MEASURING GROWTH WITH THE *IOWA ASSESSMENTS*™

A Black and Gold Paper

Abstract

The primary interpretations, statistical foundations, and data for the Iowa Growth Model are described in this overview for practitioners. Use of growth measures on individuals and groups for student and program evaluation is discussed and illustrated with sample data and reports.

Catherine Welch, PhD and Stephen Dunbar, PhD

The University of Iowa

ITP Iowa Testing Programs

March 2014

AUTHORS

Catherine Welch, PhD

Catherine Welch is a professor of Educational Measurement and Statistics at the University of Iowa. She teaches graduate-level courses in educational measurement and conducts research in the areas of test design, interpretation, and growth. Catherine has responsibilities with Iowa Testing Programs, where she is the director of statewide testing for the *Iowa Assessments* and the Iowa End-of-Course Assessments.

Stephen Dunbar, PhD

Stephen Dunbar is the Hieronymus-Feldt Professor of Educational Measurement in the College of Education at the University of Iowa, where he has taught since 1982, and also serves as Director of Iowa Testing Programs. His primary research interests are in the areas of test development and technical applications in large-scale assessment. He is a principal author of the *Iowa Tests of Basic Skills* and the *Iowa Assessments*.

MEASURING GROWTH WITH THE *IOWA ASSESSMENTS*

Understanding how students change and grow over time is becoming increasingly important as teachers and schools design education programs tied to improvement relative to core standards in Reading, Mathematics, and other key content areas. The models and approaches used to measure student growth, however, are complex and at times overwhelming to the practitioner interested in the answer to a seemingly simple question: "How is my student growing in relation to other students, and is the growth s/he has achieved in line with what should be expected?"

In an assessment world where growth is discussed with terms such as "value added," "residual gain," "student growth percentile," and "multivariate projection," it can be easy to lose sight of the idea that any approach to growth should answer the practitioner's very simple question.

The Iowa Growth Model provides answers to important questions about student growth and changes to groups over time with a descriptive framework based on many years of research and development associated with the *Iowa Assessments*. Student growth information based on the Iowa Growth Model can be readily used for a variety of purposes in which the primary interpretation involves gain and improvement over time. Growth data based on the Iowa model are also amenable to various approaches for secondary analyses and scores that feed into proprietary methods.

This paper provides an overview of the Iowa Growth Model, including its primary interpretation, its validity and statistical foundation, its growth scale metric, its data requirements, and its use in evaluation contexts. It also includes a comparison of the Iowa Growth Model to other growth models. Finally, examples are provided of growth reports and other visual displays that demonstrate the utility of the model and the simplicity of the type of information derived from it.

DESCRIPTION AND PRIMARY INTERPRETATION

The Iowa Growth Model uses an underlying vertical score scale—the National Standard Score (NSS)—that permits several approaches to describing growth. It is a metric that ranges numerically from 80 to 400 and spans a developmental continuum from Kindergarten to Grade 12 in major content domains such as Reading, Mathematics, Science, and Written Expression.

National research studies in the 2010–11 school year were conducted to validate the reference points on the NSS scale representing the medians for each grade level and the model-based inferences about the amount of growth typical of students at different achievement levels. The primary interpretations supported by the NSS scale have to do with (1) how much a student grows from one assessment occasion to the next compared to his or her assessment peers (a relative growth interpretation), and (2) how much growth would be expected for this student's assessment peers (a normative growth interpretation). This basic information about growth can be used for a variety of purposes in student and program evaluation, such as individual and group decisions about instructional interventions and responses to interventions that can be gauged by the amount of growth achieved. The development of the NSS scale is detailed in Appendix A.

Another key feature of the Iowa Growth Model and its backbone, the NSS scale, is the ability to track student growth over time to determined levels of proficiency or to research-based performance benchmarks that indicate college and career readiness. The model defines a longitudinal trajectory that, at any given point in a student's educational development, can be used to determine whether a student is on track to achieve such benchmarks. The performance benchmark for the college and career interpretation of growth is the probability of student success in credit-bearing coursework in postsecondary education (Furgol, Fina, & Welch, 2011; Welch & Dunbar 2011).

VALIDITY FRAMEWORK AND STATISTICAL FOUNDATION OF GROWTH METRICS

The validity framework for a growth model involves fundamental considerations about the content of the assessments used to measure growth, the scale and modeling requirements, the definition of targets that represent typical grade-level performance or other benchmarks such as college readiness, and the utility of information leading to sound interpretations of student growth and effective decisions about enhancing growth for individuals and groups.

Validity. The validity of the interpretation and use of information is "the most fundamental consideration in developing and evaluating tests" (AERA, APA, NCME, 1999, p. 9). In the current context, validity pertains to evidence that supports interpretations relative to growth. With the assessment imperative of college- and career-readiness at the forefront of efforts to reform education, a critical aspect of validity arguments for related claims involves the underlying model used to measure and report growth and change. Conceptual frameworks for understanding student growth are evolving rapidly, and interpretations of growth are critical for their success in statewide testing programs and the assessment consortia (e.g., Castellano & Ho, 2013; Betebenner, 2010; Reardon & Raudenbush, 2009). For any growth model, basic validity considerations encompass relevant evidence that ranges from the content definition of the domain to the utility of growth information provided in reports of results. Regardless of the approach to growth, general validation concerns remain. Table 1 summarizes several of these issues as they define a validity framework for growth.

Table 1: Examples of Validity Evidence Related to the Measurement of Growth	
Validity Evidence	**Consideration for Growth**
Content validity evidence	Content-related validity evidence is tied to test development. The proposed interpretations of growth and readiness should guide the development of a test and the inferences leading from the test scores to conclusions about a student's readiness. Content alignment studies will serve as the foundation for a trail of evidence needed for establishing the validity of growth and readiness tracking and reporting. Alignment studies will inform the interpretation of growth and readiness research findings.
Scale requirements	Scales or linking studies that allow for the longitudinal estimation and tracking of growth are a necessity in the present context. The scales need to be anchored in terms of both content and student performance within and between grades.
Definition of targets	Targets must exist that quantify the level of growth expected, observed and desired for a given period of time (i.e., fall-to-spring testing; year-to-year testing). For college readiness, targets must also exist that quantify the level of achievement where a student is ready to enroll and succeed in credit-bearing, first-year postsecondary courses. To date, these targets are currently defined by the ACT® Benchmarks, by the College Board Readiness Index, or by individual institutions of higher education.
Collection of concurrent validity evidence	Many tests will claim to measure college readiness, but a plan must be in place for validating that claim. Validity studies should be conducted to determine the relationship between the assessments and the indicators of readiness, including the content of entry-level college courses.
Utility	A primary goal of this information is that students, K–12 educators, policymakers, and higher education representatives can use it to better understand the knowledge and skills necessary for college readiness in English Language Arts and Mathematics. The information must be easily understood and actionable by a broad range of audiences.

Developing a domain and a model for growth starts with defining content standards that describe *continuous* learning. Discrete, granular descriptions of content that are the objectives of small instructional units in, for example, signed-number arithmetic, may be useful in tracking progress toward small unit objectives, but they may not be the best focus for an assessment of growth used to track progress across large spans of time, such as grade-to-grade growth across elementary school years. The five stages of development in Reading (Chall, 1996) are a good example of a learning continuum—there is an underlying construct and a progression that describes how children change from "learning to read" to "reading to learn." In this sense, the learning continuum constitutes a broad definition of the achievement domain and what it means to "grow" with respect to important content standards or guideposts of the domain. The important point is that measuring growth requires test design and development that keeps the focus on the domain.

Assessing a child's growth on a learning continuum involves developing measures aligned to broad content standards and reflecting a level of cognitive complexity appropriate for that child's stage of development. Developmental appropriateness is (1) guided by research and practice in the achievement domain (the major domains of the Common Core State Standards, English Language Arts and Mathematics, represent a good example of broadly defined achievement domains), and (2) established through extensive field testing of assessment materials *in multiple grades*. Valid and reliable measurement of growth requires both.

Statistical Foundation. The vertical scale of the *Iowa Assessments* quantifies and describes student growth over time via a growth metric. One of the defining attributes of the growth metric is that the projection of subsequent performance can be made conditional on prior performance through the vertical scale (Furgol, Fina & Welch, 2011). The expected vertical scaled scores for each grade level and content area on the *Iowa Assessments* are derived from a large national norm group and show the relative standing of students' achievement within the score distinction of students in a national probability sample (Hoover, et al., 2007).

Many tests used to measure yearly growth are vertically aligned and scaled. This means that *each successive test builds upon the content and skills measured by the previous test.* It assures that tests taken over multiple grade levels show a coherent progression in learning. They incorporate several defining technical characteristics (Patz, 2007), including:

- an increase in difficulty of associated assessments across grades,
- an increase in scale score means with grade level, and
- a pattern of increase that is regular and not erratic.

Being tested every year doesn't necessarily mean that the change in scores reflects a year's growth in student achievement. That is where vertical scaling comes in. Tests are developed for different grade levels—for example, for 4th and 5th grades—but scored on the same scale. This way, educators are assured that a change in scores represents a change in student achievement instead of differences in the tests themselves.

GROWTH METRICS

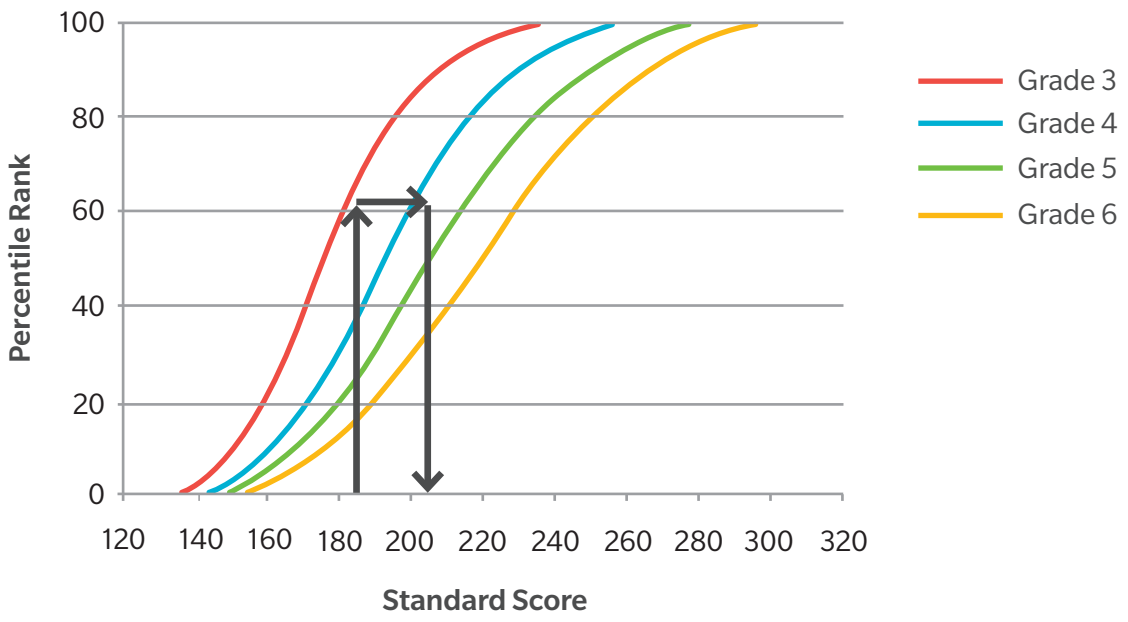
Growth metrics that allow for the longitudinal estimation and tracking of growth are a necessity. The metrics need to be anchored in terms of both content and student performance within and between grades. Three growth metrics are an integral part of the Iowa Growth Model and all three are represented in terms of the National Standard Score (NSS) scale as indicated in Table 2.

Table 2: Growth Metrics Associated with the Iowa Growth Model

Iowa Growth Metrics	Notation	Related Terminology
Expected Growth	$NSS_2 \mid NSS_1$	Estimated Growth
Observed Growth	$NSS_2 - NSS_1$	Gain Scores Change
Observed – Expected	$NSS_2 - (NSS_2 \mid NSS_1)$	Value-Added

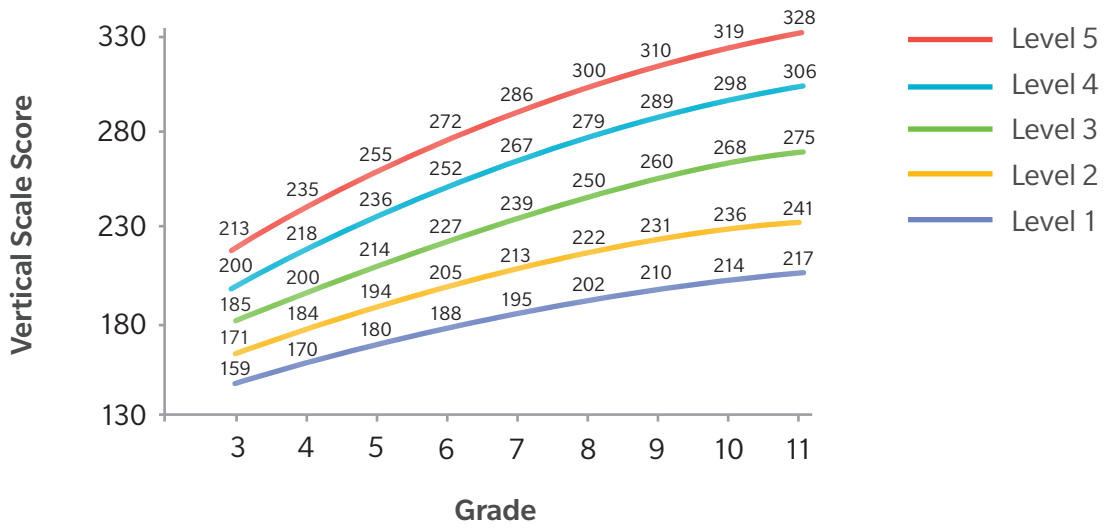
Expected Growth. The Iowa Growth Model defines expected growth as that which was obtained by a nationally representative group of students who took the appropriate assessments at the grade levels of interest. Figure 1 shows the relationship for four grade levels between NSS on the horizontal axis. As one example, consider a Grade 3 student who scores 185 on the reading assessment. As indicated in Figure 1, one year of growth represented by the horizontal arrow intersecting the Grade 4 distribution leads to an expected NSS of 204. Relationships like the one illustrated define for any student at any level of achievement in one grade the expected NSS in a subsequent grade (Cunningham, Welch, Dunbar, 2013).

Figure 1: Establishment of Expected Growth



When a student has grown as much as expected since the previous year, this student is keeping pace with other students in the nation. The growth chart in Figure 2 consists of a series of curves that illustrate the typical pace of performance for five different students that started in 3rd grade at five different points. For each of these students, the expected NSS for subsequent years is identified. For a student who started with an NSS of 159 in Grade 3, an expected growth for Grade 4 would be 170. For a student who started with an NSS of 185, an expected growth for Grade 4 would be 200. More information on the establishment of these curves is provided in Appendix A.

Figure 2: Expected Growth Curves for Five Starting Scores (Levels 1 to 5)

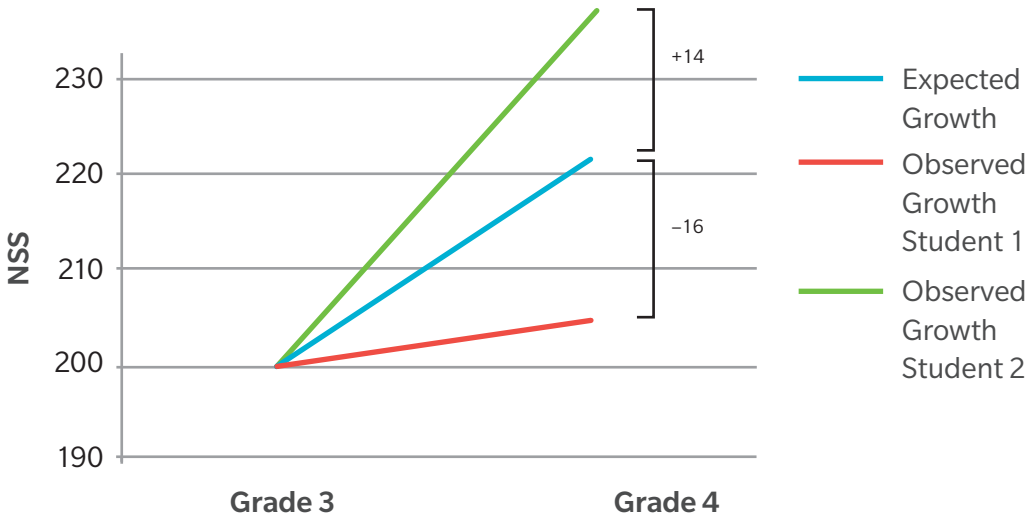


Observed Growth. The observed growth is simply the difference between the second NSS and the first NSS. Observed growth reflects the change in a student's performance between two points of time on the NSS scale. The observed growth is the absolute change in student performance between two time points. These two time points can be from fall to spring, one year to the next, or across multiple years. The sign and magnitude of the observed growth are important in indicating a student's change in performance. The magnitude of the gain indicates how much the student has changed, whereas the sign indicates if the gain is positive (signifying improvement) or negative (signifying decline) (Castellano & Ho, 2013, page 36).

Observed – Expected. The difference between the observed NSS and the expected NSS (given a student's starting point) is frequently seen as value-added. It is the increment of growth that is different than expected. As with the observed growth, the sign of this value is important. If the value is positive, then the student has exceeded expectations in growth. When the value is zero, then the student has met expectations in growth. When the value is negative, then the student has failed to meet expectations for growth.

Figure 3 illustrates the relationship between these three metrics. Two students tested in the fall of 3rd grade, and the observed reading score for both students was 200. For these two students and all other students obtaining a 200 in the fall of 3rd grade, the Iowa Growth Model says that their expected score for fall of Grade 4 is 221. One of the two students obtained a 235 in 4th grade, a 14-point gain over his expected score of 221, meaning he exceeded his growth expectations. The other student, however, obtained a 205 in 4th grade, 16 points short of the expected score of 221, and failed to meet his growth expectations.

Figure 3: Growth Example for Two Students between Grades 3 and 4



DATA REQUIREMENTS AND PROPERTIES OF MEASURES

The Iowa Growth Model supports multiple approaches to the measurement and evaluation of growth. The fundamental data requirement is a test score on the same scale at two points in time. The NSS is a meaningful metric because it is designed to place students on a developmental continuum in the domain of interest and the scale spans the continuum of learning. The *Iowa Assessments* were developed using standard scores that describe a student's location on an achievement continuum, much like a learning progression for a broadly defined content domain. Expectations for a student's annual growth (beginning at any point on the scale) can be established based on intervention and instructional strategies.

USE IN EVALUATION CONTEXTS

Among the many shortcomings of proficiency-based reporting for research and evaluation in schools is the simple labeling of children as "proficient" or "not proficient" that results from it. This labeling does not recognize the fact that all children, regardless of where they start, can show growth in achievement from year to year. How does a focus on growth and a simple growth metric change the way students, schools, districts, or entire states would be viewed in an evaluation context? In this section, several examples of how to appropriately use growth data are discussed. Appendix B contains examples of reports that focus on growth designed for different audiences and purposes.

In the example below, Reading scores in Grades 4 and 5 are illustrated for students with the same NSS score of 200 in Grade 4, a score slightly above the national average for fall testing. According to the Iowa Growth Model, students at this achievement level in Reading in Grade 4 are expected to score 215 in Grade 5 for an "expected" growth score of 15 NSS points.

Student	Grade 4			Grade 5		
	Observed NSS	Expected		Observed		Growth Group
		NSS	Growth	NSS	Growth	
Horatio	200	215	15	225	25	Exceeds
Shayna	200	215	15	212	12	Does Not
Edna	200	215	15	215	15	Meets
Ralphie	200	215	15	205	5	Does Not

The four students in this example all gained in achievement in Reading in Grade 5, but by different amounts as indicated by their observed NSSs in Grade 5. According to the Iowa Growth Model, a student at Horatio's achievement level in Reading in Grade 4 is expected to score 215 in Grade 5 for an "expected" growth score of 15 NSS points. Horatio's observed growth of 25 NSS points is not only greater than his expected growth, but also enough greater that his gain in achievement exceeds the margin of error associated with the Iowa NSS. This information can be presented at the individual student level or at the group level.

Establishing Growth Goals. Expected growth is not the same as a **growth goal** for a student. As stated earlier, expected growth occurs when a student is *keeping pace* with other students that started at a similar point. However, for low-performing students, teachers and parents may want the student to *outpace* other students in the nation. In such situations, the growth goal should be set beyond the expected growth. In other cases, growth toward a predetermined cut-score may be the goal for a student. This cut-score could be a proficiency level as determined by a state or a college readiness cut-score. In either case, growth goals may be established that allow students to be "on track" toward such a cut-score.

Table 3 illustrates the expected growth based on the observed growth for two different starting points in two different grades. For each of these starting points, the expected growth is generated from the Iowa Growth Model; these values are 170 and 194. However, as suggested in Figure 2, the student that begins Grade 3 at 159 and is expected to grow to 170 by Grade 4, may still be falling short of a desirable level of achievement. In this example, the desirable level of achievement may be a proficiency cut-score of 176, meaning a growth goal for this student may be 17 standard score points rather than the expected 11 points. The same situation is relevant for the 4th grade student who started at 184. Although the expected growth is 194, it may be desirable to reach a proficiency goal of 198. Although there are many reasons to accelerate the growth goal beyond the expected growth, the goals should always be established taking into account as much information as possible about the student. This information may include the types of resources that have been provided to the student, additional assessment information, classroom performance, observations, and conferences.

Table 3: Using Expected Growth to Establish Growth Goals – An Example					
Grade	Observed NSS Year 1	Expected NSS Year 2	Expected Growth	Proficiency Cut-score	Growth Goal
3	159	170	11	176	17
4	184	194	10	198	14

RELATIONSHIP TO OTHER GROWTH MODELS

The term "growth model" is used in many achievement contexts, and its meaning is often ambiguous. Ostensibly, different growth models may support similar or very different interpretations depending on the statistical foundation of the model and the metrics used to report its results. The results of the Iowa Growth Model have been compared to two "conditional growth" models using two large (state-level) cohorts of students between 5th grade and 6th grade and again between 6th grade and 7th grade.

The first conditional growth model is based on the Student Growth Percentile (SGP) metric, which describes the extent to which a student has grown relative to peers with similar past test scores (Betebenner, 2009). The SGP metric conditions on prior achievement to describe the rank order of the current achievement of students. The second conditional growth model is based on the Percentile Rank of Residuals (PRR) metric, which is found by using linear regression of the current test score on the past score in the same subject area. Proposed by Castellano (2012), the PRR is the percentile rank of the difference between a student's current (observed) score and the student's predicted score.

Table 4 summarizes the means, standard deviations (SDs), and sample sizes for students in 5th, 6th, and 7th grades in these student cohorts. The mean NSSs in these cohorts represent average achievement in the neighborhood of the 55th to 60th percentile nationally, and the SDs are representative of the variability in the national probability sample of the 2010–2011 norming of the *Iowa Assessments*.

The correlations across grades for the Mathematics and Reading assessments are provided in Table 5. These values are typical of correlations in matched cohorts on assessments that measure a well-defined general achievement construct. They are in the neighborhood of values obtained for test-retest reliability and provide strong support for the quantile and linear regressions needed to obtain SGPs and PRRs as indicators of growth.

Comparisons between the results from the Iowa Growth Model and the SGP and PRR approaches are provided in Tables 6 and 7 in terms of the correlations between growth indicators and the overall conceptual similarities. These correlations describe the consistency with which the Iowa Growth Model ranks student growth as compared to the SGP and PRR metrics. In both Mathematics and Reading, these results show that the Iowa Growth Model produces measures of student growth that are virtually identical to those of the other growth metrics.

Table 4: Means, Standard Deviations (SD) and Sample Sizes (N) for Comparative Study

Grade	Mathematics			Reading		
	Mean NSS	SD	N	Mean NSS	SD	N
5	222	24.6	23,452	225	28.6	23,511
6	232	28.3	27,024	231	32.0	27,046
7	250	30.6	24,024	245	34.2	27,046

Table 5: Correlations Across Grades in Mathematics and Reading

Grade	Mathematics			Reading		
	5th Grade	6th Grade	7th Grade	5th Grade	6th Grade	7th Grade
5	1.00			1.00		
6	.84	1.00		.79	1.00	
7	.81	.85	1.00	.77	.80	1.00

Table 6: Correlations between Iowa Growth Model, SGP and PRR		
	Iowa Growth Model	
	Mathematics	**Reading**
Student Growth Percentile (1 prior year)	.98	.97
Percentile Rank of Residuals	.99	.97

Table 7: Conceptual Comparison of SGPs to Iowa Growth Model		
	Iowa Growth Model	**SGP**
Starting Place	Based on Previous Performance	Based on Previous Performance
Growth Expectation	Dependent Upon Starting Place	Dependent Upon Group Performance
Growth Metric	Ranges from Negative Growth to Positive Growth Referenced to National Benchmark	Percentile Rank
Reference Group	Nationally Representative Group	Local Group

Comparison Example. The Iowa Growth Model uses the NSS vertical scale to determine (1) the expected NSS in a grade of students with the same NSS in the previous grade, and (2) the amount by which each student meets, exceeds, or fails to meet expected growth. The SGP metric simply ranks the performance of the students with the same NSS in Grade 1. Table 8 illustrates a comparison of the Iowa Growth Model results to SGP results for a group of 10 students. Each of these 10 students obtained a 200 NSS in Grade 4. With this as a starting place, each of these 10 students would have an expected growth of 215 in Grade 5. The observed scores from Grade 5 indicate that the students ranged from a low of 190 to a high of 226. Those achieving an observed score greater than 215 exceeded their expectation, while those achieving an observed score below 215 did not reach their expectation (a score of 215 met the expectation).

In the example in Table 8, the Iowa Growth scores are in the same rank order as the SGPs and the two metrics correlate perfectly in this sense. Such a result will occur whenever the SGPs are based on a locally defined cohort, such as all the students in a given grade in a school district. In fact, the only departure of the correlation coefficient from a perfect value of 1.00 would be caused by nonlinearity in the bivariate relationship between the two growth metrics. Note in this example, that the SGPs and the PRRs would be identical, so all Iowa-SGP comparisons would be the same as the Iowa-PRR comparisons.

Table 8: Comparing Iowa Growth Model Results to SGP Results

Student	Grade 4			Grade 5			
	Observed NSS	Expected		Observed		Growth Group	SGP
		NSS	Growth	NSS	Growth		
1	200	215	15	225	25	Exceeds	90
2	200	215	15	212	12	Does Not	50
3	200	215	15	215	15	Meets	60
4	200	215	15	205	5	Does Not	40
5	200	215	15	190	-10	Does Not	10
6	200	215	15	199	-1	Does Not	20
7	200	215	15	200	0	Does Not	30
8	200	215	15	226	26	Exceeds	99
9	200	215	15	218	18	Exceeds	80
10	200	215	15	216	16	Exceeds	70

Additional insight on where we consider students at different achievement levels in Grade 4 in the Iowa, SGP, and PRR metrics is provided in the following figures:

- a high-scoring group (NSS = 216, or about the 80th percentile nationally)
- a mid-scoring group (NSS = 200, or about the 60th percentile nationally)
- a low-scoring group (NSS = 180, or about the 30th percentile nationally)

Each figure that follows gives a variety of possible NSSs in Grade 5 (based on the time 1 NSS in Grade 4), the Iowa Growth score (observed time 2 minus observed time 1), the Value Added (Time 2 observed minus expected growth), and the SGP/PRR.

High-Scoring Group at Time 1

Observed Score Time 1	Observed Score Time 2	Growth	Value Added	SGP	PRR
216	242	+26	+9	99	99
	239	+23	+6	90	90
	238	+22	+5	80	80
	237	+21	+4	70	70
	236	+20	+3	60	60
	235	+19	+2	50	50
	234	+18	+1	40	40
	233	+17	0	30	30
	232	+16	-1	20	20
	231	+15	-2	10	10

Mid-Scoring Group at Time 1

Observed Score Time 1	Observed Score Time 2	Growth	Value Added	SGP	PRR
200	226	+26	+11	99	99
	225	+25	+10	90	90
	218	+18	+3	80	80
	216	+16	+1	70	70
	215	+15	0	60	60
	212	+12	-3	50	50
	205	+5	-10	40	40
	200	0	-15	30	30
	199	-1	-16	20	20
	190	-10	-25	10	10

Low-Scoring Group at Time 1

Observed Score Time 1	Observed Score Time 2	Growth	Value Added	SGP	PRR
180	193	+13	0	99	99
	192	+12	-1	90	90
	191	+11	-2	80	80
	190	+10	-3	70	70
	189	+9	-4	60	60
	188	+8	-5	50	50
	187	+7	-6	40	40
	186	+6	-7	30	30
	185	+5	-8	20	20
	182	+2	-11	10	10

As in the previous example (Table 8), students with the same NSS in Grade 4 received different NSSs in Grade 5. Although they were "assessment peers" in Grade 4, they were not after the Grade 5 results came in—they grew by different amounts between Grades 4 and 5. Student 3 in Table 8 achieved expected growth based on the Iowa Growth Model (Iowa observed-minus-expected score equals 0) and might well be judged to have met a district's growth standard. However, depending on the Grade 5 results of their assessment peers, this student might be judged to have achieved quite different results based on the SGP/PRR growth metrics. In the low-scoring group, meeting expected growth led to a growth percentile of 99, whereas in the mid- and high-scoring groups it led to growth percentiles of 60 and 30, respectively. This example shows the complexity of interpreting what on the surface appears as a simple metric,

SGP as a percentile rank, because a given SGP depends so much on the group from which it is derived. The Iowa growth metrics, in contrast, are referenced to a nationally representative group and provide empirically determined and scaled growth measures, which lead to straightforward comparisons of growth for students at different achievement levels.

APPENDIX A – SCALING OF THE IOWA ASSESSMENTS

The scaling of the *Iowa Assessments* allows for longitudinal score scales for measuring growth in achievement. Norming methods estimate national performance and long-term trends in achievement and provide a basis for measuring strengths and weaknesses of individuals and groups. Equating methods establish comparability of scores on equivalent test forms. Together, these techniques produce reliable scores that satisfy the demands of users and meet professional test standards.

Comparability of Developmental Scores Across Levels: The Growth Model

The foundation of any developmental scale of educational achievement is the definition of grade-to-grade overlap. Students vary considerably within any given grade in the types of cognitive tasks they can perform. For example, some students in 3rd grade can solve problems in mathematics that are difficult for the average student in 6th grade. Conversely, some students in 6th grade read no better than the average student in 3rd grade. There is even more overlap in the cognitive skills of students in adjacent grades—enough that some communities have devised multi-age or multi-grade classrooms to accommodate it. Grade-to-grade overlap in the distributions of cognitive skills is essential to any developmental scale that measures growth in achievement over time. Such overlap is sometimes described by the ratio of variability within grades to variability between grades. As this ratio increases, the amount of grade-to-grade overlap in achievement increases.

The problems of longitudinal comparability of tests and vertical scaling and equating of test scores have existed since the first use of achievement test batteries to measure educational progress. The equivalence of scores from various levels is of special concern in using tests "out-of-level" or in individualized testing applications. For example, a standard score of 185 earned on Level 10 should be comparable to the 185 earned on any other level; a grade equivalent score of 4.8 earned on Level 10 should be comparable to a grade equivalent of 4.8 earned on another level.

Each test in the *Iowa Assessments* is a single continuous test representing a range of educational development. A common developmental scale was needed to relate the scores from each level to the other levels. The scaling requirement consisted of establishing the overlap among the raw score scales for the levels and relating the raw score scales to a common developmental scale. The scaling test method used to build the developmental scale for the *Iowa Assessments*, Hieronymus scaling, is described in Petersen, Kolen & Hoover (1989).

The National Standard Score (NSS)

Students participated in special test administrations for scaling the *Iowa Assessments*. The scaling tests were wide-range achievement tests designed to represent each content domain in the Complete Battery. Scaling tests were developed for three groups: Kindergarten through Grade 3, Grades 3 through 9, and Grades 8 through 12. These tests were designed to establish links among the three sets of tests from the data collected. During the standardization, scaling tests in each content area were spiraled within classrooms to obtain nationally representative and comparable data for each subtest.

The scaling tests provide essential information about achievement differences and similarities between groups of students in successive grades. For example, the scores show the variability among 4th graders in science achievement and the proportion of 4th graders that score higher in science than the typical 5th grader. The study of such relations is essential to building developmental score scales. These score scales monitor year-to-year growth and estimate students' developmental levels in areas such as Reading, Language, and Math. To describe the developmental continuum in one subject area, students in several different grades must answer the same questions.

The score distributions on the scaling tests defined the grade-to-grade overlap needed to establish the common developmental achievement scale in each test area. An estimated distribution of true scores was obtained for every content area using the appropriate adjustment for unreliability (Feldt & Brennan, 1989). The percentage of students in a given

grade who scored higher than the median of other grades on that scaling test was determined from the estimated distribution of true scores. This procedure provided estimates of the ratios of within-grade to between-grade variability free of chance errors of measurement and defined the amount of grade-to-grade overlap in each achievement domain.

The relationship of standard scores to percentile ranks for each grade was obtained from the results of the scaling test. Given the percentages of students in the national standardization in one grade above or one grade below the medians of other grades, within-grade percentiles on the developmental scale were determined. These percentiles were plotted and smoothed. This produced a cumulative distribution of standard scores for each test and grade, which represents the growth model for that test. The relationship between raw scores and standard scores were obtained from the percentile ranks on each scale.

The amount of grade-to-grade overlap in the developmental standard score scale tends to increase steadily from Kindergarten to 8th grade. This pattern is consistent with a model for growth in achievement in which median growth decreases across grades at the same time as variability in performance increases within grades.

Units for the description of growth from grade to grade must be defined so that comparability can be achieved between descriptions of growth in different content areas. To define these units, achievement data were examined from several sources in which the focus of measurement was on growth in key curriculum areas at a national level. The data included results of scaling studies using not only the Hieronymus method, but also Thurstone and item-response theory methods (Mittman, 1958; Loyd & Hoover, 1980; Harris & Hoover, 1987; Becker & Forsyth, 1992; Andrews, 1995). Although the properties of developmental scales vary with the methods used to create them, all data sources showed that growth in achievement is rapid in the early stages of development and more gradual in the later stages. Theories of cognitive development also support these general findings (Snow & Lohman, 1989). The growth model for the current edition of the *Iowa Assessments* was determined so that it was consistent

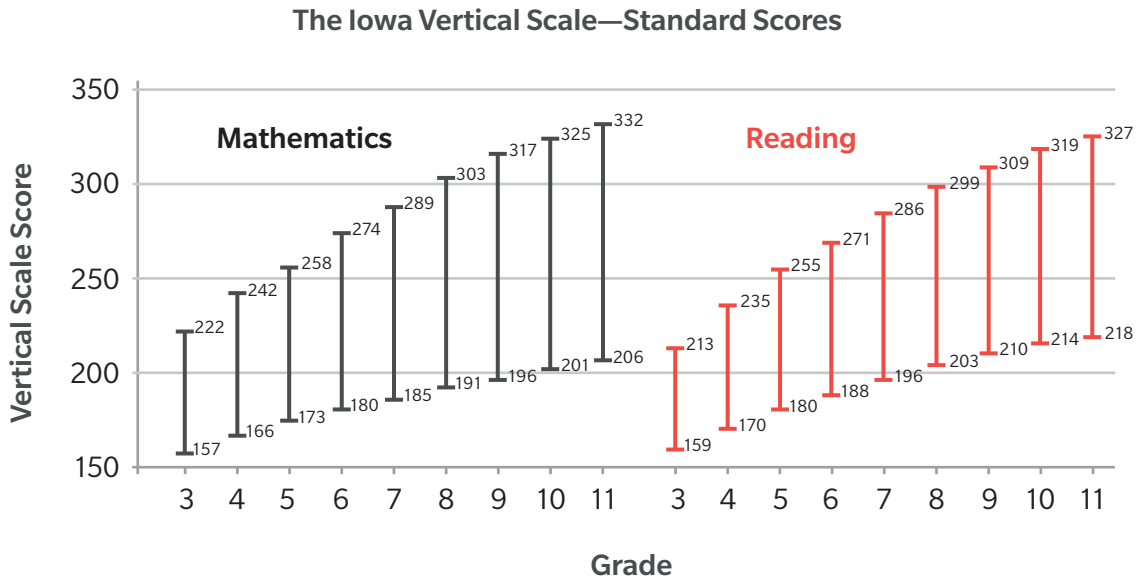
with the patterns of growth over the history of tests and with the experience of educators in measuring student growth and development.

The purpose of a developmental scale in achievement testing is to permit score comparisons between different levels of a test. Such comparisons are dependable under standard conditions of test administration. In some situations, however, developmental scores (standard scores and grade equivalents) obtained across levels may not seem comparable. Equivalence of scores across levels in the scaling study was obtained under optimal conditions of motivation. Differences in attitude and motivation, however, may affect comparisons of results from "on-level" and "out-of-level" testing of students who differ markedly in developmental level. If students take their tests seriously, scores from different levels will be similar (except for errors of measurement). If students are frustrated or unmotivated because a test is too difficult, they will probably obtain scores in the "chance" range. But if students are challenged and motivated, their achievement will be measured more accurately.

Scaling tests were used to produce a single common score scale for the *Iowa Assessments*. Two sets of data were used to produce its standard scores and grade-equivalents—standardization data and scaling-study data. Within-grade scaling is obtained from the distribution of scores in the standardization sample. Between-grade scaling is obtained from scores earned on a scaling study. For the scaling study, three scaling tests were developed and administered—one for Grades K–3, one for Grades 3–9, and another for Grades 8–12. Each test drew items from the *Iowa Assessments* test levels that were relevant to the particular grades, then each test was administered to students in all of the relevant grades. For example, the Grades 3–9 scaling test drew items from all of the test levels for those seven grades and was administered to students in all of those seven grades. Thus, Grade 3 students did, in fact, take Grade 9 items, and vice versa.

In addition to yielding direct and empirical between-grade score relationships, the scaling-test model also yields growth, or learning, curves that support what cognitive psychologists say about growth: that students in lower grades grow more than those in upper grades, and students in upper performance levels grow more than those in lower performance levels. Growth is that gain made by students between grades, or from grade to grade, which is beyond the maintenance of their within-grade status or performance levels. Figure 4 illustrates these learning curves.

Figure 4: Iowa Growth Model Attributes



APPENDIX B – REPORTING GROWTH

There is tremendous interest in measuring growth and using this information as part of making changes in education (Betebenner & Linn, 2010). With the Iowa Growth Model, teachers can set a goal and then use the data to measure progress towards that goal. With the Iowa Growth Model, administrators can summarize the growth performance of groups of students to help evaluate a program or establish growth goals for particular cohorts. With the Iowa Growth Model, policymakers can also monitor progress toward goals over time.

Using the Iowa Growth Model, the following section offers some suggestions for displaying the results for an individual student as well as aggregated results for schools or districts. These displays offer a clear way of seeing aggregate changes in student performance.

Individual Student Report. For teachers, individual growth goals, conditional upon their previous performance, can be established for a student or a classroom of students. Teachers can then monitor the progress of that student towards that goal over time. The figure on the next page provides an individual student report for Matthew Anderson. For each test that Matthew completed, the blue line represents expected growth based on his starting place in 4th grade. The black line represents the observed scores for Matthew between 4th grade and 8th grade. The gray line represents the expected growth for a student that started at the 50th NPR in 4th grade.

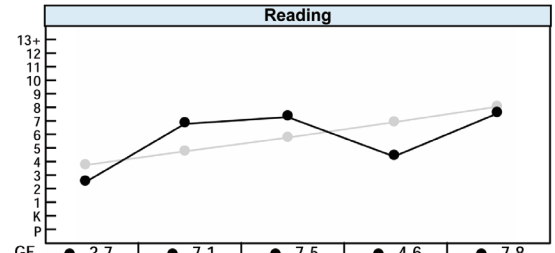
Student Profile For Matthew Andersen

Student ID: 9470123394
Iowa Assessments™

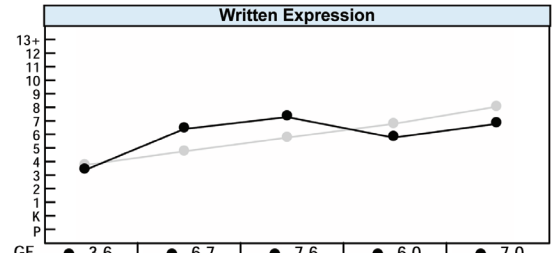
Test Date: 12/12/2012
Norms: Fall 2011

Grade: 8 Level: 14
Class: Stevenson
Building: Longitudinal Bldg

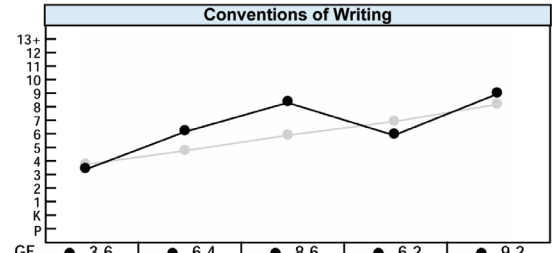
System: Training System
State: DM



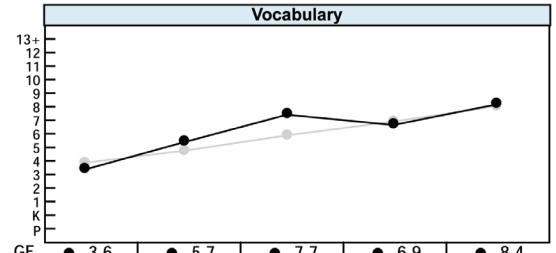
	Grade 4 Lv.10 E 08/08/2008	Grade 5 Lv.11 E 09/09/2009	Grade 6 Lv.12 E 10/10/2010	Grade 7 Lv.13 E 11/11/2011	Grade 8 Lv.14 E 12/12/2012
GE	2.7	7.1	7.5	4.6	7.8
NPR	16	78	68	21	46



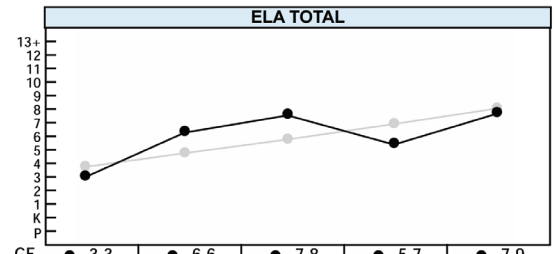
	Grade 4 Lv.10 E 08/08/2008	Grade 5 Lv.11 E 09/09/2009	Grade 6 Lv.12 E 10/10/2010	Grade 7 Lv.13 E 11/11/2011	Grade 8 Lv.14 E 12/12/2012
GE	3.6	6.7	7.6	6.0	7.0
NPR	36	70	65	38	41



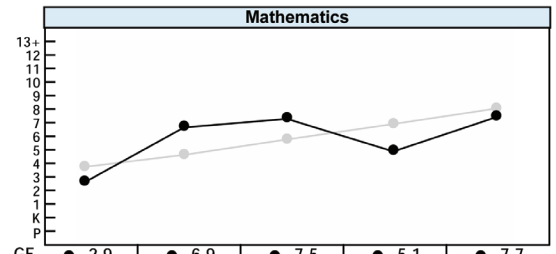
	Grade 4 Lv.10 E 08/08/2008	Grade 5 Lv.11 E 09/09/2009	Grade 6 Lv.12 E 10/10/2010	Grade 7 Lv.13 E 11/11/2011	Grade 8 Lv.14 E 12/12/2012
GE	3.6	6.4	8.6	6.2	9.2
NPR	37	68	75	37	58



	Grade 4 Lv.10 E 08/08/2008	Grade 5 Lv.11 E 09/09/2009	Grade 6 Lv.12 E 10/10/2010	Grade 7 Lv.13 E 11/11/2011	Grade 8 Lv.14 E 12/12/2012
GE	3.6	5.7	7.7	6.9	8.4
NPR	32	61	74	45	54



	Grade 4 Lv.10 E 08/08/2008	Grade 5 Lv.11 E 09/09/2009	Grade 6 Lv.12 E 10/10/2010	Grade 7 Lv.13 E 11/11/2011	Grade 8 Lv.14 E 12/12/2012
GE	3.3	6.6	7.8	5.7	7.9
NPR	29	77	74	30	46



	Grade 4 Lv.10 E 08/08/2008	Grade 5 Lv.11 E 09/09/2009	Grade 6 Lv.12 E 10/10/2010	Grade 7 Lv.13 E 11/11/2011	Grade 8 Lv.14 E 12/12/2012
GE	2.9	6.9	7.5	5.1	7.7
NPR	18	79	69	22	45

NPR = National Percentile Rank GE = Grade Equivalent

Report Date 11/13/2013

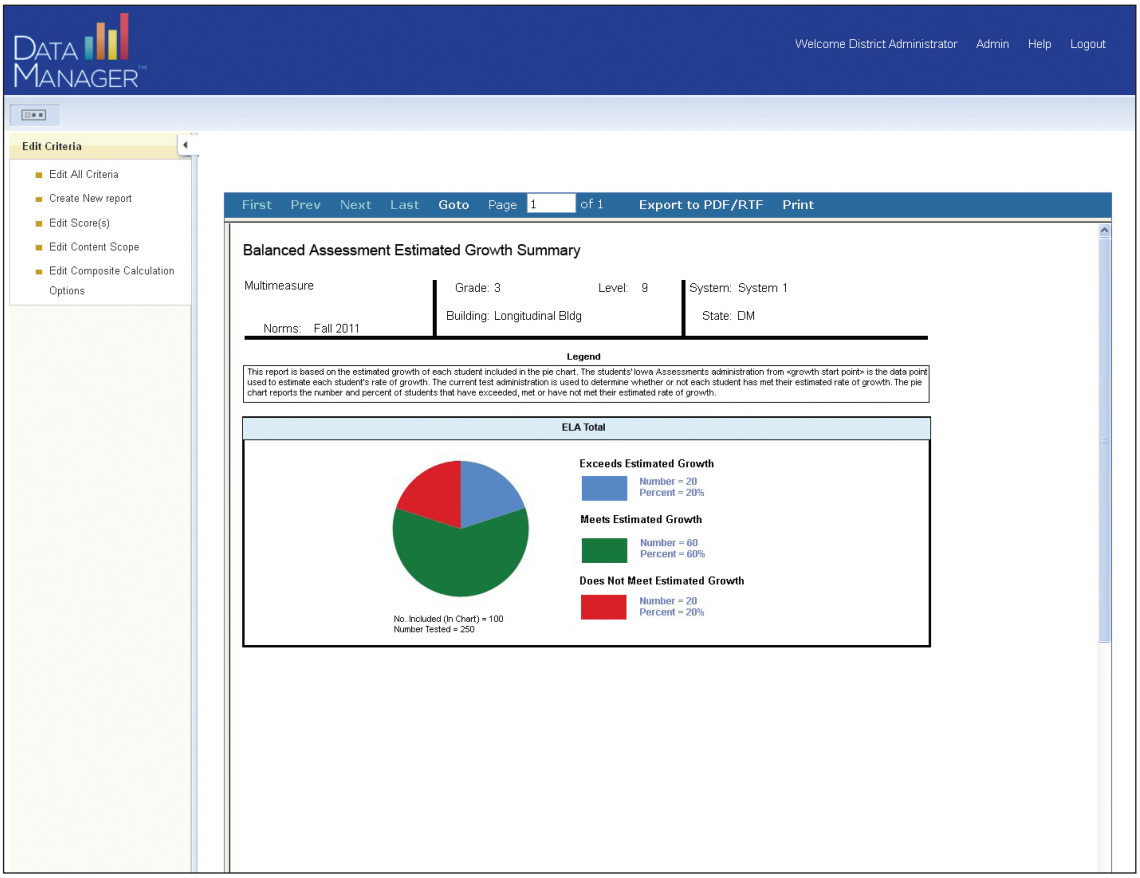
Copyright 2000,2013 The Riverside Publishing Company. All Rights Reserved.

Group Reports. The evaluation component of the Iowa Growth Model allows districts to compare the expected growth to the observed growth—enabling a determination of the value added through general instruction or perhaps the degree of response to an intervention. Figures 5 and 6 show the results of aggregating student-level data to school-level or district-level reports.

Two different displays of summarizing growth scores across years within a district are presented in Figures 7 and 8. In Figure 7, students with scores of zero and above have achieved or exceeded expected growth based on the Iowa Growth Model. In a district or group that has instituted improvements in professional development for teachers and increased the rigor of instruction, one would anticipate a histogram like the one shown in Figure 7 to shift to the right with more students exceeding expected growth. In a district or group "resting on its laurels," this histogram might shift to the left, indicating students are falling behind their achievement peers between assessment occasions. This type of plot also can illustrate value-added growth scores for students in specific educational programs.

Figure 8 is a plot of average growth for students between Grade 4 and Grade 5 in 37 elementary school buildings in a large school district in the Midwest. The plot ranks buildings from high to low in terms of the Iowa value-added metric (observed-minus-expected growth). In addition to the mean value-added for each building, the sizes of the buildings are indicated. In an actual district report to the school board, for example, other identifiers could be included for the buildings, enabling districts to better evaluate the performance of a group of students. Figure 8 allows a district to evaluate average growth across classrooms, teachers, or buildings.

Figure 5



The Estimated Growth Summary report is a highly visual report that can be used by teachers for a quick 'at a glance' understanding of a group's achievement based upon two administrations of the *Iowa Assessments*. The report includes the number and percent of students in the group that meet, exceed, or do not meet their estimated growth.

Figure 6

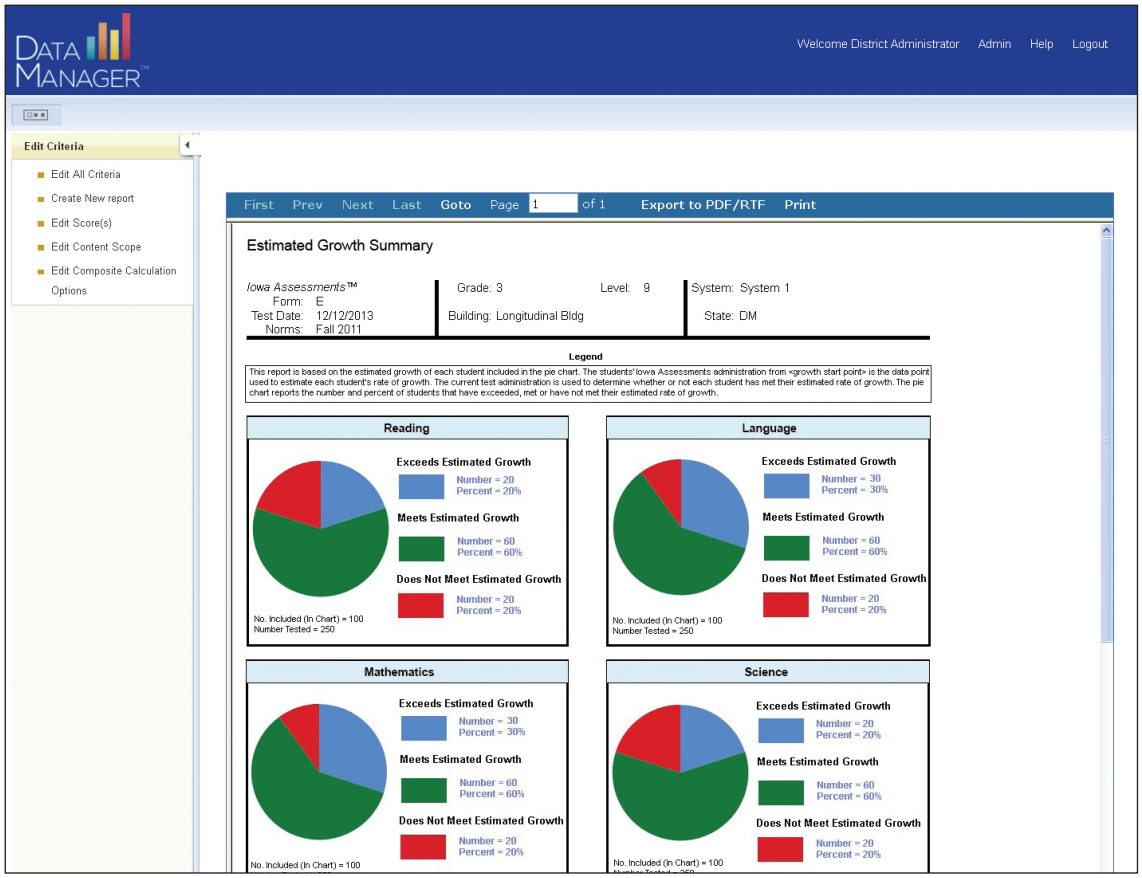


Figure 7

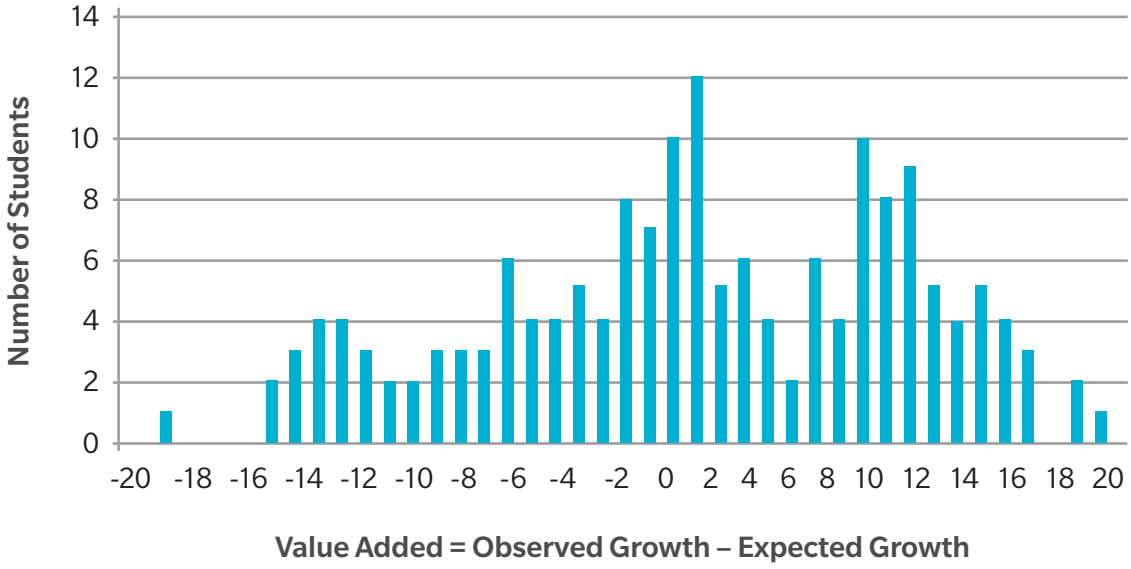


Figure 8

Average Growth = Average Observed Growth – Average Expected Growth

REFERENCES

Andrews, K. M. (1995). The effects of scaling design and scaling method on the primary score scale associated with a multi-level achievement test. Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA. [3, 4]

Becker, D. F. & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement,* 29, 341–354. [3]

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4): 42–51.

Betebenner, D.A., & Linn, R. L. (2010). *Growth in student achievement: Issues of measurement, longitudinal data analysis and accountability.* Retrieved from http://www.k12center.org/publications.html

Betebenner, D.W., (2010). *New Directions for Student Growth Models.* Dover, NH: National Center for the Improvement of Educational Assessment. Presentation dated December 13, 2010. Retrieved March 29, 2012 from http://www.ksde.org/LinkClick.aspx?fileticket=UssiNoSZks8%3D&tabid=4421&mid=10564

Castellano, K.E., & Ho, A.D., (in press). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics.*

Chall, J.S. (1996). *Stages of reading development.* New York: Harcourt Brace.

Cunningham, P., Welch, C., & Dunbar, S., (2013). *Value-Added Analysis of Teacher Effectiveness.* University of Iowa.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan. Gulliksen, H. (1950).

Furgol, K., Fina, A., & Welch, C. (2011). Establishing *validity evidence to assess college readiness through a vertical scale.* Paper presented at the 2011 Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Harris, D. J. & Hoover, H. D. (1987, June). An application of the three-parameter IRT model to vertical equating. Applied Psychological Measurement, 2, 151–159. [2]

Hoover, H.D., Dunbar, S.B., Frisbie, D.A., Oberley, K.R., Bray, G.B., Naylor, R.J., Lewis J.C., Ordman, V.L., & Qualls, A.L. (2007). *Iowa tests of basic skills: Norms and score conversions.* Itasca, IL: Riverside Publishing.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement*, 17, 179–193. [2, 3].

Mittman, A. An empirical study of methods of scaling achievement tests at the elementary grade level. Unpublished doctoral dissertation, University of Iowa, 1958.

Patz, R. J. 2007. *Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems.* Prepared for the Technical Issues in Large-Scale Assessment (TILSA) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO).

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Reardon, S.F., & Raudenbush, S.W. (2009). Assumptions of value-added models for estimating school effects. *Educational Finance and Policy,* 4(4): 492–519.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan.

Welch, C. J., & Dunbar, S. B. (2011, April). *K-12 assessments and college readiness: necessary validity evidence for educators, teachers, and parents.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

MEASURING GROWTH WITH THE *IOWA ASSESSMENTS*

A Black and Gold Paper

Leaders. Scholars. Innovators. COLLEGE OF EDUCATION, UNIVERSITY OF IOWA

Contact your **Houghton Mifflin Harcourt - Riverside** Account Executive or call Customer Service at **800.323.9540** for more information on the *Iowa Assessments*.

For more information on Iowa Testing Programs, please visit https://itp.education.uiowa.edu/

@HMHeducation Houghton Mifflin Harcourt

ACT® is a registered trademark of ACT, Inc. Houghton Mifflin Harcourt has no affiliation with ACT, Inc. and our products are not approved or endorsed by ACT, Inc. Iowa Assessments™ is a trademark of Houghton Mifflin Harcourt Publishing Company.
© Houghton Mifflin Harcourt Publishing Company. All rights reserved. Printed in the U.S.A. 02/14 MS90550

Houghton Mifflin Harcourt