National Center and State Collaborative

Guidance for Estimating and Evaluating Academic Growth

Chris Domaleski and Erika Hall

Center for Assessment

Development of this report was supported by a grant from the U.S. Department of Education, Office of Special Education Programs (H373X100002, Project Officer: Susan.Weigert@ed.gov). The contents do not necessarily represent the policy of the U.S. Department of Education, and no assumption of endorsement by the Federal government should be made.

Guidance for Estimating and Evaluating Academic Growth

Chris Domaleski and Erika Hall

Center for Assessment

December, 2016

Executive Summary

The purpose of this paper is to explore promising practices for including NCSC scores in estimates of academic growth. We begin with a brief overview of common approaches for computing individual measures of academic growth and summarize some of the key technical and administrative criteria necessary to implement them. Next we discuss factors that have traditionally limited the calculation, reporting and use of growth measures for students participating in alternate assessments and provide examples of the ways in which states are currently using alternate assessment data to calculate growth for student reporting and/or school accountability.

To help evaluate the types of approaches that might be feasible, we conduct a series of analyses with spring 2015 data in grades four and seven for ELA and mathematics. First, we produce histograms of scaled scores for all states and individually for the three largest states. Next, we produce tables of relative frequency by scale score and associated conditional standard of error (CSEM).

We found the spring 2015 NCSC assessment does not appear to exhibit ceiling effects insofar as there was not clustering near the highest and lowest scale scores. However, the error near the tails of the scale was very high for the four tests examined. This suggests the test is not very precise for disentangling degrees of performance for very low or very high achieving students. Moreover, sample size was very low for scores across the scale range except for the most dense region. This was true with all participants included and especially true for analyses by state.

Taken together, these findings suggest that methods for estimating academic growth that rely on conditioning across the full range of the scale may produce unreliable or uncertain results, particularly for any one state. However, such methods may be more promising if sample size is increased by using multiple state data and measurement error is reduced by adding discriminating items to the form in the most impoverished regions of the scale.

This does not preclude exploring approaches to growth in the near term that are less dependent on sample size and are not based on information along the full scale. For example, a value table approaches that credits progress across performance levels or even categories within

performance levels may offer a promising alternative. Another alternative may be to explore content based interpretations of academic progress.

Introduction

Many states may be interested in producing measures of academic growth for students participating in the National Center and State Collaborative (NCSC) assessment. In fact, interest is sure to be magnified given the emphasis placed on growth in the Every Student Succeeds Act (ESSA) and requirements specified in the *Final Regulations for Accountability, State Plans, and Data Reporting* released November 28, 2016.[1] The statute and supporting regulations require states to select an academic progress measure that is "valid, reliable, and comparable" and specifies that the indicator must include students who take an alternate assessment based on alternate achievement standards (AA-AAS). It is unclear if the exact same indicator must be used for the AA-AAS as with the general assessment, but it is clear that the indicator must be based on academic progress on the applicable assessment.

Accordingly, the purpose of this paper is to support states in exploring promising practices for including NCSC scores in estimates of academic growth. We begin with a brief overview of common approaches for computing individual measures of academic growth and summarize some of the key technical and administrative criteria necessary to implement them. Next we discuss factors that have traditionally limited the calculation, reporting and use of growth measures for students participating in alternate assessments and provide examples of the ways in which states are currently using alternate assessment data to calculate growth for student reporting and/or school accountability. We present NCSC data and analyses that can be used to support decisions on the types of approaches that might be feasible once multiple years of data are available. Finally, we conclude with recommendations and implications for practice.

Summary of Growth Models

We begin with a classification scheme to broadly describe the range of approaches widely used in student-level estimates of academic growth. These approaches and some of the major advantages/ limitations of each are summarized in Table 1. The approaches fall in four general groups: categorical gain, gain score, value-added, and normative. We also classify models as based on an observed score or predicted score, and discuss the application of a criterion referent to any of these approaches. We acknowledge there are multiple ways to group and categorize models (see e.g., Ho & Castellanos, 2013). However, we believe the proposed scheme effectively encompasses the variability observed across most assessment programs, including alternate assessments.

[1] The full text of the Every Student Succeeds Act (ESA) is available at: https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf and Final Regulations for Accountability State Plans and Reporting released November 2016 is available at: https://www.gpo.gov/fdsys/pkg/FR-2016-11-29/pdf/2016-27985.pdf

It should be noted that the categories are not meant to be mutually exclusive. For example, a normative growth model can be regarded as a value-added model, particularly if used to describe contributions to student learning. In addition, while most of these approaches reflect methods that generate individual measures of student growth some are not typically used to support individual growth inferences. Specifically, value-added models (VAM) may calculate a growth score for each student in a school/classroom, but utilize variables related to the target of attribution (e.g. schools, teachers, programs) to support an inference at a summary level.

Table 1. Growth Model Classifications

Method	Description	Answers what question?	Advantages	Limitations
Categorical Gain (a.k.a. Value Tables or Transition Models)	A measure of the change in performance level category from time 1 to time 2	Did the student advance or decline across performance levels?	-Straightforward to understand and implement - Clear relationship to status - Growth determinations are not constrained by sample size	-Insensitive to large magnitude growth within a performance level -Overly sensitive to small magnitude growth that crosses performance levels -Not well suited for students at the extremes. -Assumes performance levels are aligned and articulated across grades within a content area. - If performance levels change, comparability is not supported
Gain Score	Score difference between time 1 and time 2	What is the magnitude of student growth?	-Straightforward to understand and implement. - Provides results on a familiar scale with a known relationship to status	-Requires a vertical scale - There are technical concerns with vertical scales - Magnitude of growth cannot be interpreted the same for all students -Difficult to interpret in the absence of a normative reference or growth target (i.e., How much gain is expected or reasonable?)
Value-Added Model (VAM)	Regression based approach that often controls for multiple variables to determine the change in student performance, often for the purpose of attribution to an external source (e.g. teacher, school)	To what degree was the performance higher or lower than expected?	- Often accounts for multiple factors that influence growth -Typically, expectations are adjusted based on abilities and additional characteristics related to the student and/or the school.	-More complex to implement -Including background variables can be controversial because it is associated with different growth expectations -No 'built-in' relationship to status, but growth targets can account for this. -Often not designed to support individual growth inferences. -Requires data systems that allow for longitudinal tracking of all relevant variables utilized in the model. -Requires relatively large n-counts.
Normative	Regression based	To what degree is	-Provides a familiar	-More complex to implement

(e.g., Student Growth Percentiles)	approach that describes a student's current performance given that of students with similar profiles or prior test performance.	a student's performance higher or lower than expected, given the performance of students with similar academic history?	basis to interpret performance: the percentile. -Provides a definition of 'typical growth' -Expectations are adjusted for students of various abilities.	-No 'built-in' relationship to status, but growth targets can account for this -Requires relatively large n-counts to support stable estimates -Estimation is problematic for students at the extremes if not well represented in the population

Another term that is often used when describing approaches to measuring growth is 'residual gain score.' This refers to any approach that is based on the difference between predicted and actual performance (i.e., the residual). We do not include a separate row for this method in the classification table above, because it is a feature incorporated by other models, especially those associated with VAM. In fact, residual gain score models answer the same question as VAM—"Was the performance higher or lower than expected?"—without the added goal of inform attribution.

The specification of a growth model goes beyond simply producing the estimate of growth. Growth models frequently include a standard to define 'good enough' growth. Some models, in fact, are chiefly described by the associated performance standard, such as 'growth to standard' or 'criterion referenced' models. These approaches produce a growth estimate designed to evaluate whether a student showed enough growth to reach a defined standard by the end of an established period (e.g., enough growth to be "Proficient" in three years.) To be clear, this type of growth standard can be applied to most any model, such as when a norm referenced model defines the growth percentile required for a student to reach a target level of performance. In other cases, growth standards can be defined normatively, such when expectations are associated with demonstrating a minimum of 'typical' performance (e.g. the student's growth was at the 50^{th} percentile or the gain score was at or above the mean.)

There is no single correct approach to growth or method that stands-out as the 'gold-standard.' The decision regarding which analytic approach should be adopted should be informed by:

- The ability to support the technical properties and assumptions underlying the growth model (e.g., sample size; linearity, alignment of performance standards, vertical scale)
- The context and purpose for measuring growth (e.g., describe/report a measure of individual student growth, contribute to school accountability, inform educator evaluation)
- The desired model characteristics (e.g., result in a measure on the reportable scale, transparent/easy to interpret, tolerant of missing data, resilient)

In the best case, the selected model should produce outcomes that are reliable and valid for the intended uses and produce results that are clear and easily understood by stakeholders. Additionally, the model should be practically feasible to implement and maintain.

Given the challenges highlighted in Table 1, it is not surprising that there is limited information available on the achievement and growth of students participating in alternate assessment programs. It is difficult to develop alternate assessments that are both appropriate for this population and that meet the technical standards required to support growth (e.g., alignment, construct representation across administrations due to length limitations, etc.) In addition, the design of alternate assessments has changed significantly over the last decade making the computation of test-based student growth measures essentially unfeasible until recently (Tindale, 2015).

In particular, the technical requirements necessary to support the use of vertical scales or regression based approaches are more difficult to achieve for alternate assessments. As we will discuss in the subsequent section, this is related to the need for adequate n-size and precision along the full scale. In fact, many state alternate assessments do not produce scale scores that describe a broad range of performance, but rather report only performance levels. Obviously, this limits the range of approaches available to calculate academic growth.

Criteria

In this section, we address some key psychometric criteria that must be in place to produce estimates of academic growth. While most of these criteria apply to all of the methods highlighted in Table 1, they vary in importance across the different procedures.

First, regardless of the approach applied, a student must have at least two scores – a prior score and a current score- with which to estimate growth. In addition, while requirements defining the relationship between prior and current scores may differ across methods, there must always be evidence supporting use of the selected measures for estimating growth in the target content domain. For example, if a gain score is to be calculated across two consecutive mathematics assessments, in addition to the need for a vertical scale, the two tests should be developed with a clear understanding of how the math construct is defined within and across grades (i.e., through a learning progression or articulated set of standards).

Second, the range of performance underlying the assessments used to estimate growth must be sufficient. That is, the assessments must have sufficiently ‘high ceilings’ and a ‘low floors’ to measure performance across a broad range of student abilities. If the range is not sufficiently broad, the assessment will not reliably detect gains between multiple assessments for students of high or low ability. Additionally, even if the range is broad, the scale must be sufficiently precise along the full range of the score scale. If measurement error is overly pronounced at any region of the scale, especially the extremes, growth estimates for students scoring in these regions of the scale will be uncertain and unreliable.

Because computation of growth requires one or more suitable prior scores that are well correlated, the sample in the prior and current year must be adequate. Specifically, the number of examinees (n-size) earning scores along the scale should be large-enough and spread-out such that 'gaps' or 'clusters' in the distribution are minimized.

One must also consider n-size not only with respect to what is needed to calculate particular growth estimates, but also calculation for aggregate levels to which the growth inferences will apply. With very small n-sizes, sampling error can cause substantial fluctuations in scores. This is particularly relevant given that students are rarely, if ever, randomly assigned. Sampling error is directly related to the number of observations—as the sample size increases, the variability reduces.

There are many other criteria to consider when evaluating if and how measures of academic growth should be produced for an assessment that go beyond the scope of this paper and are tied to specific uses of growth measures, including school accountability, educator evaluation, and/or the reporting of student-level results. Some of these include:

- whether to include covariates other than prior scores in prediction-based methods
- establishment of growth targets (i.e. how to define 'good enough' growth)
- ability to match data from year to year and, if necessary, link to a unit of inference (e.g. teacher or school of record)
- approaches for handling missing data
- capacity to operationalize the system
- ease of understanding results and supporting appropriate interpretation and use.

In the next section, we discuss why alternate assessments tend to have difficulty meeting these criteria.

Growth for Alternate Assessments

Highlights from the Literature

It is clear from the previous sections that the approaches to evaluating growth have common as well as unique challenges and limitations. These challenges are difficult to overcome for many assessments, but they are particularly difficult to surpass for alternate assessments. This is due to the size and variability of the test taking population, which influences test design, and additional factors that influence the consistency of student participation across years. Some of the key issues, as summarized in previous research (e.g., Buzick and Latisus, 2010; Tinsdale, 2015; Ahearn, 2009), are discussed below:

- *The diversity and unique needs of the test taking population necessitates test and administration designs that may not support interpretations of student growth.*

To ensure they are appropriate, fair and do not require too much instructional time, alternate assessments, including the NCSC, are typically shorter than regular assessments, limiting the breadth and depth of the content that can be assessed. With respect to growth, this raises concerns around construct consistency as the type/range of content assessed may differ greatly from one year to the next. While this construct consistency issue impacts several of the methods described in Table 1, it is particularly problematic for observed score models that rely on changes in performance level designations or movement along a vertical scale.

Similarly, to ensure students are given the opportunity to demonstrate what they know and can do, test items and tasks must provide for maximum accessibility and utilize scoring rubrics that account for the type and degree of support required. Subjectivity in the administration and scoring of these types of tasks can greatly influence the reliability of scores, making it difficult to defend calculation of growth over years regardless of the method in play.

- *The psychometric properties of the test may not support the calculation of a growth measur*e. This can occur if a test does not span an appropriate range of difficulty (i.e. high ceiling/low floor) or if limited information leads to greater measurement error, which suppresses the precision of growth estimates.

- *The size of the test taking population tends to prohibit the use of certain types of growth measures.* Many of the methods outlined in Table 1 require large n-counts to support the estimation of stable growth scores. Specifically, normative procedures that condition on the performance of similar peer groups across multiple test scores and value-added calculations that utilize a large number of variables to estimate the contribution of a school or teacher to a student's observed performance.

- *Changes to the test design or the accommodations afforded to a student across years threaten the validity of between year growth inferences*. Research has shown that the use of different accommodations may result in differential score changes on some assessments for students with disabilities (Pitoniak & Royer, 2001; Royer, 2001; Sireci, Scarpetti and Li, 2005). Therefore, modifying the accommodations provided to a student across years, due to changes in a student's IEP team or disability category, may confound the accuracy of growth estimates. Similarly, significant changes to the test design, blueprint or administration conditions offered from one year to the next can have a negative impact on growth measures, especially if those measures are based upon previously established models or equations.

- *Changes in participation policy may result in inconsistent participation in the alternate assessment*. Due to shifts in policy or modifications to a student's IEP, a student may take an alternate assessment one year and a general assessment the following year prohibiting the estimation of growth. This is less of an issue for students with significant cognitive disabilities, but may still be a threat if a State's participation policy is unclear or unevenly applied.

State Practices

Predominantly, states do not produce measures of academic growth using results from their alternate assessments. In a recent Council of Chief State School Officers (CCSSO) survey of state department accountability leaders, 14 of 19 respondents indicated that their state did not include estimates of growth for their alternate assessment. [2]

To explore further, we examined practices in several states. While we found limited public information available for states producing growth estimates using their alternate assessments, personal communication with state leaders provided some additional insight. From a review of a sample of states it appears that those that calculate growth or plan to do so in the future tend to use one of the following procedures: categorical gain, gain score, and normative.

Categorical Gain

States that use categorical gain approaches determine adequate growth in light of changes in observed test performance over two consecutive years. For example, Nebraska uses a decision matrix to determine whether a student's performance is assigned a "growth point" for purposes of accountability, as shown below.

<table>
<tr><td></td><td colspan="6">Current Year</td></tr>
<tr><td rowspan="5">Previous Year</td><td>Performance Levels</td><td>Exceeds</td><td colspan="2">Met</td><td colspan="2">Not Met</td></tr>
<tr><td>Exceeds</td><td>X</td><td colspan="2">-</td><td colspan="2">-</td></tr>
<tr><td rowspan="2">Met</td><td rowspan="2">X</td><td>Score Gain
< 0</td><td>Score Gain
≥ 0</td><td colspan="2" rowspan="2">-</td></tr>
<tr><td>-</td><td>X</td></tr>
<tr><td rowspan="2">Not Met</td><td rowspan="2">X</td><td colspan="2" rowspan="2">X</td><td>Score Gain
≤ 0</td><td>Score Gain
> 0</td></tr>
<tr><td>-</td><td>X</td></tr>
</table>

Figure 1. Nebraska Decision Matrix for Including Growth from Alternate Assessments.

[2] Survey of Council of Chief State School Officers (CCSS) Accountability Systems and Reporting (ASR) State Collaborative on Assessment and Student Standards (Standards) presentation, June, 2016.

Another example comes from Florida, which rewarded progress on the Florida Alternate Assessment (FAA) in light of nine performance levels falling within three overarching performance categories: Emergent (Levels 1,2,3), Achieved (Levels 4,5,6), and Commended (Levels 7,8,9). To determine whether a student had demonstrated growth for purposes of accountability, a series of decision rules were established comparing performance in the current year to that of the previous year. Specifically, students who scored in Level 1, 2 or 3 on the prior assessment were labelled as having demonstrated growth if they moved up a performance level or stayed within the same performance level but increased their total score by 5 or more points. Similarly, students who scored Level 4 or higher on the prior year assessment and maintained their level or scored higher in the current year were considered to have made growth.[3]

Gain Score

There are a variety of states utilizing assessments that provide for a vertical scale, which allows for the calculation of individual gain scores (Delaware, Hawaii, New Mexico, Ohio, South Carolina, and Wyoming). However, we found no evidence that gain scores were being calculated for student-level reporting or school accountability.

Normative

Normative growth is calculated based on a student's relative position within a distribution of performance defined by his/her academic peers. Business rules are typically established to determine the range of values (e.g. growth percentiles) reflecting growth that is greater/less than that expected or considered reasonable.

Michigan's alternate assessment (MI-Access) has 3 tiers based on a student's level of disability and functioning. Student Growth Percentiles (SGPs) are calculated for students taking the highest level of the alternate assessments (i.e., the functional independence level), as this is the only level that provides for a scaled score. For a given student, the SGP is calculated relative to all students in Michigan who had comparable achievement scores on prior state-level MI-Access tests at the functional independence level.

In summary, we found examples of states that provide some growth information on their alternate assessments. For many of the examples, it is less clear if or how the state uses the growth information or if it is included in accountability determinations. It is beyond the scope of this paper to evaluate the efficacy or technical defensibility of any of the examples noted.

Analyses and Results

To evaluate some of the empirical criteria for common growth models described in previous sections we analyze the following for grades 4 and 7 in ELA and mathematics for the spring 2015 NCSC administration:

[3] In 2016, the Florida Standards Alternate Assessment (FSAA) was administered for the first time. FDE has not yet committed to a particular growth model for this new assessment.

- Histogram of scaled scores for all states that tested in spring 2015 and individually for the three largest states
- Tables of relative frequency by scale score and associated conditional standard error of measurement (CSEM) for all states and the three largest states that tested in spring 2015

These analyses are produced to inspect the sample size along the range of the scale, the extent to which data are sparse for any region of the scale and the precision of the scale across the full range.

Findings reveal:

- There are clusters of scores at the Lowest Obtainable Scale Score (LOSS), but this doesn't appear to be a floor effect as much as it reflects incomplete administrations. This will be addressed in more detail in the subsequent section.
- There is no evidence of ceiling effects, insofar as there are not clusters of scores at or near the Highest Obtainable Scale Score (HOSS).
- The density of the distribution is primarily in the range of 1220-1260 for grade four and slightly more negatively skewed—such that the density is primarily in the range of 1230-1260—for grade 7. Each of the state distributions analyzed appear to mirror the shape of the overall distribution. Theses ranges generally correspond to 'high performance level 1' to 'low performance level 2.'
- The areas of the distribution with less density correspond with the areas of the scale where measurement error is highest. Particularly in the scale region below 1210 and above 1270, the CSEMs are much higher relative to the error in the regions near the performance level cuts. The error is more pronounced for the higher scores in ELA compared to math.
- Sample sizes for the lower and upper regions of the scale are quite low for the distribution that includes all states and especially for the distributions of each of the three individual states examined. Only the most dense areas of the distribution – approximately 1220-1260 for most tests – consistently include 30 or more students per score. For any one state, that region is much narrower. There are approximately 10 or fewer score points on the distributions for only the largest states with n-sizes at or above 30.

Grade 4 ELA

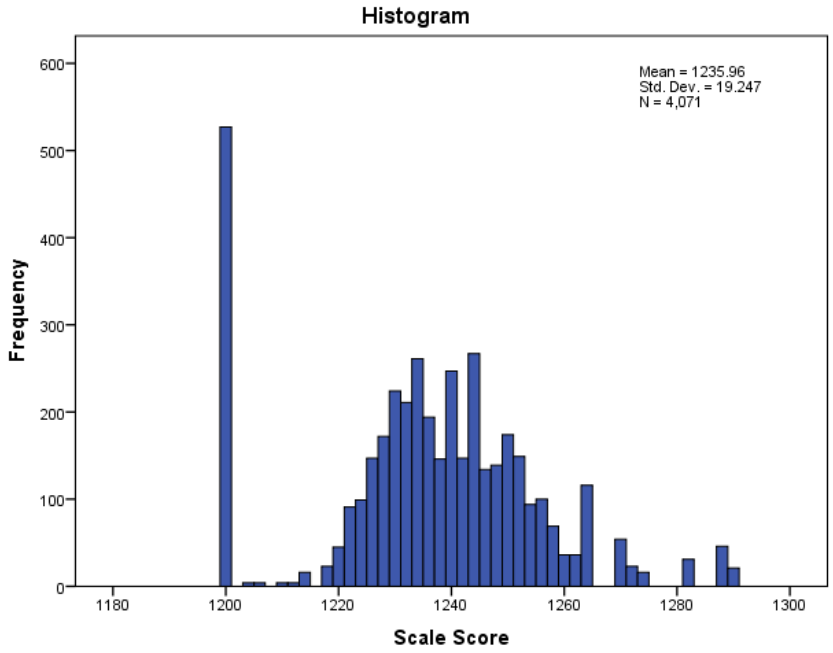

Figure 2. Grade 4 ELA Distribution - All States

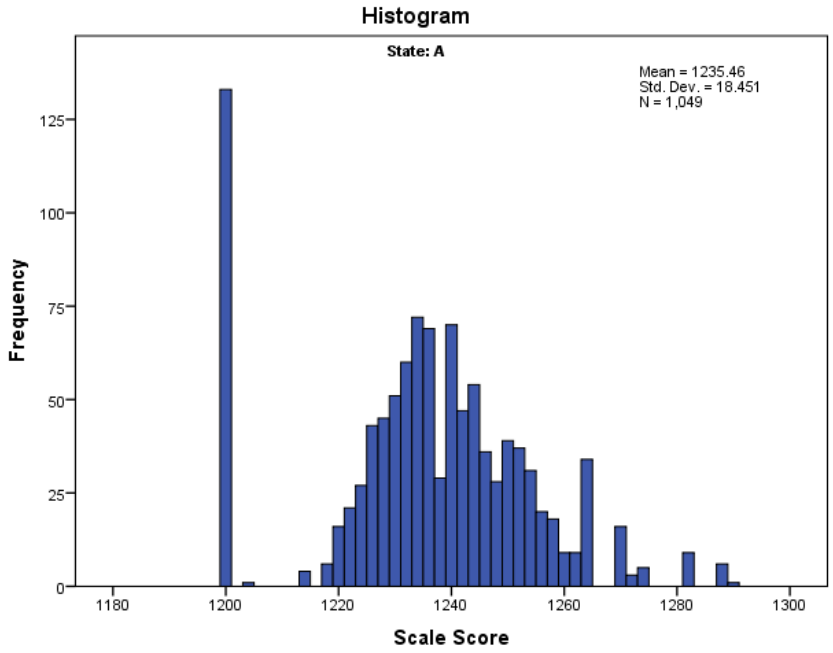

Figure 3. Grade 4 ELA Distribution - State A

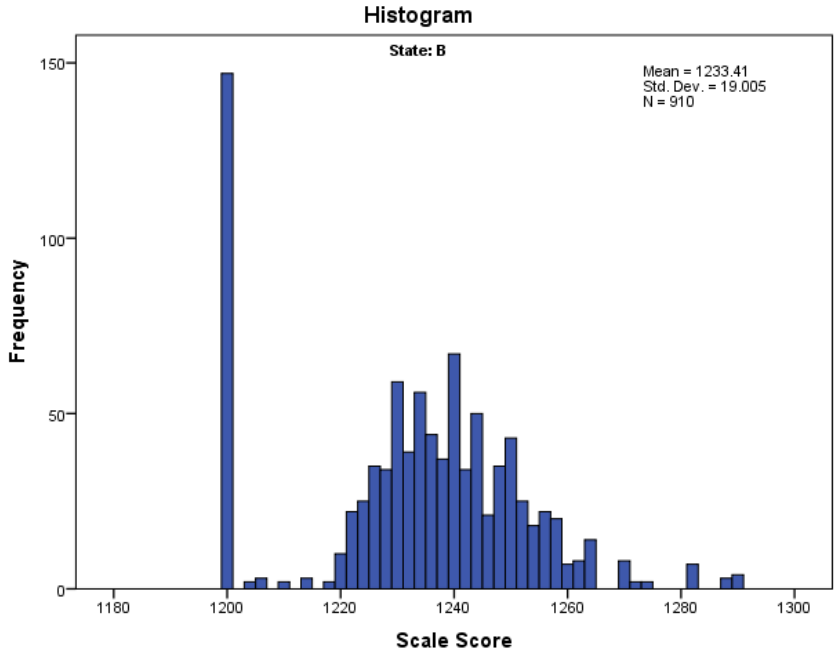

Figure 4. Grade 4 ELA Distribution - State B

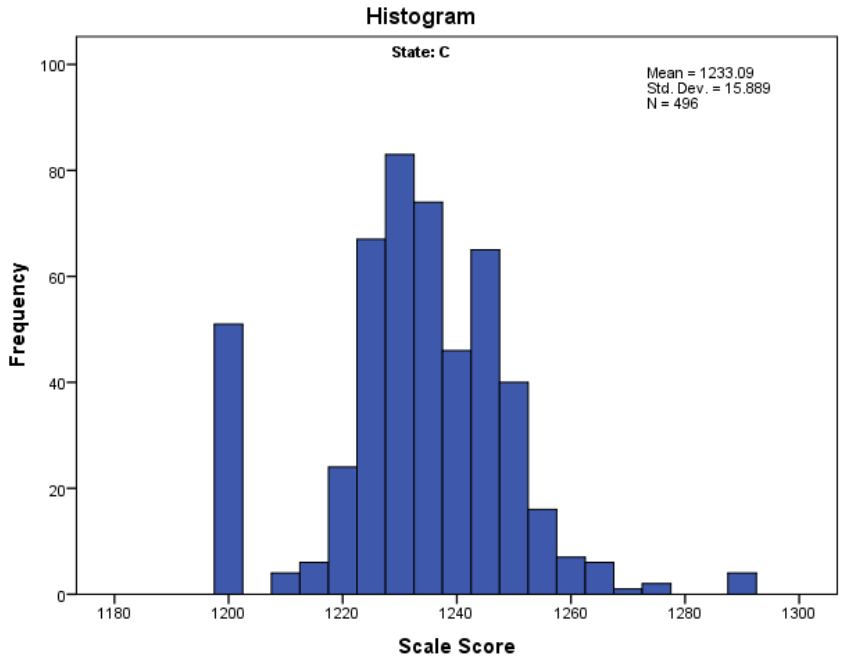

Figure 5. Grade 4 ELA Distribution - State C

Table 2. Grade 4 ELA Relative Frequency Distribution and Conditional Standard Error

ELA Score	All	State A	State B	State C	CSEM
1200	527	133	147	51	16.69
1204	4	1	2	0	8.70
1206	4	0	3	0	7.59
1210	4	0	2	2	6.71
1211	4	0	0	2	6.23
1214	16	4	3	2	5.40
1217	16	4	1	4	4.68
1218	7	2	1	1	4.64
1220	45	16	10	6	4.14
1222	91	21	22	17	3.74
1224	99	27	25	17	3.44
1226	147	43	35	25	3.22
1227	126	33	26	25	3.07
1228	46	12	8	8	2.99
1229	141	31	35	29	2.92
1230	83	20	24	16	2.78
1231	143	34	28	23	2.85
1232	68	26	11	7	2.84
1233	167	47	28	26	2.76
1234	94	25	28	15	2.70
1235	93	26	33	4	2.82
1236	101	43	11	14	2.67
1237	118	23	30	15	2.85
1238	28	6	7	4	2.64
1239	159	48	36	18	2.86
1240	88	22	31	7	3.14
1241	63	20	15	8	2.77
1242	84	27	19	9	3.30
1243	100	19	28	14	2.89
1244	167	35	22	23	3.35
1245	36	13	7	5	3.11
1246	98	23	14	11	3.35
1247	98	18	28	12	3.76
1248	41	10	7	5	3.54
1249	174	39	43	18	4.07
1251	108	27	18	15	4.07
1252	41	10	7	2	4.44

1253	49	16	7	3	5.44
1254	45	15	11	3	4.68
1255	52	9	11	5	4.72
1256	48	11	11	4	5.50
1257	33	11	7	1	6.61
1258	36	7	13	2	5.87
1259	36	9	7	3	5.54
1261	36	9	8	2	7.33
1263	31	10	4	2	8.40
1264	85	24	10	4	7.41
1269	27	5	4	0	10.48
1270	27	11	4	1	11.12
1272	23	3	2	0	9.40
1273	16	5	2	2	11.88
1281	31	9	7	0	15.94
1288	46	6	3	4	19.07
1290	21	1	4	0	19.07
Total	4071	1049	910	496	

Grade 4 Math

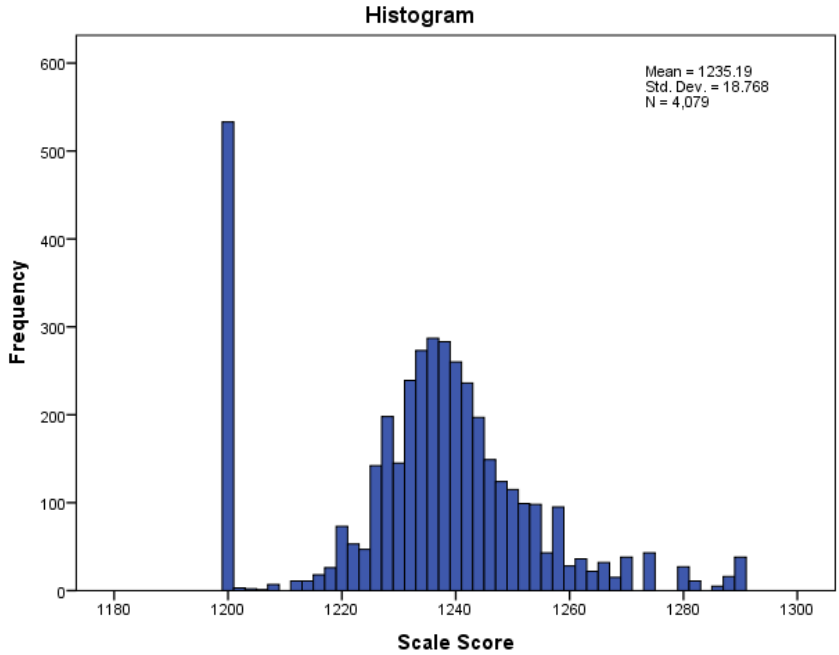

Figure 6. Grade 4 Math Distribution - All States

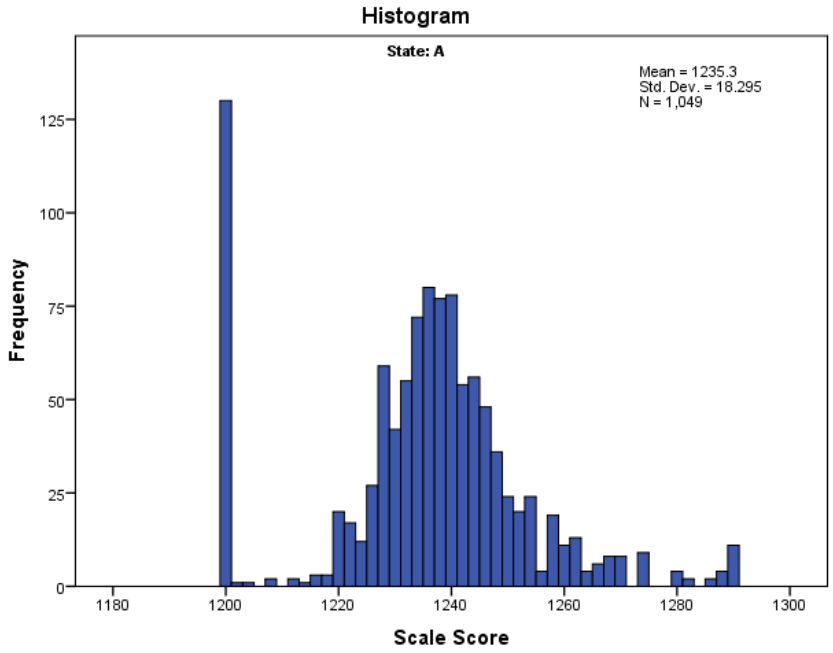

Figure 7. Grade 4 Math Distribution - State A

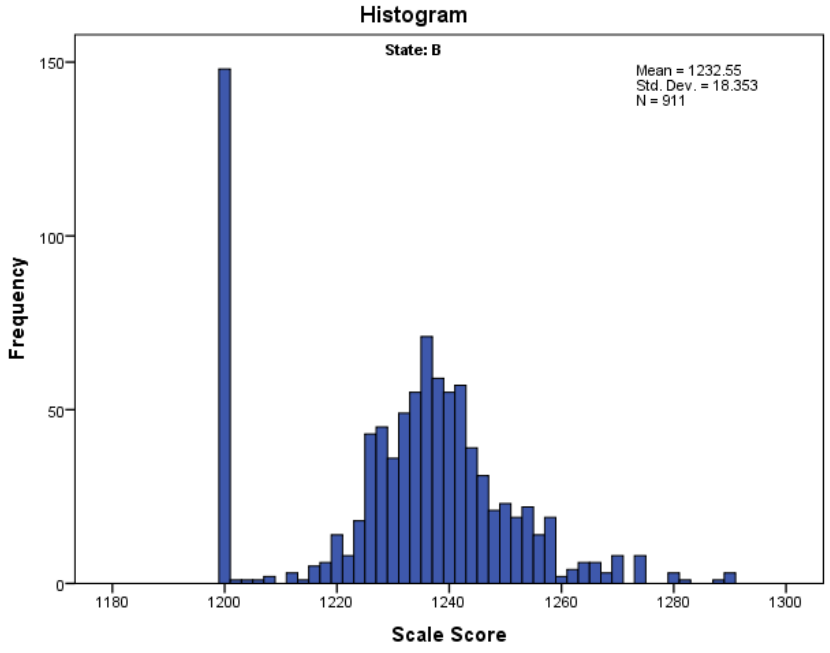

Figure 8. Grade 4 Math Distribution - State B

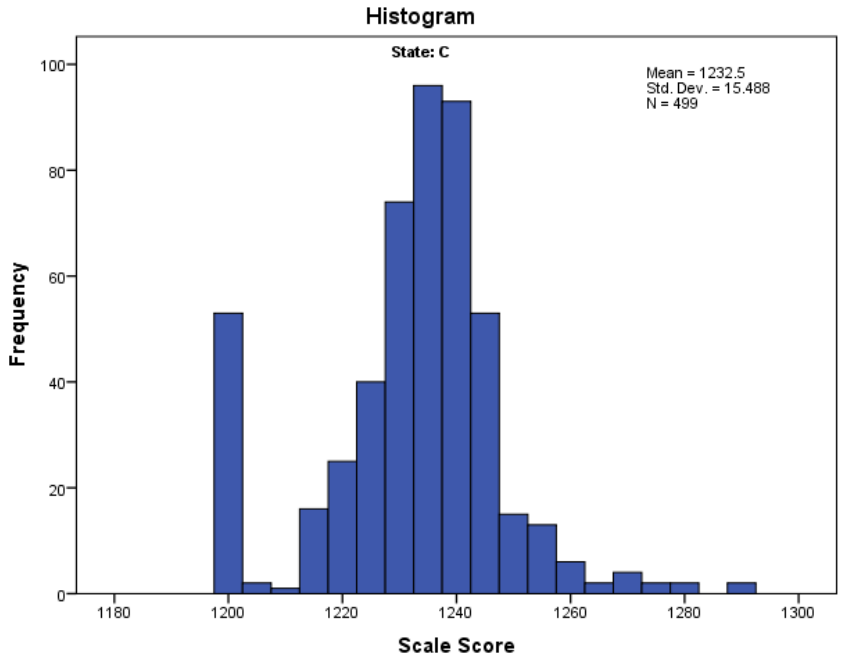

Figure 9. Grade 4 Math Distributions - State C

Table 3. Grade 4 Math Relative Frequency Distribution and Conditional Standard Error

Math Score	ALL	State A	State B	State C	CSEM
1200	533	130	148	52	18.78
1202	3	1	1	1	10.64
1204	2	1	1	0	9.80
1205	1	0	1	0	9.73
1207	3	0	1	2	9.21
1208	4	2	1	1	8.73
1211	6	1	2	0	8.42
1212	5	1	1	0	7.98
1213	11	1	1	6	7.53
1215	2	0	1	0	7.53
1216	16	3	4	4	7.14
1217	26	3	6	6	6.78
1219	37	12	8	5	6.70
1220	36	8	6	8	6.21
1222	53	17	8	12	6.21
1223	47	12	18	3	5.77
1225	102	18	35	15	5.64

1226	40	9	8	7	5.57
1227	93	30	21	15	5.50
1228	105	29	24	16	5.15
1230	145	42	36	14	5.07
1231	68	11	12	12	5.14
1232	171	44	37	32	4.86
1233	74	18	13	15	5.02
1234	199	54	42	30	4.71
1235	83	25	20	13	4.97
1236	204	55	51	28	4.61
1237	75	23	15	10	4.97
1238	208	54	44	29	4.56
1239	131	43	28	16	4.65
1240	129	35	27	13	4.71
1241	62	13	15	8	4.31
1242	174	41	42	27	4.89
1243	49	17	9	8	4.41
1244	148	39	30	21	5.01
1245	56	21	9	9	4.57
1246	93	27	22	6	5.24
1247	90	27	21	9	4.91
1248	34	9	0	2	5.17
1249	64	14	17	5	5.48
1250	51	10	6	4	5.22
1251	76	14	18	4	5.76
1252	23	6	1	0	5.44
1253	34	6	9	2	5.60
1254	64	18	13	5	6.11
1255	43	4	14	3	5.89
1257	77	15	15	3	6.50
1258	18	4	4	0	6.30
1260	28	11	2	4	7.08
1261	36	13	4	2	6.95
1264	22	4	6	2	7.80
1265	32	6	6	0	7.69
1268	15	8	3	1	8.58
1269	38	8	8	3	8.75
1273	13	2	4	0	9.78
1274	30	7	4	2	10.05
1279	9	0	1	1	11.36
1280	18	4	2	1	11.64

1281	11	2	1	0	12.22
1286	5	2	0	0	13.43
1288	16	4	1	1	14.06
1290	38	11	3	1	15.39
Total	4079	1049	911	499	

Grade 7 ELA

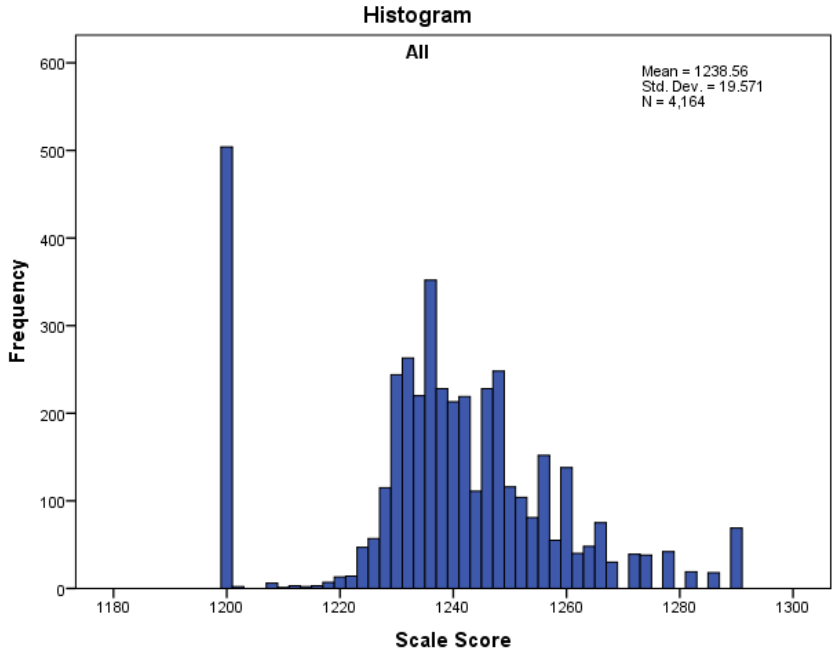

Figure 10. Grade 7 ELA Distribution - All States

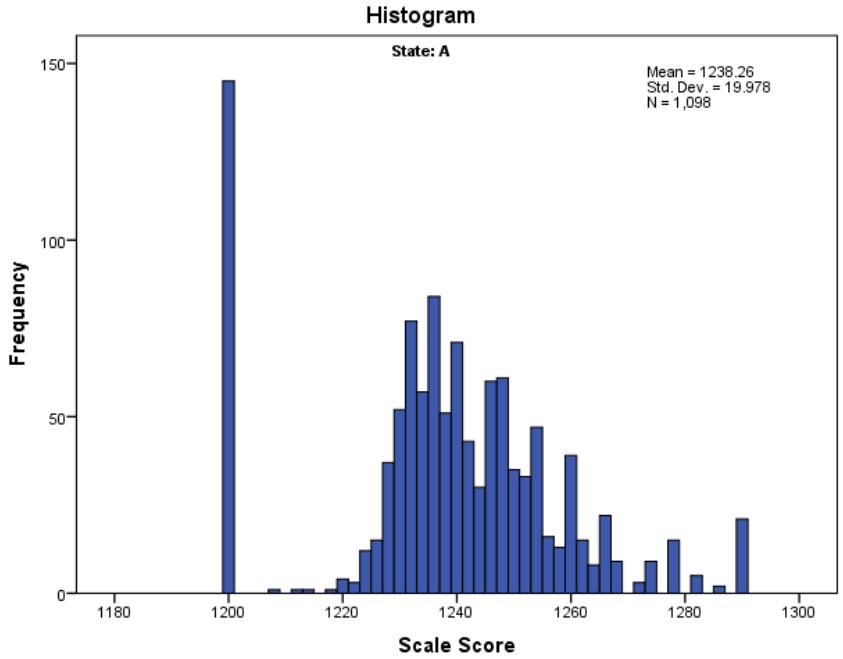

Figure 11. Grade 7 ELA Distribution - State A

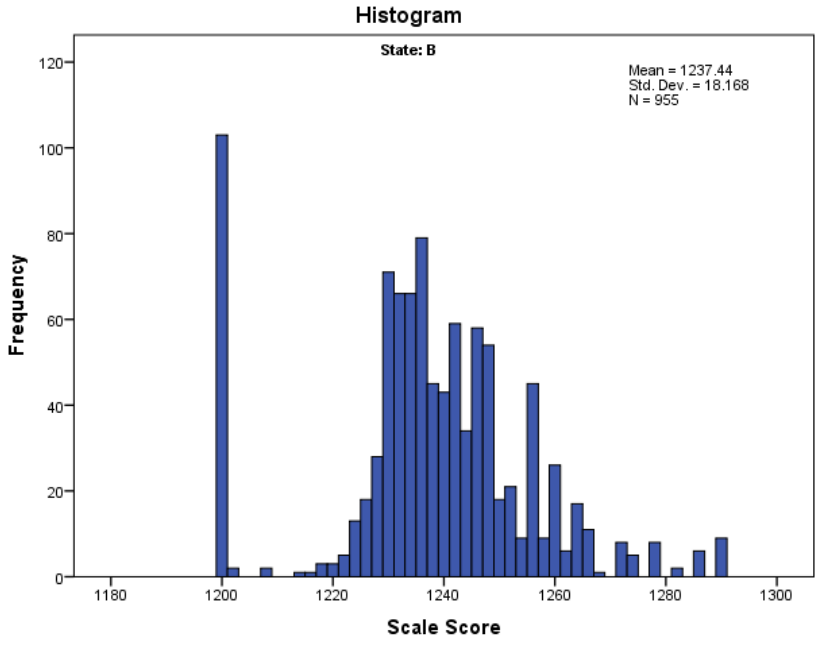

Figure 12. Grade 7ELA Distribution - State B

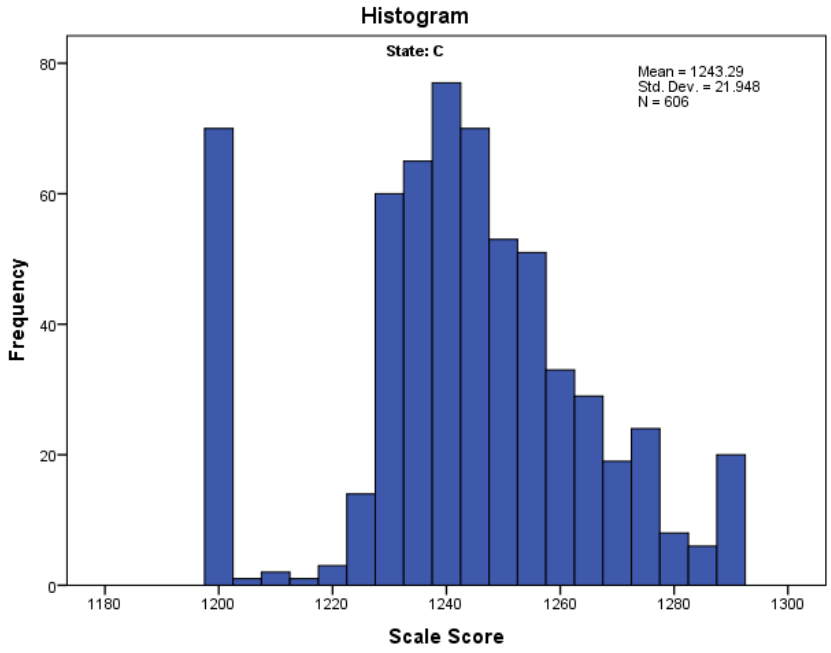

Figure 13. Grade 7 ELA Distribution - State C

Table 4. Grade 7 ELA Relative Frequency Distribution and Conditional Standard Error

ELA Score	All	State A	State B	State C	CSEM
1200	504	145	103	70	18.61
1201	2	0	2	0	10.82
1207	6	1	2	1	8.74
1209	1	0	0	0	7.58
1212	3	1	0	2	6.73
1214	2	1	1	0	6.10
1216	3	0	1	0	5.31
1217	7	1	3	1	5.15
1219	3	0	1	1	4.56
1220	10	4	2	0	4.30
1221	4	0	2	1	4.06
1222	10	3	3	1	3.77
1223	16	4	3	2	3.65
1224	31	8	10	1	3.31
1225	6	4	2	0	3.47
1226	51	11	16	3	3.03
1227	66	21	14	8	3.00

1228	49	16	14	5	3.02
1229	131	32	39	15	2.84
1230	113	20	32	13	2.69
1231	123	39	31	11	2.81
1232	140	38	35	16	2.79
1233	147	39	38	8	2.76
1234	73	18	28	7	3.01
1235	130	40	26	15	2.96
1236	222	44	53	26	3.08
1237	83	23	14	9	3.11
1238	145	28	31	16	3.41
1239	135	27	31	21	3.55
1240	78	44	12	4	3.61
1241	102	20	28	12	3.64
1242	117	23	31	24	4.11
1243	111	30	34	13	3.92
1245	189	54	47	23	4.43
1246	39	6	11	8	4.39
1247	117	30	25	26	4.71
1248	131	31	29	20	5.12
1250	116	35	18	18	5.32
1251	104	33	21	15	5.77
1253	47	13	9	8	5.81
1254	34	34	0	0	6.54
1255	152	16	45	31	6.54
1257	55	13	9	12	6.75
1259	101	28	18	22	7.53
1260	37	11	8	6	8.10
1262	40	15	6	5	8.03
1264	48	8	17	12	8.61
1266	75	22	11	17	9.85
1268	30	9	1	9	9.85
1271	39	3	8	10	10.69
1274	38	9	5	15	11.95
1277	42	15	8	9	13.20
1281	19	5	2	8	14.22
1285	18	2	6	6	15.87
1289	23	9	1	6	16.96
1290	46	12	8	14	18.84
Total	4164	1098	955	606	

Grade 7 Math

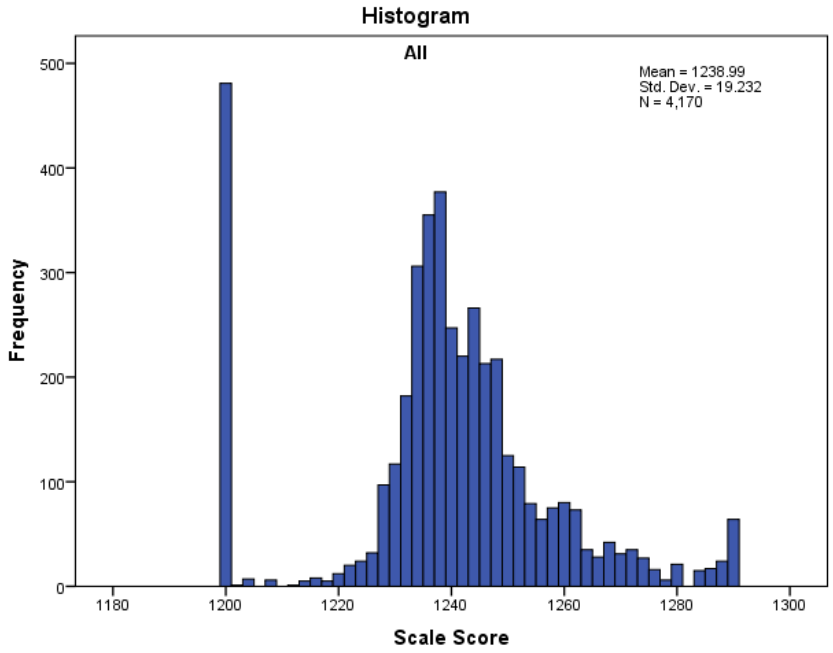

Figure 14. Grade 7 Math Distribution - All States

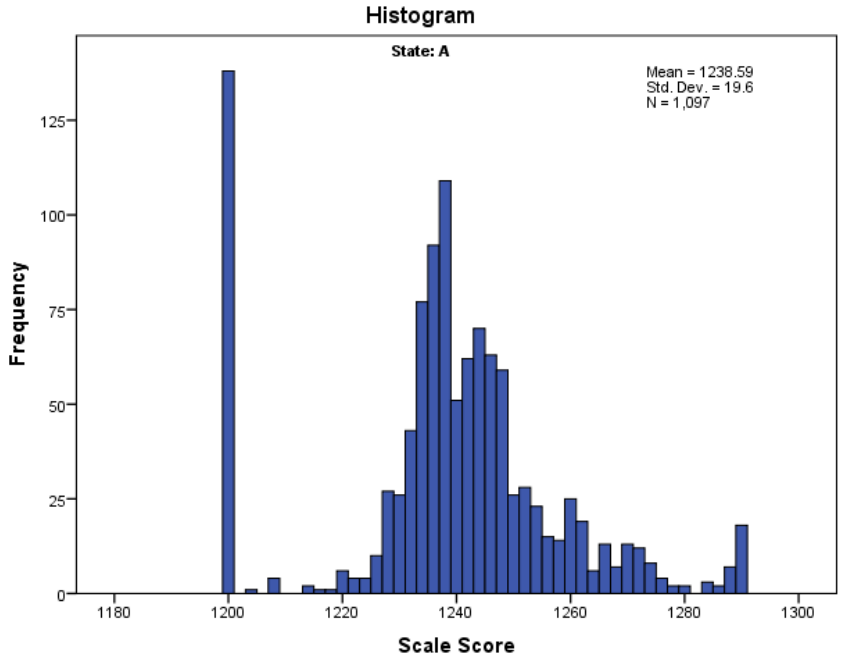

Figure 15. Grade 7 Math Distribution - State A

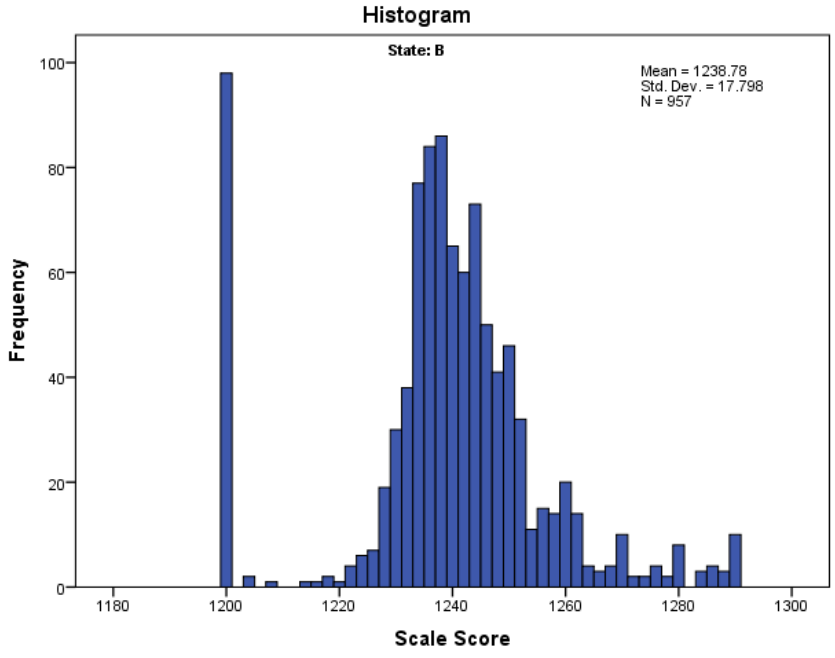

Figure 16. Grade 7 Math Distribution - State B

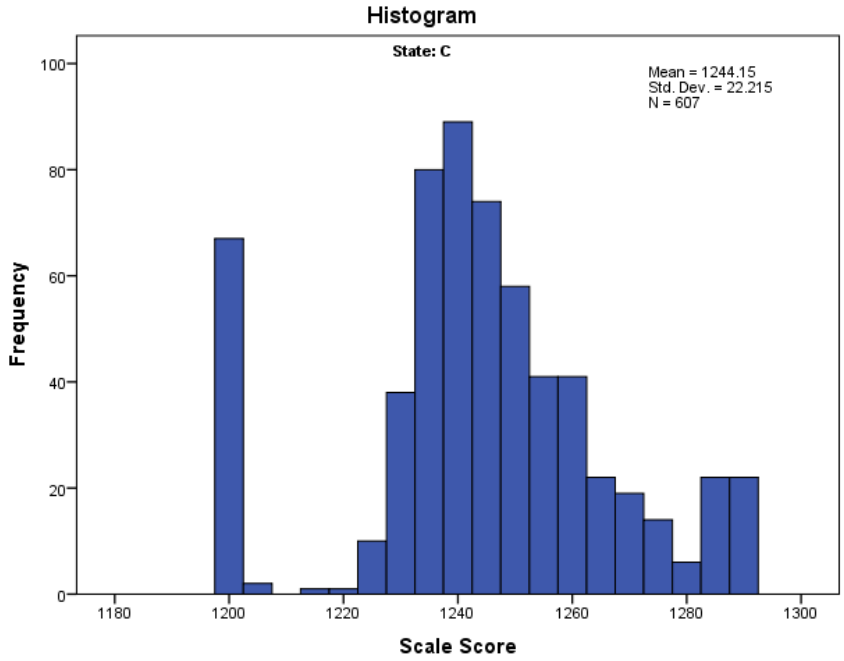

Figure 17. Grade 7 Math Distribution - State C

Table 5. Grade 7 Math Relative Frequency Distribution and Conditional Standard Error

Math Score	All	State A	State B	State C	CSSEM
1200	481	138	98	67	17.72
1201	1	0	0	0	10.58
1204	7	1	2	2	9.69
1207	4	2	1	0	8.51
1208	2	2	0	0	8.25
1212	1	0	0	0	7.27
1213	2	0	0	0	6.98
1214	3	2	1	0	6.46
1215	1	0	0	1	6.59
1216	7	1	1	0	6.33
1217	3	1	2	0	5.74
1218	2	0	0	0	5.89
1219	11	6	1	1	5.80
1220	1	0	0	0	5.24
1221	6	0	1	0	5.41
1222	14	4	3	0	5.24
1224	24	4	6	5	5.12
1225	10	5	2	0	4.61
1226	22	5	5	3	4.93

1227	20	6	3	2	4.64
1228	77	21	16	6	4.60
1229	28	5	7	2	4.73
1230	89	21	23	6	4.48
1231	49	5	7	10	4.63
1232	133	38	31	14	4.39
1233	172	43	44	18	4.31
1234	134	34	33	18	4.58
1235	120	35	33	6	4.15
1236	235	57	51	29	4.37
1237	73	24	17	9	4.26
1238	304	85	69	34	4.34
1239	181	28	51	23	4.39
1240	66	23	14	8	4.26
1241	159	42	47	16	4.42
1242	61	20	13	8	4.28
1243	216	50	62	37	4.44
1244	50	20	11	3	4.26
1245	149	48	34	18	4.59
1246	64	15	16	7	4.59
1247	98	30	18	9	4.57
1248	119	29	23	15	4.79
1249	30	9	12	5	4.64
1250	95	17	34	22	4.92
1251	28	6	5	4	4.82
1252	86	22	27	12	5.09
1253	12	12	0	0	5.17
1254	67	11	11	15	5.17
1255	40	9	8	5	5.34
1256	24	6	7	7	5.32
1257	57	9	13	14	5.65
1258	18	5	1	6	5.60
1259	57	17	16	7	5.86
1260	23	8	4	4	5.99
1261	21	5	2	7	6.01
1262	52	14	12	17	6.37
1263	19	4	2	4	6.48
1264	16	2	2	5	6.51
1265	28	13	3	7	6.89
1267	22	5	2	6	7.14
1268	20	2	2	8	7.18

1269	31	13	10	3	7.58
1271	18	5	0	5	8.06
1272	17	7	2	3	8.11
1273	27	8	2	9	8.56
1276	16	4	4	4	9.46
1277	6	2	2	1	9.53
1279	21	2	8	6	10.06
1284	15	3	3	8	11.91
1285	17	2	4	7	11.98
1287	24	7	3	7	12.69
1290	64	18	10	22	16.08
Total	4170	1097	957	607	

Discussion and Implications

In this paper, we set out to examine promising practices for including NCSC scores in measures of academic growth. We acknowledge this is a priority for many states, particularly in light of requirements under ESSA that call for valid, reliable, and comparable growth measures that include results from the AA-AAS.

We opened with a typology for growth models based on four categories. This established a foundation for a review of growth criteria, highlights from the literature on growth with AA-AAS, and an illustration of some current state practices. This was followed by data analyses to examine the 2015 NCSC results with respect to the empirical criteria associated with scale based models. In this section, we will review the results of those analyses and offer some recommendations, reflecting on the four growth categories. We will close with some suggestions for further research.

The analyses show that the spring 2015 NCSC assessment does not appear to exhibit ceiling effects insofar as we do not observe clustering near the HOSS. Although we find clustering at the LOSS, these scores appear to be overwhelmingly associated with students who attempted zero or very few items, as opposed to students who consistently responded incorrectly. For this reason, the cluster at the lowest score point is probably better described as non-participants, rather than evidence of a floor effect.

However, the error near the LOSS and HOSS was very high for the four tests examined. This suggests the test is not very precise for disentangling degrees of performance for very low or very high achieving students. Moreover, sample size was very low for scores across the scale range except for the most dense region—approximately in the range of 1220-1260. This was true with all participants included and especially true for analyses by state.

Taken together, these findings suggest that methods for estimating academic growth that rely on scale scores or conditioning across the full range of the scale may produce unreliable or uncertain results, particularly for any one state. Because VAM and normative approaches condition across the full range of scale scores, we do not recommend including NCSC results in these models absent additional information that model criteria are satisfied. Furthermore, gain score models typically utilize the full scale range and also require a vertical scale, which is not a feature of NCSC at this time. For this reason, the gain score approach is not a promising alternative for NCSC until and unless the necessary criteria are met.

In the near term, we believe an approach that is less dependent on sample size and not based on information along the full scale is the most promising alternative. Categorical growth approaches (e.g. value tables or transition models) as presented earlier in this paper and in use by several states, can be created to satisfy these requirements. If this option is implemented with NCSC, we recommend constructing the model with a relatively modest number of categories, especially for levels one and four, where the error is particularly limiting.

Moving forward, we believe scale-based models may be feasible if sample size is increased by using multiple state data and measurement error is reduced by adding discriminating items to the form in the most impoverished regions of the scale.

Whatever approach is selected, we suggest additional analyses with multiple years of data to evaluate outcomes. Some important questions to evaluate include the following:

- *Are results reliable?* Reliability refers to the stability of a measure. An evaluation plan should include tracking the consistency of growth estimates for students and across aggregate units that will be reported (e.g. school, district, if applicable) across years.
- *Are results sensitive to meaningful differences?* This refers to evidence that growth estimates differentiate outcomes for students where there is a credible expectation of differential performance (e.g. we would expect students who receive strong instruction and support to grow at increased rates.)
- *Are results related to variables that should not be associated with effectiveness?* Growth outcomes should not have strong positive relationships with factors unrelated to academic progress. For example, growth estimates should be weakly related to a student's prior year achievement.

Additional evaluation of growth alternatives should take into account the extent to which results are easy to understand and are useful for improvement.

Finally, we recommend exploring content-based approaches to describing student progress on NCSC. This refers to an approach that seeks to identify knowledge and skills associated with performance in specified ranges of the NCSC scale (more 'granular' than performance levels). By so doing, a student's progress can be reported qualitatively in terms of skills and expertise the student exhibits with respect to the content assessed each year. Certainly, such an approach must

be informed by a program of research that includes identifying the items that most discriminate in each score region to support claims about the nature of knowledge and skills exhibited in each score category.

We acknowledge that content-based growth information may not fit into a state's federal accountability model. However, the information could be helpful to better understand and improve student achievement.

References

Ahearn, E. (2009). Growth models and students with disabilities: Report of state interviews. Alexandria, VA: Project Forum, National Association of State Directors of Special Education.

Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. Educational Researcher, 39, 537–544.

Castellano, K. E., & Ho, A. D. (2013). A practitioner's guide to growth models. Washington, DC: Council of Chief State School Officers.

Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. (2005). Using value tables to explicitly value student growth, conference on longitudinal modeling of student achievement. Dover, NH: Center for Assessment. Retrieved from http://www.nciea.org/publications/MARCES_RH07.pdf

Karvonen, M., Flowers, C., & Wakeman, S. Y. (2013, April/May). An exploration of methods for measuring academic growth for students with significant cognitive disabilities. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. Review of Educational Research, 71(1), 53–104.

Sireci, S. G., Scarpeti, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. Review of Educational Research, 75, 457–490.

Tindal, G., Nese, J.F.T., Farley, D., Saven, J.L., Elliott, S.N. (2015) Documenting Reading Achievement and Growth for Students Taking Alternate Assessments, Exceptional Children, vol. 82(3), 321-336.

Appendix – Index of Tables and Figures