score was used) for the i^{th} student belonging to the j^{th} high school cohort. The variable $TIME$ represents the number of years since the first measure (EXPLORE) was obtained; thus, $TIME$ is coded as 0 (EXPLORE), 2 (PLAN), 3.5 (ACT taken in 11th grade), or 4 (ACT taken in 12th grade). The parameters γ_{0ij} and γ_{1ij} represent the student-specific intercept and slope, respectively. The residual error term is given by e_{ij} . At the second

level of this model, we assume that the mean of students' intercepts and slopes vary by high school cohort.

Equation 8: Level 2 of Hierarchical Model for EPAS Growth Trajectories

$$\begin{aligned}\gamma_{0ij} &= \lambda_{0j} + r_{0ij} \\ \gamma_{1ij} &= \lambda_{1j} + r_{1ij}\end{aligned}$$

In Equation 8, λ_{0j} and λ_{1j} represent the mean intercept and slope, respectively, at the j^{th} high school cohort and r_{0ij} and r_{1ij} are normally distributed error terms. Thus, students' intercepts and slopes are assumed to be random deviations from the high school cohort's mean intercept and mean slope. Finally, at the third level of this model, we assume that high school cohorts' mean intercepts and slopes are random deviations from an overall intercept and slope, as in Equation 9.

Equation 9: Level 3 of Hierarchical Model for EPAS Growth Trajectories

$$\begin{aligned}\lambda_{0j} &= \mu_0 + s_{0j} \\ \lambda_{1j} &= \mu_1 + s_{1j}\end{aligned}$$

In Equation 9, s_{0j} and s_{1j} represent deviations from the overall intercept and slope, respectively, for the j^{th} high school cohort. Thus, s_{1j} is the value that represents how much the j^{th} high school cohort contributed to students' academic growth.

We attempted to fit the three-level hierarchical model using the SAS[®] MIXED procedure. However, due to the size of the data set and the complexity of the model, there was insufficient memory to fit the model. Thus, we considered a different approach that retained the distinctive features of the three-level hierarchical model (i.e., student-specific and high school cohort-specific intercepts and slopes), but was less computationally intensive. From Equation 7, γ_{1ij} (and γ_{0ij}) can be estimated by regressing EPAS test scores on *TIME* for each student, yielding

$\hat{\gamma}_{1ij}$ (and $\hat{\gamma}_{0ij}$). Using these least-squares regression estimates and combining Equation 8 and Equation 9, the model for student slope is approximated with Equation 10.

Equation 10: Approximation to Three-Level Hierarchical Model

$$\hat{\gamma}_{1ij} = \mu_1 + s_{1j} + r_{1ij}$$

This model is of the same form as the ACT-VAM model (see Equation 5). Thus, the three-level model can be approximated for each of the four EPAS subject area tests, resulting in estimates of the high school cohorts' effects on student growth in English, mathematics, reading, and science.

We observed that students' initial level of academic performance was positively related to their growth in academic performance. For example, the correlation of EXPLORE English score and change in English score from EXPLORE to ACT was 0.11 and the correlation of EXPLORE Mathematics score and change in Mathematics score from EXPLORE to ACT was also 0.11. This suggests that expected growth depends on level of initial achievement and that EXPLORE scores (X_1, X_2, X_3, X_4) should be used as covariates in the model for the least-squares slope estimate ($\hat{\gamma}_{1ij}$). The model for deriving estimated school effects on EPAS growth trajectories is given in Equation 11 - this model is later referred to as the "EPAS-VAM" model.

Equation 11: Value-Added Model for Deriving Estimated School Effects on EPAS Growth Trajectories

$$\hat{\gamma}_{1ij} = \mu_1 + \sum_{p=1}^4 \beta_p X_p + s_{1j} + r_{1ij}$$

To adjust for contextual factors, Equation 11 is extended by including additional covariates (student-level and school-level); the context-adjusted model is given in Equation 12 and is later referred to as the "EPAS-CAVAM" model.

Equation 12: Value-Added Model for Deriving Context-Adjusted Estimated School Effects on EPAS Growth Trajectories

$$\hat{\gamma}_{1ij} = \mu_1 + \sum_{p=1}^4 \beta_p X_p + \sum_{q=1}^7 \lambda_q B_q + \sum_{r=1}^4 \theta_r S_r + s_{1j} + r_{1ij}$$

The additional covariates and parameters from Equation 12 are the same as those described in Equation 6.

In Table 24, we summarize the distributions of the EPAS-VAM measures for the 1,019 high school cohorts in the sample. The variation in school effects on EPAS growth trajectories was greatest for English (standard deviation of 0.23) and smallest for science (standard deviation of 0.13). This result is consistent with the results observed earlier for the school effects on ACT scores (Table 21); both results suggest that there is greater variability in schools' effects on performance in English relative to performance in science.

TABLE 24

Distributions of Estimated School Effects on EPAS Growth Trajectories

Subject	Estimate of school effect on EPAS growth trajectories					
	Min	P ₂₅	Med	P ₇₅	Max	SD
English	-0.67	-0.15	0.00	0.15	0.66	0.23
Mathematics	-0.65	-0.14	-0.01	0.13	0.64	0.20
Reading	-0.47	-0.10	0.00	0.09	0.55	0.15
Science	-0.43	-0.09	0.00	0.08	0.42	0.13

Note: $n = 1,019$ high school cohorts

Under both types of value-added models, the mean school effect is 0. The 25th and 75th percentiles of the estimated school effects on EPAS growth trajectories in English were -0.15 and 0.15, respectively. Thus, 0.15 could be considered a “good” score for the number of EPAS English trajectory points that could be attributed to a high school, over and above what could be expected of an “average” high school. Similarly, -0.15 could be considered a “poor” score for

the number of EPAS English trajectory points that could be attributed to a high school. Using these rules of thumb, 25% of the high school cohorts would have good scores, 25% would have poor scores, and 50% would have average scores.

Table 25 summarizes the distributions of the EPAS-CAVAM measures generated by Equation 12. The distributions of the EPAS-CAVAM measures are very similar to those of the EPAS-VAM measures. As expected, there is slightly less variation in the context-adjusted measures than the corresponding unadjusted effects.

TABLE 25

Distributions of Context-Adjusted Estimated School Effects on EPAS Growth Trajectories

Subject	Estimate of context-adjusted school effect on EPAS growth trajectories					
	Min	P₂₅	Med	P₇₅	Max	SD
English	-0.67	-0.15	-0.01	0.14	0.69	0.22
Mathematics	-0.58	-0.13	0.00	0.12	0.58	0.18
Reading	-0.45	-0.09	0.00	0.09	0.43	0.14
Science	-0.37	-0.07	0.00	0.07	0.34	0.11

Note: n = 1,019 high school cohorts

Table 26 contains the intercorrelations of the estimated school effects on EPAS growth trajectories. The context-adjusted effects are highly correlated with the unadjusted effects, with correlations ranging from 0.87 to 0.90. Consistent with the ACT-VAM and ACT-CAVAM models, this suggests that value-added measures are less influenced by contextual factors. However, the correlations are slightly smaller than those observed for the ACT-VAM and ACT-CAVAM models. Still, schools that are considered above average using the context-adjusted model will most likely be considered above average using the non-context-adjusted model. Across subject areas the EPAS-VAM and EPAS-CAVAM value-added measures are also highly

correlated, suggesting that high school cohorts that score well in one area will likely score well in other areas.

TABLE 26

Intercorrelations of School Effects on EPAS Growth Trajectories

Estimated school effect on growth trajectory in ...	1.	2.	3.	4.	5.	6.	7.	8.
1. English	1.00							
2. Mathematics	0.57	1.00						
3. Reading	0.68	0.59	1.00					
4. Science	0.61	0.70	0.76	1.00				
5. English (context-adjusted)	0.90	0.40	0.55	0.49	1.00			
6. Mathematics (context-adjusted)	0.50	0.88	0.46	0.57	0.51	1.00		
7. Reading (context-adjusted)	0.60	0.39	0.88	0.62	0.67	0.47	1.00	
8. Science (context-adjusted)	0.53	0.49	0.62	0.87	0.60	0.60	0.71	1.00

Note: $n=1,019$ high school cohorts

Uncertainty of Estimated School Effects

Earlier, we discussed how the sampling error of status measures (e.g., proportion proficient) is inversely related to sample size. The same is true for value-added measures. As we have demonstrated, value-added models can be used to produce estimates of school effects. Value-added models can also be used to produce measures of the *uncertainty* of the estimated school effects. Typically, standard errors and/or confidence intervals are used to quantify the uncertainty of estimates. If the standard error is larger, the confidence interval is wider, and there is greater uncertainty about the estimate. Reporting the uncertainty about estimates of school effects is crucial for an accountability system because the estimate may only be appropriately interpreted if there is adequate certainty about the estimate. How much certainty is adequate? Statisticians often use *p-values* as a measure of uncertainty. In our case, the p-value represents the probability that the estimate would have resulted if the “true” value was actually 0. Therefore, smaller p-values imply greater certainty that an estimated school effect is different

than that for the average school. P-values of .01 and .05 are commonly used thresholds for certainty.

For each of our estimated school effects, a p-value reflects the degree of certainty that the estimated school effect is greater or less than that for the average school. In Table 27, we present classifications for the 1,019 high school cohorts in the sample for the value-added measures generated from the ACT-VAM, ACT-CAVAM, EPAS-VAM, and EPAS-CAVAM models. Each high school cohort is classified as below average (estimated effect < 0 , p-value $< .05$), above average (estimated effect > 0 , p-value $< .05$), or uncertain (p-value $\geq .05$). The estimated school effects from the ACT-VAM and ACT-CAVAM models usually have p-values suggesting that school effects cannot be classified as “below average” or “above average” with certainty. For example, for value-added measures generated from the ACT-CAVAM model, 67% (for English) to 86% (for science) are classified as uncertain. For the ACT-VAM model, 66% (for English) to 83% (for science) of the school effects are classified as uncertain. Similar results are obtained for the EPAS-VAM and EPAS-CAVAM models. This finding has important implications for how value-added measures can be used: Because most school effects cannot usually be distinguished from “average” with certainty, the most common scenario for a high-stakes decision based on value-added measures is that no action (rewarding or sanctioning) should be taken. This problem is not unique to EPAS-based value-added measures, but reflects the reality that most school effects are not significantly different from the “average” school effect.

TABLE 27

Classifications of School Effects

Estimated school effect on	Pct. uncertain	Pct. below average	Pct. above average
<i>ACT scores</i>			
English	66	16	18.5
Mathematics	70	14	16
Reading	82	8	10
Science	83	7	10
English (context-adjusted)	67	17	16
Mathematics (context-adjusted)	72	14	14
Reading (context-adjusted)	85	8	7
Science (context-adjusted)	86	7	8
<i>EPAS growth trajectory</i>			
English	68	14	18
Mathematics	70	12	18
Reading	83	7	10
Science	82	7	11
English (context-adjusted)	69	15	16
Mathematics (context-adjusted)	73	14	13
Reading (context-adjusted)	86	7	7
Science (context-adjusted)	86	7	7

Note: $n=1,019$ high school cohorts, *Below Average* implies school effect < 0 with a p-value $< .05$, *Above Average* implies school effect > 0 with a p-value $< .05$.

Estimated school effects for large cohorts tend to have less sampling error (i.e., smaller standard errors); larger schools are therefore more likely to be classified with certainty. Accordingly, larger schools are more likely to be subject to high-stakes decisions. A possible remedy to this problem would be to combine data across multiple years for smaller schools. To determine how many years would be needed for a specific school, it is necessary first to consider how many students are needed in order to estimate value-added measures with greater certainty. Earlier (Table 21), we observed that the 25th and 75th percentiles of the ACT-VAM measures were -0.61 and 0.61, respectively. In order for estimates as extreme as 0.61 to be classified as “very certain” (p-value less than 0.01), the standard error of the estimate must be equal to or less

than 0.23. Thus, by requiring a standard error of 0.23 or smaller, we can be assured that 25% of the schools could be classified as “good”, 25% could be classified as “poor”, and 50% could be classified as “average” – and all “good” and “poor” classifications would be made with certainty. Generally, standard errors are proportional to \sqrt{n}^{-1} , where n is the number of students. For the ACT-VAM measures, we used a regression model (with no intercept) to find that the standard errors were approximately equal to $2.71/\sqrt{n}$ for English. Thus, for English, 139 students are needed in order to estimate values as extreme as 0.61 with certainty. We applied this same approximation procedure to the other subject areas and also to the ACT-CAVAM measures, finding that fewer students are needed for value-added measures with greater variation (English) relative to those with smaller variation (Science). In fact, about 388 students are needed to estimate values as extreme as the 25th and 75th percentiles with certainty for the ACT-CAVAM science measures. Thus, assuming that all students in a school are EPAS-tested, a school with average grade level enrollments of 100 students would need to combine four years of EPAS data to be able to estimate school effects on ACT scores with certainty across all four subject areas.

Even with our limited sample sizes, some of the estimated school effects are classified as above or below average with certainty (Table 27). The percentage of high school cohorts that are classified with certainty is directly related to the variability of estimated school effects. For example, from Table 21 we see that the ACT-VAM English measures have the greatest variation across high school cohorts; likewise, from Table 27 we see that the ACT-VAM English measures are most likely to have a classification made with certainty. Similar results are obtained for the EPAS-VAM and EPAS-CAVAM models.

Reliability of Value-Added Measures

Table 28 summarizes the autocorrelations of the value-added measures for adjacent cohorts (one year apart), as well as for cohorts that are 2 and 3 years apart. The correlations are weighted according to the average sample size (across cohorts) for each high school.

TABLE 28**Autocorrelations of Value-Added Measures**

Estimated school effect on	Years between cohorts		
	1	2	3
<i>ACT scores</i>			
English	0.78	0.73	0.70
Mathematics	0.77	0.74	0.71
Reading	0.60	0.56	0.61
Science	0.60	0.62	0.63
English (context-adjusted)	0.75	0.70	0.67
Mathematics (context-adjusted)	0.70	0.68	0.66
Reading (context-adjusted)	0.58	0.54	0.59
Science (context-adjusted)	0.57	0.60	0.58
<i>EPAS growth trajectory</i>			
English	0.76	0.70	0.66
Mathematics	0.79	0.75	0.72
Reading	0.62	0.55	0.58
Science	0.62	0.63	0.63
English (context-adjusted)	0.72	0.65	0.62
Mathematics (context-adjusted)	0.70	0.67	0.65
Reading (context-adjusted)	0.55	0.49	0.51
Science (context-adjusted)	0.55	0.58	0.55

Note: $n=422$ high schools for 1 year between cohorts, 279 for 2 years, 161 for 3 years.

Generally, the correlations are larger for adjacent cohorts and decrease as time between cohort increases. It appears that the ACT-VAM and ACT-CAVAM models generate value-added measures with comparable reliabilities. Comparing the autocorrelations of the value-added measures to those of the status measures (Table 10), one can see that the status measures have

slightly greater consistency over time. Still, the correlations in Table 28 suggest that schools that have high value-added scores for one cohort are likely to have high value-added scores for future (and past) cohorts.

Value-Added Measures: Relationships with Prior Mean Academic Achievement and School Contextual Factors

We now assess the associations of value-added measures with prior mean academic achievement and high school characteristics. Table 29 contains beta weights obtained by regressing each value-added measure on these characteristics using a multiple linear regression model.

We begin with the value-added measures generated by the ACT-VAM and ACT-CAVAM models. Surprisingly, prior mean academic achievement level (mean EXPLORE Benchmarks met) is negatively related to the value-added measures. Thus, cohorts with *higher* entering student achievement levels had significantly *lower* value-added scores. This relationship is in direct contrast to the relationships observed for status, improvement, and growth measures. Further research is needed to understand why higher entering student achievement levels are associated with lower estimated school effects in this sample. Aside from prior mean academic achievement, grade 11 enrollment was positively related to the value-added measures and poverty level was inversely related to the mathematics and science value-added measure. This suggests that larger schools in the sample tend to have greater effects on ACT scores. Because the ACT-CAVAM model adjusts for school characteristics (Equation 6), the beta weights relating the school characteristics to the value-added measures are forced to be 0. Therefore, by definition, the measures generated by the ACT-CAVAM and EPAS-CAVAM models are unrelated to grade 11 enrollment, proportion of students tested, poverty level, and proportion of

racial/ethnic minority students. However, as with the unadjusted value-added measures, prior mean academic achievement is negatively related to the context-adjusted value-added measures.

TABLE 29**Beta Weights for Predicting Value-Added Measures**

Estimated school effect on	Grade 11 enrollment	Proportion tested	Poverty level	Proportion minority	Mean number of EXPLORE Benchmarks met
<i>ACT scores</i>					
English	0.23	-0.13	-0.10	-0.08	-0.37
Mathematics	0.23	-0.01	-0.18	-0.10	-0.22
Reading	0.24	-0.08	-0.09	-0.15	-0.28
Science	0.20	-0.08	-0.23	-0.12	-0.23
English (context-adjusted)	0.00	0.00	0.00	0.00	-0.39
Mathematics (context-adjusted)	0.00	0.00	0.00	0.00	-0.24
Reading (context-adjusted)	0.00	0.00	0.00	0.00	-0.29
Science (context-adjusted)	0.00	0.00	0.00	0.00	-0.24
<i>EPAS growth trajectory</i>					
English	0.28	-0.08	-0.14	-0.09	-0.40
Mathematics	0.28	0.04	-0.21	-0.11	-0.24
Reading	0.31	-0.01	-0.12	-0.16	-0.30
Science	0.25	-0.02	-0.25	-0.13	-0.24
English (context-adjusted)	0.00	0.00	0.00	0.00	-0.42
Mathematics (context-adjusted)	0.00	0.00	0.00	0.00	-0.26
Reading (context-adjusted)	0.00	0.00	0.00	0.00	-0.31
Science (context-adjusted)	0.00	0.00	0.00	0.00	-0.25

Note: $n = 1,019$ high school cohorts. The ACT score value-added measure is the number of score points attributed to a school. The EPAS growth trajectory value-added measure is the score gain (from grade eight to grade 10 to grade 12) attributed to a school.

For the value-added measures generated by the EPAS-VAM and EPAS-CAVAM models, the relationships with prior mean academic achievement and school characteristics are

very similar to those observed for the ACT-VAM and ACT-CAVAM models. Generally, poverty level and proportion of racial/ethnic minority students are not significantly related to the ACT-VAM and EPAS-VAM measures. The most extreme beta weight is for poverty level and the school effects on ACT Science score and EPAS Science trajectory ($b=-0.23$ and $b=-0.25$, respectively), suggesting that wealthier schools have slightly greater effects on science achievement. To a lesser degree, the measures generated by ACT-VAM and EPAS-VAM for mathematics are affected by poverty level ($b=-0.18$ and $b=-0.21$, respectively). These findings are in contrast to the findings for status measures (Table 11), which are clearly related to poverty level and proportion of racial/ethnic minority students in the school.

Summary: Value-Added Models

- Value-added models are not compatible with the current rules of NCLB.
- To implement value-added models, the minimum data requirement is an entry score (e.g., EXPLORE score) and an exit score (e.g., ACT score) obtained on individual students.
- Two types of value-added measures were examined: measures of school effects on ACT scores, which do not require vertically-scaled assessments; and measures of school effects on EPAS growth trajectories, which require vertical scaling. For each type, we also examined *context-adjusted* value-added measures, which adjust the school effects for certain student and school characteristics.
- In most cases, estimated school effects do not differ significantly from the “average” school effect. Thus, the most common scenario for a high-stakes decision based on value-added measures is that no action (rewarding or sanctioning) should be taken.

- EPAS-based value-added measures have relatively high autocorrelations, suggesting that the measures are reliable and that schools that are “above average” for one cohort are also likely to be “above average” for future and prior cohorts.
- Compared to status measures, value-added measures have smaller associations with school poverty level and proportion of racial/ethnic minority students. In fact, context-adjusted value-added measures have no association with school characteristics that are included in the adjustment. Thus, value-added measures are much more likely to be accepted as fair measures of school effects.
- Cohorts with higher entering student achievement levels had significantly lower value-added scores. Further research is needed to understand why higher entering student achievement levels are associated with lower estimated school effects.

Case Examples: EPAS-Based Accountability Measures for Two High School Cohorts

In this section, examples of accountability measures are provided for two high school cohorts in the sample: A high-poverty, high-minority high school and a low-poverty, low-minority high school. This is an example of how the accountability measures can be used by schools to inform decision making.

A High Poverty, High Minority High School

The first school under consideration has a high poverty rate (50% of the students are eligible for free or reduced lunch) and a large concentration of racial/ethnic minority students (50%). This school had 93 students in their 2006 graduating class who had taken EXPLORE, PLAN, and the ACT. In Table 30, we present selected accountability measures for this high

TABLE 30

Accountability Measures for a High Poverty, High Minority High School

	Accountability measure	Score	SE	PR
<i>Status measures</i>	Proportion meeting Benchmark on...			
	PLAN English	0.84	0.04	54
	PLAN Mathematics	0.31	0.05	32
	PLAN Reading	0.65	0.05	67
	PLAN Science	0.17	0.04	32
<i>Growth measures</i>	Proportion vertically-projected to meet Benchmark on...			
	ACT English	0.72	0.05	47
	ACT Mathematics	0.30	0.05	37
	ACT Reading	0.49	0.05	66
	ACT Science	0.24	0.04	62
	Proportion WSR-projected to meet Benchmark on...			
	ACT English	0.80	0.04	68
	ACT Mathematics	0.23	0.04	32
	ACT Reading	0.54	0.05	56
	ACT Science	0.14	0.04	40
<i>Value-added measures</i>	Estimated school effect on ACT scores			
	English	1.68	0.33	96
	Mathematics	-0.33	0.30	35
	Reading	1.11	0.35	97
	Science	0.46	0.28	83
	Context-adjusted estimated school effect on ACT scores			
	English	1.22	0.33	92
	Mathematics	-0.34	0.30	32
	Reading	0.90	0.35	95
	Science	0.40	0.27	83
	Estimated school effect on EPAS growth trajectories			
	English	0.39	0.09	96
	Mathematics	-0.12	0.08	28
	Reading	0.25	0.09	95
	Science	0.09	0.07	78
	Context-adjusted estimated school effect on EPAS growth trajectories			
	English	0.29	0.09	90
	Mathematics	-0.10	0.08	30
	Reading	0.22	0.09	94
	Science	0.09	0.07	80

Note: SE=standard error, PR=percentile rank among 1,019 high school cohorts

school cohort. Status (as of 10th grade), growth (as of 10th grade), and value-added (as of 11th/12th grade) measures are presented. In reality, an accountability report would probably include

accountability measures *as of* the same grade level (i.e., status and value-added measures as of 11/12th grade, or status and growth projections as of 10th grade). But, for demonstration purposes, the complete mix of accountability measures is presented.

From Table 30, one can see that the cohort's 10th grade status (based on PLAN score) is slightly above average in reading (percentile rank of 67 among the 1,019 high school cohorts in the sample), about average in English (percentile rank of 54), and below average in mathematics and science (percentile ranks of 32). Using the WSR projection method, growth measures are obtained by projecting ACT scores based on EXPLORE and PLAN scores. Therefore, as of 10th grade, 80% of the students are projected to meet or exceed the College Readiness Benchmark in English, 23% in Mathematics, 54% in Reading, and 14% in Science. The growth measures, particularly those obtained from the WSR-growth model, are closely aligned with the status measures. The growth measures based on the Vertical-growth model are slightly different; this is probably because they are measured with greater error.

Perhaps the most informative accountability measures in Table 30 are the value-added measures. These suggest that the school has well-above-average effects on English, reading, and science performance, but slightly below-average effects on mathematics performance. For the value-added measures representing school effects on ACT scores (generated from the ACT-VAM and ACT-CAVAM models), the school has especially large effects on English (1.68 ACT English score points more than average) and reading (1.11 ACT Reading score points more than average). For valued-added measures representing school effects on EPAS growth trajectories (generated from the EPAS-VAM and EPAS-CAVAM models), the school effects on English and reading are 0.39 and 0.25, respectively. Importantly, the two types of value-added measures are not directly comparable because they are on different scales: The first value-added measure

represents schools effect on ACT scores while the second represents school effects on growth trajectories (i.e., yearly *change* in test score). The percentile ranks assigned to the value-added measures indicate that the school's effects are well above average in all areas except mathematics, where the school effect is slightly below average. The percentile ranks for the context-adjusted effects mirror those of the unadjusted effects.

Importantly, the standard errors of each accountability measure are reported, allowing one to gauge the certainty of the accountability measure. For example, the context-adjusted estimated school effect on ACT English score is 1.22, with a standard error of 0.33. One can then form a 95% confidence interval as $1.22 \pm 2(0.33) = 1.22 \pm 0.66 = [0.56, 1.88]$. The interpretation of this estimate and confidence interval is: "We estimate that the context-adjusted school effect is 1.22 and we are 95% certain that the effect is between 0.56 and 1.88." Because the lower bound of this interval (0.56) is still larger than the "average" effect (0), we are quite certain that the school has an above-average effect on English performance. As we have stressed in this report, the uncertainty of accountability measures can be especially troublesome for small schools and for subgroups of students, due to larger standard errors. For this particular cohort, 93 students were tested and there was palpable uncertainty reflected in the standard errors.

The accountability measures for this particular high school cohort highlight some important features of status, growth, and value-added measures. If we had only considered the cohort's status or growth measures, we might have concluded that this school's performance was average, or perhaps slightly below average. By also considering the value that the school added to academic performance, we were able to see that this school actually has shown above average effects in all subject areas except mathematics. This information might lead the school to study their mathematics curriculum and perhaps devise a strategy for improvement. Further, the school

can take pride in their strong effects on English, reading, and science and perhaps identify the teachers and practices that have contributed to their success.

A Low Poverty, Low Minority High School

The next school we consider has a low poverty rate (4% of the students are eligible for free or reduced lunch) and a small concentration of racial/ethnic minority students (4%). This school had 175 students in their 2003 graduating class who had taken EXPLORE, PLAN, and the ACT. Included in Table 31 are selected accountability measures for this high school cohort.

The 10th grade status of this high school cohort is well above average in each subject area, with percentile ranks ranging from 84 (English) to 94 (in the other three areas). Related to this high status, the growth measures reveal that the projected proportions meeting the ACT Benchmarks are also well above average. The value-added measures, on the other hand, tell a different story. For example, the school's estimated effects on ACT scores are about average, with percentile ranks ranging from 34 (for English) to 57 (for Science). So, even though the cohort has large proportions of students meeting the College Readiness Benchmarks, the school's effects on improving academic performance appear to be modest. Because the school has relatively few low-income and racial/ethnic minority students, it is not surprising that the context-adjusted value-added scores are lower than the unadjusted scores. The percentile ranks for the context-adjusted value-added measures range from 21 to 27, suggesting that the school is not performing as well as schools serving similar groups of students. Results based on the EPAS-VAM and EPAS-CAVAM models mirror those based on the ACT-VAM and ACT-CAVAM models.

Contrasting the two schools, we can see that value-added and status models can lead to different conclusions. Specifically, the low poverty school *looks like* a higher performing school

when comparing schools by status or growth measures, but the high-poverty school actually has greater effects according to the value-added measures.

TABLE 31

Accountability Measures for a Low Poverty, Low Minority High School

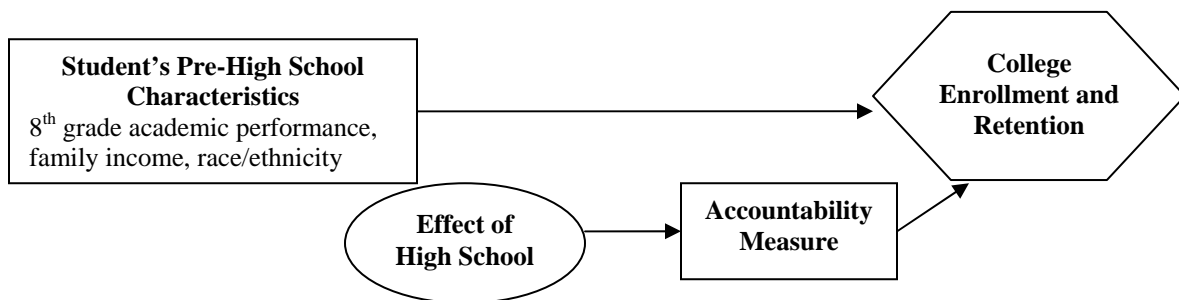
Accountability measure		Score	SE	PR
<i>Status measures</i>	Proportion meeting Benchmark on...			
	PLAN English	0.90	0.02	84
	PLAN Mathematics	0.63	0.04	94
	PLAN Reading	0.79	0.03	94
	PLAN Science	0.42	0.04	94
<i>Growth measures</i>	Proportion vertically-projected to meet Benchmark on...			
	ACT English	0.83	0.03	84
	ACT Mathematics	0.51	0.04	88
	ACT Reading	0.55	0.04	82
	ACT Science	0.29	0.03	80
	Proportion WSR-projected to meet Benchmark on...			
	ACT English	0.90	0.02	93
	ACT Mathematics	0.62	0.04	97
	ACT Reading	0.79	0.03	97
	ACT Science	0.44	0.04	97
<i>Value-added measures</i>	Estimated school effect on ACT scores			
	English	-0.38	0.25	34
	Mathematics	-0.02	0.23	49
	Reading	-0.11	0.27	42
	Science	0.09	0.21	57
	Context-adjusted estimated school effect on ACT scores			
	English	-0.72	0.25	21
	Mathematics	-0.46	0.23	27
	Reading	-0.38	0.27	24
	Science	-0.26	0.21	26
	Estimated school effect on EPAS growth trajectories			
	English	-0.09	0.07	35
	Mathematics	0.00	0.06	53
	Reading	-0.03	0.07	41
	Science	0.04	0.06	61
	Context-adjusted estimated school effect on EPAS growth trajectories			
	English	-0.19	0.07	19
	Mathematics	-0.12	0.06	26
	Reading	-0.11	0.07	21
	Science	-0.06	0.06	27

Note: SE=standard error, PR=percentile rank among 1,019 high school cohorts

Relation of EPAS-based Accountability Measures and College Enrollment and Retention Rates

In order for an accountability system to work, the performance indicators (i.e., accountability measures) must be meaningful and relevant (Fast & Hebbler, 2004). In order for accountability measures to be valid, they must be tied in a meaningful way to the overarching goals of the accountability system. For example, if one of the goals of a state's accountability system is to prepare students better for college, it would be desirable for the accountability measures to reflect schools' effects on college readiness. One way to assess the validity of EPAS-based accountability measures as markers of schools' effects on college readiness is to study their relationships to college enrollment and retention rates. In Figure 2, we present a conceptual model for validating EPAS-based accountability measures. If the accountability measure is truly measuring the high school's effect on students' college readiness, then the accountability measure should be predictive of college enrollment and retention, even after adjusting for students' pre-high school characteristics.

FIGURE 2. Conceptual Model for Validating Accountability Measures



College Enrollment and Retention Data

Data from the National Student Clearinghouse (NSC) was used to identify the students who enrolled in college the fall after high school graduation (first year enrollment) and who re-

enrolled at the same or a different postsecondary institution the second fall after high school graduation (retention). NSC enrollment data are available for at least 85% of enrolled freshmen who took the ACT. Thus, the data capture the overwhelming majority of freshman enrollees. Of the 1,019 high school cohorts in the sample, retention data are only available for 835 cohorts; retention data were not available for the 2002 and 2007 cohorts.

When students register for the ACT, they specify their first-choice college. Students whose first-choice college was not among those included in the NSC data were identified and an indicator variable was created to represent whether or not a student's first choice college was excluded. By doing so, the analysis was adjusted to accommodate for the fact that not all enrollments were included in the NSC data set. Overall, 68% of the students in the sample enrolled at an NSC institution their first year after graduation. Of those whose first choice college was not included in the NSC data, only 42% enrolled at an NSC institution. Likewise, 83% of the students returned to an NSC institution their second year after graduation; for those whose first choice college was not included in the NSC data, the retention rate was 77%.

Up to this point in this report, we have considered subject-specific accountability measures (e.g., proportion meeting the Benchmarks in English, Mathematics, Reading, and Science). Because we are interested in relating the accountability measures to college enrollment and retention rates, we now consider accountability measures that encompass all subject areas. For status measures, we consider the mean number of PLAN and ACT Benchmarks met. For improvement measures, we consider the year 2014 projected mean number of PLAN and ACT Benchmarks met. For growth measures, we consider the mean number of projected ACT Benchmarks met, where the projections are based on the WSR-growth model (Equation 1) and the VP-growth model (Equation 2). These *composite* accountability measures are equal to the

sum of the four subject-specific accountability measures. For value-added measures, we consider the mean of the four subject-specific value-added measures. Recall that we considered two general types of value-added measures: school effects on ACT scores (ACT-VAM) and school effects on EPAS growth trajectories (EPAS-VAM). For each type, we also considered context-adjusted effects (ACT-CAVAM and EPAS-CAVAM). So, there are four variants of composite value-added measures.

Analysis of Aggregated College Enrollment and Retention Rates

For each high school cohort in the sample, the proportions of students who enrolled the fall after high school graduation and re-enrolled at any postsecondary institution the next fall were tabulated. In Table 32, the distributions of the composite accountability measures and aggregated college enrollment and retention rates are summarized. There is considerable variability across the high school cohorts for all measures. For status measures, the mean of the mean number of PLAN Benchmarks met is 2.01, with standard deviation 0.46. The distributions of the mean number of ACT Benchmarks met are similar to that of the number of PLAN Benchmarks met. The means of the improvement measures mirror those of the status measures; however, the standard deviations of the improvement measures are substantially larger due to the “fanning out” caused by projecting status several years into the future. The means of the growth measures are similar to the mean of the ACT status measure; this is to be expected because the growth measures are the mean number of projected ACT Benchmarks met. Relative to the growth measure based on vertical projections, there is greater variation in the growth measure based on the WSR projection method. The value-added measures representing school effects on ACT scores have means of 0 by design (recall that the “average” school effect is always 0, according to Equation 5 and Equation 11). The proportion enrolled in college also varies

substantially across high school cohorts. The median first-year college enrollment rate is 0.68, but ranges from 0.13 to 1.00. The median retention rate is 0.82 and ranges from 0.29 to 1.00.

TABLE 32

Descriptive Statistics for Measures Related to College Enrollment Rates

Measure	N	Mean	SD	Min	Med	Max
<i>Status measures</i>						
Mean number of PLAN Benchmarks met	1,019	2.01	0.46	0.20	2.05	3.26
Mean number of ACT Benchmarks met	1,019	1.78	0.50	0.08	1.81	3.15
<i>Improvement measures</i>						
Year 2014 projected mean number of PLAN Benchmarks met	272	2.06	1.08	0.00	2.04	4.00
Year 2014 projected mean number of ACT Benchmarks met	272	1.78	1.08	0.00	1.85	4.00
<i>Growth measures</i>						
Mean number of WSR-projected ACT Benchmarks met	1,019	1.73	0.54	0.00	1.75	3.33
Mean number of vertically-projected ACT Benchmarks met	1,019	1.71	0.39	0.35	1.74	2.74
<i>Value-added measures</i>						
Estimated school effect on ACT scores	1,019	0.00	0.58	-1.93	0.00	1.79
Estimated school effect on growth trajectory	1,019	0.00	0.17	-0.55	0.00	0.49
Context-adjusted estimated school effect on ACT scores	1,019	0.00	0.53	-1.82	0.01	1.59
Context-adjusted estimated school effect on growth trajectory	1,019	0.00	0.14	-0.48	0.00	0.40
<i>College enrollment and retention rates</i>						
First-year enrollment rate	1,019	0.66	0.16	0.13	0.68	1.00
Retention rate	835	0.80	0.11	0.29	0.82	1.00

If the accountability measures are valid as markers of a school's effect on college readiness, they should have statistical relationships with enrollment and retention rates. Further, if the accountability measures are truly measuring the high school's contribution to college readiness, the statistical relationships should persist after adjusting for the high school cohort's prior mean academic achievement (mean number of EXPLORE Benchmarks met), as well as contextual factors (high school poverty level and proportion of racial/ethnic minority students in the school). As described earlier, it is also prudent to adjust college enrollment rates for students

whose first choice college is excluded from the college enrollment data. The mean proportion with first college choice excluded ranges from 0.00 to 0.67, with a median of 0.09.

Table 33 shows statistical relationships between the composite accountability measures and aggregated college enrollment and retention rates. The table includes simple correlations and correlations (beta weights) adjusted for prior mean academic achievement, school poverty level and proportion of racial/ethnic minority students, and proportion of students whose first choice college is excluded from the data. The correlations are weighted according to the average sample size (across cohorts) for each high school. Each of the accountability measures is correlated with first-year college enrollment and retention rates. The correlations for the status measures range from 0.39 to 0.42 with first-year enrollment rates, and from 0.54 to 0.55 with retention rates. The status measure based on the ACT (grades 11 and 12) has the highest correlations; this is to be expected because it is a proximal measure of aggregated college readiness. The improvement measures have smaller correlations with enrollment and retention rates; this may be a product of introducing additional measurement error in improvement measures (relative to status measures) by projecting status several years into future. The growth measures (mean number of projected ACT Benchmarks met) have correlations with enrollment rates of the same magnitude as those based on the status measures. The growth measure based on the WSR projections has slightly larger correlations than those based on the vertical projection; the smaller correlations of the growth measure based on vertical projection may be a product of extra attenuation associated with the larger SEMs of the vertically-projected ACT scores. The ACT-VAM value-added measures are also correlated with college enrollment and retention rates. The context-adjusted measures have smaller correlations with college enrollment and retention rates than do the unadjusted measures (0.20 versus 0.26 for first-year enrollment and 0.09 versus 0.24 for

retention). Correlations for the EPAS-VAM and EPAS-CAVAM value-added measures are similar to those for the ACT-VAM and ACT-CAVAM measures.

TABLE 33**Statistical Relationships of Accountability Measures and College Enrollment Rates**

Accountability measure	Relationship with college enrollment rates			
	Correlations		Beta weights	
	Enrollment	Retention	Enrollment	Retention
<i>Status measures</i>				
Mean number of PLAN Benchmarks met	0.39	0.54	0.27	0.35
Mean number of ACT Benchmarks met	0.42	0.55	0.39	0.31
<i>Improvement measures</i>				
Year 2014 projected mean number of PLAN Benchmarks met	0.13	0.20	0.03	0.03
Year 2014 projected mean number of ACT Benchmarks met	0.21	0.28	0.08	0.01
<i>Growth measures</i>				
Mean number of WSR-projected ACT Benchmarks met	0.39	0.52	0.32	0.32
Mean number of vertically-projected ACT Benchmarks met	0.37	0.50	0.18	0.20
<i>Value-added measures</i>				
Estimated school effect on ACT scores	0.26	0.24	0.20	0.12
Estimated school effect on Growth trajectory	0.25	0.26	0.19	0.14
Context-adjusted estimated School effect on ACT scores	0.20	0.09	0.20	0.11
Context-adjusted estimated school effect on growth trajectory	0.19	0.09	0.20	0.12

Note: $n = 1,019$ high school cohorts for first year college enrollment, $n = 835$ high school cohorts for second year college enrollment; beta weights represent standardized regression coefficients for accountability measure, where enrollment rate is regressed on the accountability measure, mean number of EXPLORE Benchmarks met, proportion of students whose first choice college is not included in enrollment data, proportion eligible for free or reduced lunch, and proportion minority

Table 33 shows that the composite accountability measures are predictive of college enrollment rates beyond what is already predicted by a high school cohort's prior mean academic achievement, contextual factors, and proportion of students whose first choice college is

excluded from the enrollment data. Again, the status measure based on the ACT Benchmarks is the most predictive, with a beta weight of 0.39 for first-year enrollment rates and 0.31 for retention. The improvement measures offer little or no prediction of college enrollment and retention rates. The growth measure based the VP-growth model is incrementally predictive, with beta weights of 0.18 and 0.20 for enrollment and retention, respectively. The growth measure based on the WSR-growth model is more incrementally predictive, with beta weights of 0.32 for both enrollment and retention. The value-added measures representing school effects on ACT scores are also incrementally predictive of college enrollment and retention rates, with beta weights of 0.20 and 0.12, respectively. The context-adjusted effects appear to be as predictive as the unadjusted effects. The statistics for the EPAS-VAM and EPAS-CAVAM value-added measures mirror those based on ACT-VAM and ACT-CAVAM measures.

These results suggest that the accountability measures, with the exception of improvement measures, are correlated with, and incrementally predictive of, college enrollment and retention rates. Hence, the results support the proposition that the accountability measures are valid markers of schools' effects on college readiness.

Discussion

In this study, we demonstrated that EPAS data could be used as the basis for high school accountability models. While EPAS was not specifically designed to accommodate an accountability system, it has important features (e.g., pre, during, and near-end high school assessments; content standards most relevant to skills needed for college success) that make it a valuable source of information that can be used to implement accountability models measuring school effects on college readiness. This demonstration was based on 1,019 high school cohorts and over 70,000 students with test scores from three time points (8th, 10th, and 11th or 12th

grades). Our sample was not representative of all public high schools in the United States. In particular, most of the high school cohorts were located in Midwest and south-central states. Further, high-racial/ethnic minority and large-enrollment high schools were under-represented. It is unlikely that this under-representation affected the primary findings of this study. However, it is likely that the under-representation would affect the normative accountability scores that were assigned to high school cohorts in our sample. For example, in Table 30, we reported that a specific high school cohort in our sample had a normative score of 35 (percentile rank) for the school's effect on ACT Mathematics score. If our sample had been more nationally representative, the resulting normative score could have been different.

Our findings highlight how status, improvement, growth, and value-added models can lead to very different conclusions about school effectiveness. Clearly, status, improvement, and growth measures can be heavily influenced by factors outside of the school's control – specifically, the entering achievement level and socioeconomic status of the students served by the school. By using value-added models, the school's effect is better isolated and measured. Thus, we found that value-added measures have smaller associations with prior mean academic achievement and, by extension, school contextual factors such as poverty level (proportion receiving free or reduced lunch) and proportion of racial/ethnic minority students. In Table 34, we see that the composite status (mean number of ACT Benchmarks met) and growth (mean number of vertically-projected ACT Benchmarks met and mean number of WSR-projected ACT Benchmarks met) measures are highly correlated with one another and also highly correlated with school poverty level and proportion minority. The value-added measures from the ACT-VAM and EPAS-VAM models have much smaller correlations with poverty level and proportion of racial/ethnic minority students; the context-adjusted value-added measures have the smallest

correlations with poverty level and proportion of racial/ethnic minority students. Because value-added models better isolate the effects that schools have on student learning, they are less likely to be strongly related to school contextual factors and are more likely to be perceived as fair accountability measures.

TABLE 34

Intercorrelations of Composite Accountability Measures and School Contextual Factors

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Mean Number of ACT Benchmarks Met	1.00								
2. Mean Number of Vertically-Projected ACT Benchmarks Met	0.88	1.00							
3. Mean Number of WSR-Projected ACT Benchmarks Met	0.91	0.89	1.00						
4. Estimated School Effect on ACT Scores	0.50	0.45	0.25	1.00					
5. Estimated School Effect on Growth Trajectory	0.48	0.45	0.24	0.97	1.00				
6. Context-Adjusted Estimated School Effect on ACT Scores	0.32	0.30	0.08	0.90	0.84	1.00			
7. Context-Adjusted Estimated School Effect on Growth Trajectory	0.29	0.28	0.06	0.88	0.98	0.85	1.00		
8. Poverty level	-0.63	-0.57	-0.59	-0.27	-0.33	-0.01	-0.01	1.00	
9. Proportion minority	-0.46	-0.43	-0.46	-0.08	-0.07	-0.07	-0.07	0.49	1.00
<i>Note: n = 1,019 high school cohorts</i>									

This study also demonstrated some of the practical problems encountered when implementing accountability models. Perhaps the most obvious requirement of a value-added accountability model is longitudinal test score data for students. We required high school cohorts included in our analysis to have at least 50% of the cohort with EXPLORE, PLAN, and ACT test scores. Maximizing student representation is a crucial element of any accountability system. If data are not available for a significant portion of students in a school, there could be concern that the resulting accountability measures are not an accurate reflection of the school's effects.

Moreover, the standard errors of accountability measures will be larger when many students are missing from the analysis – the consequence of this is greater uncertainty about the school’s effects.

The standard errors of accountability measures can be quite large, even when all students are counted in the calculations. Naturally, this problem is more pervasive at smaller schools. Because of this problem, standard errors of accountability measures should be reported, especially when the measures are used for high-stakes decisions. By doing so, stakeholders can better understand that accountability measures are simply estimates, and that some estimates are rather imprecise. By reporting the uncertainty about accountability measures, stakeholders are more likely to use the data properly. For example, stakeholders would be less likely to harshly judge a school with a small effect on ACT Mathematics score (a value-added measure) if they understood that the estimate of the effect was imprecise (i.e., the estimate had a wide confidence interval). The problem of large standard errors of accountability measures becomes magnified when results are reported for student subgroups. For this reason, it is often difficult to draw strong conclusions about a particular school’s effects on certain subgroups.

In this study, we considered students who had tested in 8th, 10th, and 11th or 12th grade. In order to measure the effect that high schools have on student learning, an entry and an exit score are necessary. Because students typically take EXPLORE in 8th grade, EXPLORE scores are the natural choice for an entry score; likewise, ACT scores are the natural choice for exit scores. Ideally, EXPLORE would be taken at the end of 8th grade; otherwise, the measured high school effect would include the portion of learning that took place in grade 8 after EXPLORE was taken. Similarly, the ACT would ideally be taken at the end of 12th grade; otherwise, the measured high school effect would not include the portion of learning that took place in grade 11

or 12 after the ACT was taken. Because of the requirements of college applications, very few students choose to take the ACT at the end of grade 12. Therefore, it is likely that measures of high school effects would only include the portion of learning that took place through the time of ACT testing. In order for accountability measures to be truly comparable across schools, it is necessary for the assessments to be spaced in a similar fashion. For example, it might be misleading to compare academic growth from the beginning of grade 8 to the beginning of grade 12 at school “A” to academic growth from the end of grade 8 to the middle of grade 11 at school “B”: In this case, students at school “A” might be expected to show greater growth due to the larger time span. When implementing a value-added model, care should be taken to account for different time spacing of assessments across schools. This problem could be addressed by introducing a covariate in the models that accounts for varying time spans.

Because accountability measures are often used as the basis for rewarding or sanctioning teachers or schools, it is implied that accountability models are actually measuring the effects that teachers or schools have on student learning. Most educational researchers and policymakers agree that status models do not actually measure effects on students learning. However, there is considerable debate on whether value-added models actually measure teacher or school effects on student learning. Some authors (Raudenbush, 2004; Rubin, Stuart, & Zanutto, 2004) argue that value-added models do not adequately measure the complex ways in which teachers or schools affect student learning. Others (Martineau, 2006) suggest that value-added models that rely on vertically-scaled assessments can lead to distorted conclusions about the effectiveness of teachers or schools. Clearly, more work is needed to understand better the relevance of value-added models. As a starting point, one must define *what is desired* to be measured by value-added models. Then, one must tailor the specifics of the assessment system and the statistical

models to estimate what is desired. In our analysis of value-added models, we considered specific forms of hierarchical linear models, which are widely used in practice to implement value-added models.

In this report, we presented value-added measures that were adjusted for contextual factors (i.e. school poverty level, school's proportion racial/ethnic minority students) and those which were not. We found that the context-adjusted measures were highly correlated with unadjusted measures (Table 23 and Table 26). Ballou, Sanders, and Wright (2004) discuss how adjusting for contextual factors could distort the measurement of teacher or school effects. They write, "If better teachers are able to obtain jobs in schools serving an affluent population, or if more affluent parents seek the best schools and teachers for their children, demographic and SES variables become proxies for teacher and school quality. Because they are correlated with otherwise unmeasured variation in school and teacher quality, the coefficients on these variables will capture part of what researchers are trying to measure with residuals" (pp.38-39). In other words, value-added models cannot always distinguish the effect of contextual factors from the effects of schools. Another practical reason for not adjusting for contextual factors is that the adjustment requires additional student-level data, such as parent's income, parent's education level, and race/ethnicity. Such data may not be readily available or reliably measured.

Entities responsible for implementing and reporting accountability models should take great care to ensure that results are properly interpreted. Users must understand the limitations of specific accountability measures, and care must be taken not to interpret accountability measures outside of their intended purposes. Guiding policymakers to make the appropriate use of accountability measures could help alleviate concerns with the possible adverse impacts of accountability systems. For example, such guidance could prevent policymakers from making

high-stakes decisions on the basis of status or growth (projection) models, or on value-added measures with large standard errors.

References

- ACT. (1999). *PLAN Technical Manual*. Iowa City, IA: Author.
- ACT. (2006). *The ACT Technical Manual*. Iowa City, IA: Author.
- ACT. (2007a). *ACT high school profile report: Colorado*. Iowa City, IA: Author.
- ACT. (2007b). *ACT high school profile report: Illinois*. Iowa City, IA: Author.
- ACT. (2007c). *EXPLORE Technical Manual*. Iowa City, IA: Author.
- ACT. (2008). *ACT high school profile report: National*. Iowa City, IA: Author.
- Allen, J. & Scoring, J. (2005). Using ACT Assessment Scores to Set Benchmarks for College Readiness. (ACT Research Report 2005-3). Iowa City, IA: ACT.
- Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, 37, 65-75.
- Ballou, D. (2002). *Sizing Up Test Scores*. Retrieved September 13, 2007, from <http://www.hoover.org/publications/ednext/3398961.html>.
- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, 29, 37-65.
- Callender, J. (2004). Value-Added Assessment. *Journal of Educational and Behavioral Statistics*, 29, 5.
- Choi, K., Goldschmidt, P., & Yamashiro, K. (2005). Exploring models of school performance: From theory to practice. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (NSSE Yearbook, Vol. 104, Part 2, pp. 119-146). Chicago: National Society for the Study of Education. Distributed by Blackwell Publishing.
- Fast, E.F. & Hebbler, S. (2004). *A Framework For Examining Validity In State Accountability Systems*. A Paper in the Series: Implementing the State Accountability System Requirements Under the No Child Left Behind Act of 2001. Washington, DC: Council of Chief State School Officers.
- Goldschmidt, P. & Choi, K. (2007, Spring). *The practical benefits of growth models for accountability and limitations under NCLB* (CRESST Policy Brief 9). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

- Hershberg, T., Simon, V.A., & Lea-Kruger, B. (2004). Measuring What Matters: How value-added assessment can be used to drive learning gains. Retrieved September 13, 2007, from http://www.cgp.upenn.edu/ctr_pubs.html.
- Howley, C. B., Strange, M., & Bickel, R. (2000). *Research about School Size and School Performance in Impoverished Communities* (ERIC Digest). Charleston, WV: ERIC Clearinghouse on Rural Education and Small Schools. (ERIC Document Reproduction Service No. ED 448 968).
- Linn, R. L. (2001). *The Design and Evaluation of Educational Assessment and Accountability Systems*. CSE Technical Report (CSE_TR-539). Los Angeles, CA: Center for Research and Evaluation, Standards, and Student Testing.
- Linn, R.L. (2006). Educational Accountability Systems. (CSE Technical Report 687). Los Angeles: Center for the Study of Evaluation.
- Martineau, J.A. (2006). Distorting value added: The use of longitudinal, vertically-scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, 31, 35-62.
- National Center for Education Statistics. (2007). *Mapping 2005 State Proficiency Standards Onto the NAEP Scales* (NCES 2007-482). U.S. Department of Education. Washington, DC: Author.
- Raudenbush, S.W. (2004). What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice? *Journal of Educational and Behavioral Statistics*, 29, 121-129.
- Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.
- Rubin, D.B., Stuart, E.A., & Zanutto, E.I. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics*, 29, 103-116.
- Sable, J., Thomas, J.M., & Sietsema, J. (2006). *Documentation to the NCES Common Core of Data Public Elementary/ Secondary School Universe Survey: School Year 2004-05*, (NCES 2006-339). U.S. Department of Education. Washington, DC: National Center for Education Statistics.
- U.S. Department of Education, Office of the Under Secretary. (2002, September). *No Child Left Behind: A desktop reference*. Washington, DC: Author.
- U.S. Department of Education. (2007). *No Child Left Behind. Growth models: Ensuring grade-level proficiency for all students by 2014*. Retrieved July 24, 2007, from www.ed.gov/admins/lead/account/growthmodel/proficiency.pdf.

U.S. Department of Education. (2008). *U. S. Secretary of Education Margaret Spellings Approves Additional Growth Model Pilots for 2007-2008 School Year*. Retrieved November 12, 2008, from www.ed.gov/news/pressreleases/2008/06/06102008.html.

Wright, S. P., Sanders, W. L., & Rivers J.C. (2005). Measurement of Academic Growth of Individual Students toward Variable and Meaningful Academic Standards, in R.W. Lissitz (ed.) *Longitudinal and Value Added Models of Student Performance*, Maple Grove, MN. JAM Press.

Appendix A
States and Locales of High School Cohorts Studied

State	Location of school						Total
	Large city	Mid-size city	Urban fringe of city	Large town	Small town	Rural	
Alabama	0	0	1	0	0	0	1
Arkansas	0	15	13	0	55	138	221
Colorado	0	10	10	0	10	28	58
Florida	0	0	0	0	0	3	3
Georgia	0	0	1	0	0	0	1
Iowa	0	0	4	0	8	10	22
Illinois	0	18	50	1	25	77	171
Kansas	0	1	8	0	23	47	79
Louisiana	6	7	7	0	6	53	79
Michigan	0	0	8	0	3	31	42
Minnesota	0	0	0	0	2	1	3
Missouri	4	3	5	0	7	20	39
Mississippi	0	0	0	0	0	1	1
Montana	0	0	0	0	0	4	4
North Dakota	0	0	0	0	0	4	4
Nebraska	0	0	3	0	1	43	47
New Mexico	0	0	0	0	0	2	2
Ohio	0	0	4	0	0	5	9
Oklahoma	10	1	13	0	46	115	185
South Dakota	0	0	1	0	0	0	1
Texas	0	0	0	0	0	5	5
West Virginia	0	0	5	0	14	23	42
Total	20	55	133	1	200	610	1,019
Sample %	2	5	13	<1	20	60	100
Population %	10	10	28	1	11	40	100

Note: Population total derived from 2004 Common Core of Data (Sable et al., 2006)

Appendix B

Projection Parameters from WSR Method for Projecting ACT Scores

Parameter	Corresponding mean / predictor	ACT score			
		English	Mathematics	Reading	Science
M_Y	ACT score	20.429	20.143	20.932	20.632
M_1	EXPLORE English	15.849	15.849	15.849	15.849
M_2	EXPLORE Mathematics	16.079	16.079	16.079	16.079
M_3	EXPLORE Reading	15.783	15.783	15.783	15.783
M_4	EXPLORE Science	17.314	17.314	17.314	17.314
M_5	PLAN English	18.220	18.220	18.220	18.220
M_6	PLAN Mathematics	18.216	18.216	18.216	18.216
M_7	PLAN Reading	17.629	17.629	17.629	17.629
M_8	PLAN Science	18.800	18.800	18.800	18.800
b_1	EXPLORE English	0.325	0.024	0.150	0.044
b_2	EXPLORE Mathematics	0.082	0.361	0.006	0.163
b_3	EXPLORE Reading	0.135	-0.003	0.287	0.068
b_4	EXPLORE Science	0.062	0.104	0.144	0.178
b_5	PLAN English	0.478	0.070	0.303	0.105
b_6	PLAN Mathematics	0.147	0.557	0.046	0.255
b_7	PLAN Reading	0.124	0.000	0.317	0.082
b_8	PLAN Science	0.080	0.184	0.199	0.291

Note: Estimated based on sample of 17,740 students (approximately 25% of sample)