Final Report

Growth Model Working Group

Submitted to:

Christopher Koch, State Superintendent

April 14, 2011
The Growth Model Working Group (GMWG) was appointed for a 12-month period beginning May 2010. Membership included representatives of key professional organizations, ICCB, IBHE, ISBE, local superintendents, principals, central office administrators, and technical advisors (see Attachment A for a complete listing of members). The GMWG was asked to address the following charge (see Attachment B, original charge letter):

1. Identify and coordinate the various uses and timelines for assessing and using student growth information including, but not limited to, Public Act 096-0861, the State Fiscal Stabilization Fund Phase II, the Race to the Top (RTTT) application, and the Race to the Top Assessment (RTTT-A) consortia;

2. Develop a rigorous process for evaluating different approaches to assessing student growth utilizing peer benchmarking, empirical analysis, stakeholder input, and expert review;

3. Advise the Superintendent and the State Board on the selection and use of growth information for instructional improvement; evaluation of teachers, principals, and school effectiveness; resource allocation; school and district accountability; policy making; and

4. Evaluate progress and effectiveness of the development and use of growth information at state, regional, and local levels.

The GMWG has met five times thus far and engaged in the following activities (see Attachment C for meeting notes):

June 16, 2010—Introductions; review of charge; overview of uses and timelines for assessing and using student growth, including ESEA reauthorization, RTTT, RTTT-Assessment, PA 96-0861, State Fiscal Stabilization Fund; establish meeting schedule.

July 26, 2010—Exploration of the Tennessee Value-Added Assessment System, the Colorado Growth Model, and the CPS/WCER Value Added Model.

August 30, 2010—Update on PEAC activity and connection to GMWG; overview of Illinois Race to the Top application, Illinois Longitudinal Data System and SIS with implications for GMWG; reports on local uses of student growth information for decision-making, instructional improvement, and accountability.

October 22, 2010—Revision of charge to address only the use of student growth information for accountability purposes; update on Race to the Top; overview of the PARCC Consortium; discussion of pros and cons of different approaches to growth models.

April 7, 2011—Review and approval of final report; update on PARCC Consortium and PEAC activity; report on preliminary analysis of three candidate growth models using ISAT data.

The GMWG established a website (www.isbe.net/gmwg) which contains meeting notes, agendas, membership, presentations, and background reading for the group including:

- Case Studies of Data Use: A Series of Reports Developed by the CCSSO Accountability Systems Reporting (ASR) Collaborative (Gallagher, WestEd [pdf])
- Growth, Standards, and Accountability (Betebenner, Center for Assessment) [pdf])
Growth Model Working Group

- Implementer’s Guide to Growth Models (Auty, ASR Working Group, CCSSO [pdf])
- Policymakers’ guide to Growth Models for School Accountability: How do Accountability Models Differ? (Goldschmidt et al, ASR Working Group [pdf])
- Progress on Development and Reporting Measures of Growth (Blank, CCSSO [pdf])
- Hierarchical Modeling Approach (Meyer & Dokumaci, 2010)
- Coming Together to Raise Achievement (ETS, 2011)

We have revised our charge to reflect the Superintendent’s request to focus exclusively on the use of student growth information for accountability purposes and outcome of the Race to the Top competition. The revised charge is as follows:

1. Identify and coordinate the timelines for assessing and using student growth information for accountability purposes including, but not limited to, ESEA reauthorization, the State Fiscal Stabilization Fund Phase II, and the PARCC consortia;
2. Develop a rigorous process for evaluating different approaches to assessing student growth for accountability purposes, including peer benchmarking, empirical analysis, stakeholder input, and expert review;
3. Advise the Superintendent and the State Board on the selection and use of growth information for school and district accountability; and
4. Evaluate progress and effectiveness of the development and use of growth information for accountability purposes at state, regional, and local levels.

Progress to date, conclusions, recommendations, and next steps will be discussed in relation to each element of the charge.

1. Identify and coordinate the timelines for assessing and using student growth information for accountability purposes, including, but not limited to, ESEA reauthorization, the State Fiscal Stabilization Fund (SFSF) Phase II, and the PARCC consortia.

With an uncertain timeline for ESEA Reauthorization and RTTT funds not forthcoming, the most pressing timeline for assessing and using student growth models for accountability purposes is dictated by the state’s SFSF Phase II application. As indicated in Part 3A, Indicator (b)(2), by October 31, 2011, ISBE must have in place a system for reporting student growth data in reading/language arts and mathematics to schools and LEAs for accountability purposes. Development of the student growth model for accountability runs concurrently with the development of a teacher and principal performance evaluation process in which ISBE must provide “student growth data on their current students and the students they taught in the previous year to, at a minimum, teachers of reading/language arts and mathematics in grades in which the State administers assessments in those subjects in a manner that is timely and informs instructional programs.” Implementation of the teacher and principal performance evaluation process begins September 1, 2012
In addition to SFSF Phase II, the State is engaged in other activities that influence its capacity to meet the deadlines for reporting student growth data for accountability and teacher impact. Illinois has been fortunate to receive two federal grants, the SLDA Cooperative Agreement and the SLDA Expansion Grant (ARRA) to create a P–20 Illinois Longitudinal Data System (ILDS) to support the assessment and reporting purposes. The new system will build upon and greatly enhance the current Student Information System (SIS) by assigning unique student, teacher, and administrator identifiers that enable linking of student data with teacher, administrator, and school data over time and across educational systems from early childhood to postsecondary. Since 2006, ISBE has assigned students a unique identifier (SID) which enables tracking of students and their ISAT, IMAGE, PSAE, IAA, and ACCESS test results over time and across all K-12 public schools in Illinois. In 2009, students in public early childhood programs were added to SIS. ISBE will have the capacity to match teachers to students and courses and teachers-toeacher preparation programs by September 30, 2011. The ability to link teachers, students and courses over time is an essential element in most growth models. Federal funding for the development of the ILDS runs through the fourth quarter of 2013, at which time the data system and support infrastructure will be fully functional and integrated into state and local reporting and accountability systems. If well designed and implemented, the ILDS could greatly enhance the efficiency and utility of student growth models for accountability purposes.

In 2010, ISBE voted to adopt the Common Core State Standards and to join the PARCC assessment consortium as a governing member. Each of these events has implications for design and implementation of a student growth model for accountability since the standards and the ways in which they are assessed have changed or are changing, thus breaking the trend and significantly compromising the state’s ability to reliably gauge changes in student performance over time. In discussing accountability implications of the shift from Illinois Learning Standards (ILS) to the Common Core, GMWG members immediately raised the questions, “On what standards will schools be held accountable for student growth?”; “To what extent are current state assessments (ISAT, ACCESS, IAA, PSAE) valid and reliable measures of student growth on the Common Core State Standards?”; and “What are the strengths and limitations of growth models?” Credibility and validity of any accountability system based upon student growth will depend on ISBE developing psychometrically sound and rational answers to those questions and effectively disseminating them to families, teachers, and administrators throughout the state.

As a member of PARCC, Illinois is on track to adopt new state assessments aligned with the Common Core State Standards in mathematics and literacy/ELA in grades 3–high school. Piloting and field testing will take place in 2011–2012, 2012–2013, and 2013–2014. As a governing board member, Illinois will participate in the pilot and field tests. Full statewide administration will be in the 2014–2015 school year, with 2015–2016 being the first year in which student growth can be calculated. The ability to assess student growth over time for instructional improvement, assessment of teacher impact, and accountability is an underlying premise of the PARCC proposal. Once again, if well designed and implemented, the PARCC assessments could greatly advance the use of student growth models for accountability purposes. As a governing board member, Illinois is in an excellent position to influence their development.


Table 1. Key Dates Affecting Growth Model Implementation

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 30, 2011</td>
<td>State Assessment data available through IDLS linking teachers with courses, students and teacher preparation programs.</td>
</tr>
<tr>
<td>October 31, 2011</td>
<td>SFSF requires that ISBE must have in place a system for reporting student growth data in reading/language arts and mathematics to schools and LEAs for accountability</td>
</tr>
<tr>
<td>September 1, 2012</td>
<td>Implementation of the teacher and principal performance evaluation process begins</td>
</tr>
<tr>
<td>Spring 2012</td>
<td>Piloting of PARCC assessment begins</td>
</tr>
<tr>
<td>2012–2013</td>
<td>Piloting and field testing of PARCC assessment</td>
</tr>
<tr>
<td>December 31, 2013</td>
<td>IDLS fully implemented</td>
</tr>
<tr>
<td>2013–2014</td>
<td>Field testing of PARCC assessment</td>
</tr>
<tr>
<td>2014–2015</td>
<td>Full statewide administration of PARCC</td>
</tr>
<tr>
<td>2015–2016</td>
<td>Second full statewide administration of PARCC; Growth model based on PARCC data begins</td>
</tr>
</tbody>
</table>

Conclusions

The timeline for implementing a student growth model for accountability purposes is complicated by recent adoption of Common Core Standards, uncertainty about ESEA reauthorization, an immature longitudinal data system, and a new state assessment system on the horizon. While some argument might be made to delay adoption until the new data and assessment systems are in place, SFSF requirements necessitate immediate consideration of student growth for state accountability purposes.

Recommendations

1.1 ISBE should design and begin to implement a student growth model for state, school, and district accountability purposes based upon ISAT, IAA, and PSAE; aligned with Common Core State Standards; and dependent upon the current capabilities of SIS and the IDLS for implementation by the SFSF deadline of October 31, 2011. Use of this model will continue until a new growth model based upon the PARCC assessment can be implemented as early as 2016 (or when at least two years of PARCC assessment data are available). Some members of the committee have expressed concerns that changing the growth model in five years may be disruptive for LEAs and the state. At this point, we do not know how possible it will be to link PARCC assessment results with prior state assessments. All attempts should be made to equate the two systems in order to maintain trends and preserve continuity in the growth models.
1.2 ISBE should develop strong connections between ILDS development and state needs for student growth models for accountability by appointing a member of the GMWG or a TAC member knowledgeable about growth models to serve on the State Education Data Advisory Committee (DAC).

1.3 ISBE should commission an independent content validity/alignment study between the current state assessments (ISAT, IAA, and PSAE) and the Common Core State Standards. Based on this study, ISBE should develop a plan for how the state assessments will be modified or bridged to represent the Common Core. ISBE should aggressively disseminate this plan to teachers, administrators, and families via the communication plan discussed in 1.5 below.

1.4 ISBE, as a governing member of PARCC, should actively influence the development of the student growth model components of the assessment system in ways that align with Illinois accountability needs. ISBE should report regularly on PARCC progress at TAC meetings and discuss implications for growth model development and the need for state and local supports for implementation. ISBE should ensure that growth models are explored during field testing and piloting of the PARCC assessments so that a student growth model based upon PARCC assessments can be implemented for accountability purposes in 2016.

1.5 ISBE should develop and implement a communication plan for informing parents, teachers, principals, superintendents, legislators, and other key stakeholders about the process and impact of implementing a growth model for state, school, and district accountability, including the use of growth data to supplement AYP reporting.

2. Develop a rigorous process for evaluating different approaches to assessing student growth for accountability purposes, including peer benchmarking, empirical analysis, stakeholder input, and expert review.

Since 2005, the USED has been supporting the incorporation of a student growth component into state accountability systems. All growth models approved for federal accountability purposes must meet seven core principles set by the USED:

1. Ensure 100% student proficiency by 2014, and closing of the achievement gap for all student groups;
2. Establish high expectations for low-achievement students without setting annual achievement expectations based upon student background or school characteristics;
3. Produce separate accountability decisions for reading and mathematics;
4. Include all students, schools, and districts and hold schools and districts accountable for the performance of student subgroups;
5. Have been operational for more than one year and have received approval through the NCLB Peer Review Process;
6. Track student progress and include annual assessments in grades 3–8 and high school in reading and mathematics; produce comparable results from grade to grade and year to year;
7. Include student participation rates and student achievement on an additional academic indicator.
Growth Model Working Group

In the Federal Register announcement for Race to the Top funds, the USED defines “student growth” quite specifically:

Student growth means the change in achievement data for an individual student between two points in time. Growth may be measured by a variety of approaches, but any approach used must be based on student achievement (as defined in this notice) data, and may also include other measures of student learning in order to increase the construct validity and generalizability of the information. (pp. 37811–37812)

Given this context, the GMWG reviewed the results from the USED Growth Model Pilot Program, read a number of scholarly critiques of growth models, and examined several existing student growth models spanning the three types most commonly used for accountability purposes (Gallagher, 2009):

- **Growth to proficiency, growth to a standard or trajectory models** are intended to show if a student is “on track” to reach a proficiency target at some future point in time (usually a future grade level), the gain required to reach the proficiency target, and the annual increments necessary to make that gain. Students are considered “on track” if their score gains match or exceed the annual increments required to reach proficiency in the future grade. For AYP, students are counted as “proficient” if their gains match or exceed the annual increment. Colorado, Alaska, Arizona, Arkansas, Florida, Missouri, and North Carolina use this approach. The GMWG held a video-conference with Colorado State Department of Education to learn more about their model.

- **Value table or transition models** evaluate growth by tracking student transitions across performance levels (or sublevels) with the goal of moving students to higher performance levels. Because growth is associated with change in performance level, transition models do not require a vertical scale (CCSSO, 2007). What is required is a sophisticated understanding of what is required to change performance levels across grades. Delaware, Iowa, and Michigan use value tables.

- **Projection models**, including value-added models, predict future performance (generally within 3–4 years) based upon past performance and the performance of a normative sample of peers (prior cohorts) in target grades and then compare predicted performance to the proficiency standard for the target grade. Students are counted as “proficient” if they are predicted to reach proficiency in the future. In value-added models (VAM), students’ past performance is used to estimate a projected score for that student. Attained scores are compared to projected scores. A student whose actual score exceeds the projected score has demonstrated “growth.” Students whose scores meet or exceed the score needed to reach proficiency by the target date are counted as proficient. Pennsylvania uses projection models in their accountability system. Tennessee and Chicago Public Schools employ value-added models for school and district accountability. The GMWG heard presentations on the Tennessee Value-Added Assessment System (TVAAS) and the CPS Systems.

The GMWG evaluated these models from two perspectives:

---

1 Academically challenging, technically sound annual standards-based assessments.
1. Appropriateness for the Illinois growth model; to begin implementation by **October 31, 2011**, based upon current state assessments; and

2. Implications for a future accountability system, based upon PARCC assessments, to be operational in 2016.

**Conclusions**

There is no “best” student growth model for accountability purposes. Each approach has strengths and limitations and is only as good as the assessment, standards, and longitudinal data systems upon which it is based. Incorporating a student growth model will place additional demands on these systems, and any approach selected must take into consideration existing strengths and limitations, such as:

- Lack of a vertical scale for IAA and PSAE which mitigates against use of mean-based or gain score models such as those used in Tennessee and Colorado;
- Weak alignment between the ISAT (grades 3–8) and the PSAE (high school) in terms of “percent meets and exceeds” and underlying standards which may limit the accuracy of projection models based on these assessments;
- Strong relationship between the PSAE and college readiness benchmarks;
- Ability to track student demographics and state assessment results across schools over time since AY 2006. No current link between students and teachers, although full ILDS implementation in 2013 will greatly enhance the ability to track students, archive data over time, and maintain and store large data sets securely.

Different growth models may produce different results. The selection of a student growth component should reflect the specific purpose(s) for which the growth information can be used, such as rewarding schools who produce the most growth, targeting resources to schools or students who fail to reach growth targets, identifying interventions to enhance growth for specific grade levels or subgroups, and providing information on student learning beyond AYP. A clear understanding of the intended use of the model will guide the tradeoffs and key decisions made about the model, such as what type of growth will be measured, how much growth will be sufficient, and if students with different starting points will be expected to grow at the same rate (Gong, Perie, and Dunn, 2006).

By focusing on changes in performance of individual students over time, growth models can provide more detailed information than status models, such as AYP, regarding the extent to which specific groups of students are making progress toward achievement targets, changes in student achievement over time, cumulative effects of schooling, and program effectiveness. These benefits warrant the effort required to design and implement a credible, technically adequate student growth model based on our current assessment, standards and data systems in the short term and to plan for continued use of student growth in future accountability systems.

Based upon peer benchmarking, expert review, and stakeholder input, and given the strengths and limitations of Illinois current standards, assessment, and data systems, the following approaches to assessing student growth for state, district, and school accountability purposes have emerged as viable and worthy of empirical investigation:
Growth Model Working Group

Value tables or transition models (see Delaware and Michigan descriptions, Attachment D) which answer the question “Are students (and subgroups) making adequate progress across performance levels?” This approach does not require a vertical scale and permits inclusion of students who take alternate assessments. It allows for non-linear growth across grade levels. Performance level descriptors and cut scores will likely be revised with the adoption of the Common Core State Standards. Revision could be done with attention to cross-grade alignment and weighting and sublevels within a performance level.

A simple Post-on-Pre (using adjacent grades), two-period, value added model (Meyer & Dokumaci, 2010—Attachment E) that makes use of two years of consecutive grade longitudinal assessment data and answers the question, “To what extent do schools contribute to student growth, controlling for other sources such as prior achievement, and student and family characteristics?” Value added systems facilitate comparisons of student growth across schools serving different student populations. While typically based upon residual gain scores in a particular content area, VAMs could be expanded to include prior achievement in other areas and other student measures, such as attendance and demographics. While there may be some advantage of expanding the model to include three or more periods, an initial investigation of the feasibility and utility of a two-period model would represent a significant step forward in providing growth information for accountability purposes. Though some multi-level VAMs are dependent upon vertical scales, this simple model for use at the school level could work within the parameters of the current ISAT in grades 3–8. CPS has been using VAM with ISAT at the grade and school level for the past few years in all elementary schools controlling for selected student factors.

Student Growth Percentiles or other normative growth analysis techniques (Betebenner, 2009—Attachment F) which answer the question, “How much did students progress in relation to their academic peers (i.e., students with identical prior achievement)?” Because past scores are used solely for conditioning purposes, use of student growth percentiles does not require a vertical scale. They differ from the previous approaches by focusing on a student’s growth relative to other students rather than the magnitude of growth itself. Normative models, such as student growth percentiles, are helpful for determining “reasonable” growth as compared to the very high rates of growth in trajectory or projection models based on NCLB targets of 100% proficient by 2014.

Much of the discussion of the GMWG thus far has focused on the use of growth models based upon ISAT data. We have not spent as much time grappling with how to incorporate other assessments such as ACCESS, IAA, and the various components of PSAE, but given the limited data on alignment across these various measures and the gap between 8th grade ISAT and PSAE, considerable challenges exist. One promising option is that many schools and districts across the state are using EPAS. In those districts, there will be annual assessment data in grades 3 through 11 which could serve as the basis for an empirical investigation of potential growth models which span elementary and secondary levels. WCER has examined the use of EPAS for growth modeling in CPS.

Recommendations

ISBE should explicitly state and aggressively communicate a rationale for incorporating the selected growth model component into the state accountability system, intended short- and long-term uses of growth data for accountability, and limitations of the growth model.
2.1 ISBE, in consultation with the state assessment TAC, other experts, and contractors, should conduct empirical analyses to examine the comparability, feasibility, accuracy, and propriety of the three candidate approaches (i.e., Values tables, Post-on-Pre VAM, and Student Growth Percentiles) to assessing student growth for accountability purposes in the short term. In these analyses, strategies should be tested for maximizing valid inclusion of students who take IAA and English language learners (see Attachment G).

2.2 ISBE should conduct an exploratory study of the feasibility of developing local use student growth models across elementary and secondary levels with existing aggregated data from schools using the Educational Planning and Assessment System (EPAS).

3. Advise the Superintendent and the State Board on the selection and use of growth information for school and district accountability;

The recommendations contained in this report are intended to advise the Superintendent and the State Board as they plan for implementation of a student growth component in state, district and school accountability systems. In addition to the work of the GMWG, the TAC should be asked to make recommendations about including measures of student characteristics in a value added model and effective ways of representing and reporting student growth data at the school, district and state levels. The TAC should to receive and respond to findings from the studies recommended above to aid in the final selection of a growth model to be implemented by October 31, 2011.

Recommendations

3.1 Representatives of the GMWG should meet jointly via a telephone conference call with the state assessment TAC in spring 2011 to discuss findings from their respective work and advise on the final selection of a student growth model.

4. Evaluate progress and effectiveness of the development and use of growth information for accountability purposes at state, regional, and local levels.

As GMWG considered student growth models for accountability purposes, other groups, such as PEAC, have planned for the use of student growth data for principal and teacher evaluations. At the classroom, school, and district levels, there is great interest in accessing student growth data for instructional improvement. While it is unlikely that a single standards, assessment, and data system could equitably meet all these needs, it is important to maximize the use of shared data and coordination across all sectors of use so that representations of student growth are consistent and support instructional improvement and student attainment.

Recommendations

4.1 ISBE should coordinate activity and facilitate communication among various working groups and committees planning for the collection and use of student growth data for different purposes such as the PEAC, TAC, and Data Advisory Committee.

4.2 ISBE should appoint a stakeholder working group made up of end users such as teachers, principals, superintendents, assessment directors, special educators, and ELL experts to continue to provide input on the implementation and use of growth models.
Growth Model Working Group

We hope that this report proves useful as you move forward with this important endeavor. Please do not hesitate to contact the chair of the committee with questions or concerns.

Respectfully submitted by,

Lizanne DeStefano, Chair

On behalf of the Members of the Student Growth Working Group:

Carmen Acevedo
Ellen Cwick
Amy Alsop
Susy Woods
Malinda Aiello
Debbie Meisner-Bertauski
Mark Doan
Kathlene Shank
Robin Ehrhart
Colleen Legge
Mark Mitrovich
Amy Nowell
John Byrne
Laura Cresap
Robert Grimm
Joseph Matula

ISBE staff: Joyce Zurkowski, Connie Wise, Linda Tomlinson, Susie Morrison, Wilma Vanscyoc, Jim Palmer, Dan Brown
Attachment A

Membership of Growth Model Working Group
GROWTH MODEL WORKING GROUP
(6/14/10)

ISBE
Joyce Zurkowski
Connie Wise
Linda Tomlinson
Susie Morrison
Legal Representation: (Wilma Vanscyoc)

Organizations
Illinois Advisory Council on Bilingual Education:
Ms Carmen Acevedo
3124 Venard Road
Downers Grove, IL 50616
Dsmace4@att.net
630-258-4555(c)
630-968-4555(h)

Illinois Principal’s Association
Dr. Ellen Cwick(pronounced Swick), Principal
Vernon Hills High School
145 north Lakeview Parkway
Vernon Hills, IL 60061
Cwick.e@district128.org
847-932-2043

Illinois Federation of Teachers
Ms. Amy Alsop
IPT Union Professional Development Director
Illinios Federation of Teachers Metro-East Administrative Offices
Four Executive Drive
Fairview Hts, Il 62220
618-624-4373
aalsop@ift-aft.org

Illinois State Advisory Council on the Education of Children with Disabilities
Ms. Susy Woods, M.A.
Visiting Director, ODS
Office of Disability Services
One University Plaza, HRB 87
Springfield, Illinois 62703
217-638-8411(c)
217-206-6666(w)
Swood06s@uis.edu

Illinois Community College Board
Ms. Malinda Aiello
Director of Academic Affairs
401 East Capitol Avenue
Springfield, IL 62701-1711
217-524-5503
Malinda.aiello@illinois.gov

Illinois Board of Higher Education
Ms. Debbie Meisner-Bertauski
431 East Adams, 2nd Floor
Springfield, IL 62701
217-557-7347
Meisner-bertauski@ibhe.org
IBHE Back-up Rep ~ Michael Afolayan
217-557-7348
afolayan@ibhe.org

Illinois Association of School Administrators
Mr. Mark Doan, Supt.
Farmington Central CU #265
212 North Lightfoot Road
Farmington, IL 61531
309-245-1000
mdoan@dist265.com

IAASE
Dr. Kathlene Shank
Professor and Chair, Dept of Special Education
Eastern Illinois University
600 Lincoln Avenue
Charleston, IL 61920
217-581-5315
ksshank@eiu.edu

Illinois Education Association
Robin Ehrhart, Research Specialist
3440 Liberty Drive
Springfield, IL 62704
217-321-2330
Robin.ehrhart@ieanea.org
Colleen Legge  
Superintendent  
Kankakee SD 111  
240 Warren Avenue  
Kankakee, IL 60901  
815/933-0779  
colleen-legge@ksd111.org

Mark Mitrovich  
Superintendent  
Naperville CUSD 203  
203 W. Hillside Road  
Naperville, IL 60540-6589  
mmitrovich@naperville203.org

Ms. Amy Nowell  
Office of Performance Management  
Chicago Public Schools 299  
16th Floor  
125 South Clark Street  
Chicago, IL 60603  
amnowell@cps.k12.il.us

Mr. John Byrne  
Superintendent  
Community High School 218  
10701 South Kilpatrick Avenue  
Oak Lawn, IL 60453  
708/424-2000, ext 2500  
john.byrne@chsd218.org

Other
Dr. Laura Cresap  
Assistant Superintendent  
Mahomet-Seymour School District #3  
101 North Division Street  
PO Box 229  
Mahomet, Illinois 61853  
217-586-4995 (w)  
lcresap@ms.k12.il.us

Dr. Robert Grimm  
Assistant Superintendent for Personnel  
Township High School District 211  
1750 South Roselle road
Palatine, Illinois 60067-7379
847-755-6618
rgrimm@d211.org

Joseph J. Matula, Ph.D.
10642 Ridgewood Drive
Palos Park, IL 60464
708-361-6964
j-matula@govst.edu
Assistant Professor
Governors State University
Alternate email: jmatula@comcast.net

Technical Advisors
Richard Hill, Ph.D.
National Center for the Improvement of Educational Assessment, Inc.
P.O. Box 351
Dover, NH 03821-0351
603-516-7900
rhill@nciea.org

Michael Kolen, Ph.D.
Professor
University of Iowa
7 Penn Circle
Iowa City, IA 52242-1529
319-335-6429
michael.kolen@uiowa.edu
Attachment B

Sample Charge Letter
May 19, 2010

Dr. Robert Grimm  
Assistant Superintendent for Personnel  
Township High School District 211  
1750 South Roselle Road  
Palatine, IL  60067-7379  

Dear Dr. Grimm,  

I am pleased to have the opportunity to appoint you to the Student Growth Working Group that will meet regularly over the next 12 months to address the following charge:  

1. Identify and coordinate the various uses and timelines for assessing and using student growth information including, but not limited to, Public Act 096-0861, the State Fiscal Stabilization Fund Phase II, the Race to the Top application, and the Race to the Top Assessment consortia;  

2. Develop a rigorous process for evaluating different approaches to assessing student growth utilizing peer benchmarking, empirical analysis, stakeholder input and expert review;  

3. Advise the Superintendent and the State Board on the selection and use of growth information for instructional improvement, evaluation of teachers, principals and school effectiveness, resource allocation, school and district accountability, policy making; and  

4. Evaluate progress and effectiveness of the development and use of student growth information at state, regional and local levels.  

The Student Growth Working Group will be chaired by Dr. Lizanne DeStefano, University of Illinois, and will be made up of knowledgeable and influential members of key stakeholder groups, teachers, principals and superintendents, technical experts and ISBE and other state agency staff. The first meeting of the Working Group is scheduled for June 16, 2010 from 10:00 a.m. to 4:00 p.m. at the Illinois State Board of Education, 100 N. First Street, Springfield, Illinois. Lunch will be provided and mileage will be reimbursed.  

We look forward to seeing you on June 16th. Please RSVP to Jeannie Schulze at 217-782-0354.  

Sincerely,  

Christopher A. Koch, Ed.D.  
State Superintendent of Education
STUDENT GROWTH MODEL MEETING
June 16, 2010

NOTES

Introduction of the Working Group

Review of Charge
To understand the various needs for/uses of Growth Data in local, state and federal contexts
To advise ISBE on the design and use of growth data in Illinois

Use of Growth Measures
- ESEA Reauthorization
- RTTT (Accountability and Teacher and Leader Evaluation)
- RTTT-Assessment
- PA 96-0861 Performance Evaluation Act of 2010
- State Fiscal Stabilization Fund

ESEA and RTTT
Although ESEA Reauthorization is not final, the “Blueprint for Educational Reform” gives a good idea of what lies ahead. Illinois has adopted Common Core Standards and has entered into the PARCC consortium as a governing member. PARCC has most of the larger states. It includes the Achieve and the American Diploma Project and is consistent with many Illinois initiatives.

Within the Consortium, states will share standards and performance standards. The PARCC assessments include performance based events, through course and end of course assessments, and makes good use of technology.

1. 2010-2011 Common Core Standards/No new test
2. 2011-2012 Development
3. 2012-2013 Piloting items
4. 2013-2014 Piloting assessments; ISBE’s contract with ACT runs through 2014
5. 2014-2015 New Assessment operational

Massive amounts of professional development on the Common Core Standards and new assessments will be needed for teachers and principals. At the moment, there is uncertainty about what these changes will mean for them and suspicion that growth data will be used for evaluation when the standards and the assessments are changing. Neither ISAT or PSAE are aligned with Common Core standards. Until the new assessments are available, LEAs will need resources and TA to do performance assessment and growth analysis using local data.

Faculty in higher education will also have to be trained on the Common Core, new assessments and new evaluation systems.
ESEA, the Common Core and the assessments are based upon college and career readiness. There is some concern that students may be tracked into careers too early.

It is important for the Working Group to know what other states and localities are doing for growth assessment. We will have a report on that at the next meeting.

It is not entirely clear how SWD and ELLs are included in the Common Core and RTTT-Assessment. It is hoped that growth models will do a better job of representing SWD and ELL achievement than performance levels, so it is important that they are have opportunity to learn the Core Standards and are included in the assessment systems.

**PA 96-0861 Performance Evaluation Act of 2010**
- By September 1, 2012, teachers will be evaluated using 4 ratings: Excellent, Proficient, Needs Improvement, Unsatisfactory
- ISAT can’t be used for teacher evaluation
- Growth data must be reported for every subject that is taught
- Student performance/growth data must count as a significant part (50%) of evaluation for teachers/principals (for RTTT and other funded programs)
- Can use ISAT for school improvement for evaluating principals.

Beginning in 2012 there will be no waivers given to districts. All must report growth data.

Some districts will be using Explore/PLAN/ACT for this. LEAs are having trouble establishing the reliability and validity of their growth data. They could use TA from the state on how to do this.

*This working group could produce a set of guidelines for using student growth data at the local level. Performance Evaluation Advisory Council (PEAC) is also working to develop rules for LEAs. The Working Group and PEAC should coordinate, hold joint meetings, and issue joint papers.*

*The working group could commission an evaluation study on the reliability and validity of local evaluation systems.*

Given the emphasis on College and Career Readiness, other indicators are important:
- % students graduating from HS
- % entering postsecondary education
- % needing remediation
- % graduating from higher education

**Timeline for Implementation:**
- 2010-2011 beginning of evaluation system for principals and RTTT (tier 1 and 2 schools--super LEAs)
- 2011-2012 statewide training on template
• 2012-2013 beginning of evaluation system for teachers under RTTT. Data collection and support system developed
• 2014-2015 All teachers
• 2015-2016 all LEAs must evaluate based on student growth re: legislation

ISBE will develop a state template for assessment of teacher/principal and will offer statewide training on that template. There is also a tremendous need for examples of data systems that can be adopted locally and for web-based systems or tools to facilitate local use.

There are several models that are being used locally:
• Danielson model
• DuPage model
• Plainfield is implementing common assessments.
• Chicago
  • TAP model for teacher eval.
  • Gone down to the grade level at each school
  • Going down to teacher level is difficult and CPS is/may not be going there
  • Pay for performance
  • 50% growth and other 50% is observation

**State Fiscal Stability Fund (SFSF)**—Includes our goals as a group
10/31/10 School/District growth dates for info purposes

10/31/11 School/District growth data available for accountability purposes

10/31/11 Teacher input program for school/district use

**MOST IMPORTANT THINGS TO ADDRESS**

A. Gain understanding of Growth Models and Growth Data
   a) Understand differences in models
   b) What works for other states, what would work for Illinois?
   c) What would work at local level?

B. SFSF Requirements
   a) What is shared
   b) How it is shared
   c) Timeline

C. The Longitudinal Data System
   a) Contents
   b) Use for Growth Models

D. Elements of the Assessment Consortium.
E. Guidelines and Training for LEAs Re: Performance Evaluation and Growth Models
F. List of Student characteristics that need to be accounted for in the system.

**July Meeting ~ What are other states/localities doing?**
- Readings posted on website
- Presentations: Chicago Value Added, TN, CO
- TAC Participation on Working Group

**August Meeting ~ Data Systems in the State**
- SFSF ~ What should be presented for 10/31/10

**September Meeting ~ Assessment Consortium and How it is moving forward re: growth**
- September ~ Long term plans
- Student characteristics for a growth system

**To Do:**
**Establish a website and post materials from meeting including notes.**
Present at Meeting: Lizanne De Stefano, Chair; Kathlene Shank, recording; Amy Alsop; Carmen Acevedo; Ellen Cwick; Joseph Matula; Debbie Meisner-Bertauski; Robin Ehrhart; Mark Doan; John Byrne; Robin Ehrhart; Brian Durham; Linda Tomlinson, ISBE; Laura Cresap; Joyce Zurkowski, ISBE; Colleen Legge; Amy Nowell; Travis McQuire; Elizabeth Freemen; Sue Walter;

Item #1: Website

www.isbe.net/gmwg is the site access; readings and schedule of the meetings are available on the site. Meeting summary will also be on the site; these will not be formal minutes as such.

Item #2: Meeting Topics

Chair DeStefano shared that today we will discuss the Tennessee Model; the Chicago Value-Added Model; and the Colorado model.

In August we will look at the longitudinal data system; we will also talk about what an Illinois model should have in it.

Item #3 PEAC

Joseph Matula is on this Growth Model group and PEAC.

July 16th PEAC met; heard information on a survey that is in progress. The Superintendent from Evanston talked about how they are using tests in Evanston Skokie #65. The Charlotte Danielson Model and Learning Points also gave presentations. Another survey system was also shared that is used to evaluate principals.

A more thorough summary of this meeting is available on the ISBE website for this group. August 20th is the next meeting. The focus of this group is on principal and teacher evaluation. 50% of teacher and principal evaluation must be based on “student growth”. The state designed system will be used if a district and the teacher representative group (union) can not agree on a system then they will default to the state designed model.

One question asked was “what will happen with districts that have invested considerable resources in data models?” Chair DeStefano said she would be sure to ask the ISBE person sharing the longitudinal data system at the next meeting to address this question.
Members shared what their districts were doing relative to common core standards and components of evaluation systems. One member shared that we need to remember that districts across the state are in varying stages with some who have not had the resources to do as much as other districts have. A question was asked about districts continuing to align to the common core standards and state standards. The unofficial answer is to continue to align with the common core that is already done.

Another question was about having a timeline that showed the areas we would have common core in and when. Linda Tomlinson explained that we already have numerous dates pushing this work that must be met (See summary for June 16, 2010).

Chair DeStefano suggested it may be helpful if ISBE would provide information to districts on the common core standards and where the State is going so districts could plan and move forward so that work is not done that would ultimately be undone.

Linda Tomlinson suggested that it would be good to look at assessments and how these relate to principal and teacher evaluation. It will be necessary to have assessments that are reliable and valid that are aligned to the common core standards. Valid assessments are a key element.

Chair DeStefano shared her personal professional thought that districts will find it useful to continue to discuss the common core and alignment to it. That the common core will be around for a significant time; the ones that have been developed will not go away any time soon. Focusing on the core will provide a framework.

Concern about teachers’ being fully involved and feel like the work they have done is valued was shared. The statement was made that it is important that teachers can not feel disenfranchised.

Debbie Meisner-Bertauski shared that higher education will be bringing a group together to discuss what we need to do to bring post-secondary and high school professionals together to discuss how to move forward with curriculum mapping to make it more seamless between higher education and secondary education.

**Item # 4: Tennessee Growth Model (TVAAS)**

Chair DeStefano shared about this model. One reason Tennessee was picked was because it is one of the first systems approved, 2006. It is probably one of the most mature; has been used since 1998 with implementation beginning in 1992. It is a value-added model. Tennessee was one of the first schools to be given a “race to the top” grant.

Value added is a subset of how we can represent growth. “Status” is the way we currently represent growth in Illinois. (AYP is a status approach) “Improvement” looks at are students
doing better than the group last year; looking at cohorts so it is a way of looking at change. Status and Improvement approaches are not really growth models.

True growth models look at how much on average did students change; it looks at the same group and must be longitudinal.

Value-added is a form of growth model; it predicts where students should be if they continue on the same trajectory. If the student exceeds the prediction this is the “value added”.

The Illinois legislation does not same we must have a value-added model it says we must look at growth.

The Tennessee growth model (TVAAS) system uses value added analysis as a means of a meaningful way to calculate the influence a district and school have on the academic progress rates of their students. Chair DeStefano says this is a “giant” which produces significant amounts of data. Tennessee is known as keeping as many students in as possible in the system; it is known to be very inclusive. Three years of student data is used when looking at a teacher’s performance; five years of data is used to look at a given student’s performance. Students who take the alternate assessment are not included in the system. Tennessee has ways to look at different forms of delivery such as team teaching, resource room, departmentalization, etc.

Tennessee does not adjust for socio-economic, race, or other factors. It controls for prior achievement. DOE likes that there are no factors considered.

The strength in this system is that it is multivariate, longitudinal analyses.

Teachers must verify annually that the students attributed to them in the system are the ones they are responsible for; the students must also be in the school for a minimum amount of time.

This system requires, in Dr. DeStefano’s professional judgment, a “vertical scale”. The Tennessee tests have been stable and in place for a considerable length of time. It has been used in grades 3-8; it has been used in math, reading, science, and social sciences in Tennessee. (See page 19, Guide to United States Department of Education Growth Model Pilot Program 2005-2008, CCSSO.) It takes a considerable amount of psycho-metric expertise; Tennessee has really invested itself in psycho-metric expertise.

The Tennessee system does not share how data is analyzed; it is a very traditional value added model. The software is “proprietary”. Tennessee has a very strong statistical foundation.
Data-driven conversations have as a foundation 15 years of data. The system can drill down to a group of 45 or more. Teacher effect reports are generated. Individual student achievement data is also available; the school can give this to parents but parents can not access the data on their own.

Tests to work in this growth model must be “broad”; it must be sensitive at both ends of the continuum. Meaning the standard error of measurement is the same for low achievers and high achievers.

Under AYP “status” has been used but since 2006 “growth models” have also been useable. Three years of data must be used if “growth” is used. Early returns from seven states that used growth models under the NCLB pilot program indicate that growth models produce little to no difference in AYP ratings using growth model from the status models (see page 6, Implementer’s Guide to Growth Models, CCSSO). In Tennessee 7 schools did better with growth measure over status measure of AYP. These 7 schools tended to have lower performing students.

Dr. DeStefano shared that the powerful aspect is the availability of data overtime.

The data is based on “projection” of student performance and not actual performance of students. Three years of data is used. Gain of a school, teacher, or student is measured against state averages. It is linked to high stakes decisions such as “Hope” scholarship and “teacher evaluations”.

Joyce will follow-up relative to three questions: mobility of students and use of the student test scores in evaluation of special education teachers. Another question would be that 8% of teacher evaluation was based on student growth prior to “Race to the Top”, what is the percentage now.

There are aspects of the Tennessee model that do not match Illinois especially the vertical nature of the tests used. One assumption of the Tennessee model is that students are randomly assigned to teachers. This is a piece to keep in mind. It is not a super sensitive system. It is also based on “projection” rather than actual growth. It is a statistical model. There is a 33% swing between groups of students relative to exceeds, meets, does not meet.

The Tennessee model costs $2-3.00 per student.

Slides provided by Tennessee and shared by Dr. DeStefano will be on the group website.

Item #5 Chicago Value-added
This presentation was made by Amy Nowell.

This model is a value-added model. WCER (Wisconsin Center for Educational Research) is working with CPS; the focus has been on whether schools are performing or are effective. The slides from this presentation are also on the website for this group. Conditions are controlled for in this model. Real data is used and based on conditions that can be controlled. It looks at adjusted growth. The assumption is that what remains after controlling for conditions beyond control what is left is the "growth". ISAT is being used; it assumes it is a vertical scale test. Chicago control factors in the model are students’ prior performance; free/reduced lunch status; ELL; IEP/Special Education status; (primary disability); gender and this year race is being added. It is used in principal evaluation but not teacher evaluation. There is a swing between groups of students that adds to the standard error of measurement that makes this less useful for teacher evaluation.

This model uses grades 4-8 with grade 4 being the baseline with performance measured by the ISAT. The value added is the average student gain on ISAT over and above the district average over 3 years.

It is a multi-regression model; it is "means" based. It makes "correction for pretest measurement error". It also corrects for "shrinkage" which relates to school size. The bias is that there is higher growth for higher achieving schools.

Every year data uses that year's data; there is no projection. It is not based on multiple years of data; it is based on one year. Pre-test is always Spring; the pre-test is the spring ISAT.

This model controls for pre-test measurement error and student's changing schools (dosage).

This is a classroom level model; there are data issues linking students to teachers. There is also difficulty with multiple teacher effects such as special education resource rooms. Data is an overall school and grade level value added in reading and math. Data does provide that CPS can differentiate between top 10% schools and lowest 10% schools but rest are in the middle.

Value added is used in "performance policy". It is used as 25% in the principal evaluation. Trend is used in the principal evaluation. CPS will need to move to use it in teacher evaluation.

CPS concerns include "stability of value added". Movement between categories wasn't as great as once thought might be true.

"Growth" was not part of the CPS focus whereas in Tennessee it was; also, Tennessee is based on "projection" whereas CPS uses "real" data. CPS has determined what it will control for
whereas Tennessee does not control for any factors. Student is at the heart of the Tennessee model.

Item # 6: The Colorado Model

This presentation was made by a representative of the Colorado Department of Education, Bill Bonk.

Joyce shared that the Colorado model is being used by 12-15 states; not all states are using it in the same way but rather in varying ways. This conversation is just in the initial stages.

The basis of the growth model is a “normative comparison”. Test scores in themselves have no real meaning. The point of the Colorado growth model is that data must be considered in a human context. Test scores should not be viewed as precise. The best way to understand scores is not in terms of a vertical scale but rather to look at change in scores and what this means for learners that started in differing places. The model uses the current test score as the dependent variable. It is a regression growth model (Additional information, see page 15-16 of the CCSSO document, Guide to United States Department of Education Growth Model Pilot Program 2005-2008).

It is a norm model and the scores change every year. It does not need or assume vertical scores. This model is based on the set of previous scores and looks at where the new score fits. It is expressed as a percentile score. It gives you “growth percentiles”. As it is a norm model there will always be the lower percentages and the higher percentages. Colorado is using its Colorado tests across grades 3-10 in reading, math, and writing plus the Colorado ACT at grade 10.

Current growth is seen as a leading indicator. A goal is that this data is communicable to stakeholders. How to visually represent data has been a very important consideration and commitment of time in Colorado. The system has been explained on a website which is being re-launched this fall. Colorado’s goal is to have the data be transparent at all levels and understandable at all levels.

The essence is to look at test scores in one year and compare to one, two, and three years later; only scaled scores are used. It compares the score with other students who started in the same place. Scores are only compared to others who started at the same point across the state. Two years data across the state are used to arrive at average percentiles then three years is used to arrive at average percentiles. For each individual student the maximum number of years of data
is used (could be for each grade 3 through 10). A quantile regression is used with the 
expectation that all students will reach proficiency within 3 years of by 10th grade.

Colorado is not using a value-added model; no demographics are considered. It does not know 
student attributes or which school. It is normative at its base and produces individual 
interpretations. This is not a cohort model; the group is mathematically constructed.

The model produces individual interpretations. Compares growth to a peer group that was 
mathematically determined; 50th percentile is the norm. 50% is not the target; it is not the 
standard. Median is used not "means". Medians must be used as it is percentile based. Medians 
are used to account for differences in intervals.

This model is used at the district and school level. Colorado does not currently have an 
application of this model to teacher evaluation.

Data is displayed and made available to stakeholders. There are two versions. School and 
district growth summary reports are provided. There is a "SchoolView.org" website; the 
protected level for number is 20. Data in the new version will go back to 2003.

The model was approved by the DOE in the growth model pilot program process but Colorado 
elected not to use it in place of AYP. Colorado is currently using the status AYP measurement 
for NCLB not this growth model.

Although factors are not controlled for group data can be broken out by school, grades, race, etc.

A question asked of Dr. Bonk was "capacity" necessary to do the work he is describing. Dr. 
Bonk responded that they have had multiple sources of funding, including private. He said there 
is an open source called "R" where Illinois could run its data. Oracle itself costs $1 million. If a 
State signs an article of understanding with the Colorado State Department of Education the 
State' can use what has been developed by Colorado for free.

Professional development has been focused at the district level not the teacher level.

Data for spring testing is reported by mid-August.

Colorado will be introducing the concept of "adequate growth". This is based on "catch up and 
keep-up growth". Since it is norm referenced there is always 50% growth. "Catch-up Growth" 
refers to crossing into proficiency; "Keeping up" represents maintenance. The performance 
framework looks at these factors; this framework is being used for districts, schools, and varying 
groups (free and reduced lunch, ELL, special education excluding those being test through 
alternative assessments, plus two others). The "growth is normed" not the scores.
Dr. Bonk knew of one district using the data as “one factor” in the teacher evaluation process. Colorado as a state uses this information as description information. The data is not used as causal information. It is not viewed as precise and as a singular information point but rather one place to begin the conversation.

The metric is “growth” not test scores.

In Colorado discussion of growth models began at the district level in 2001; the State adopted the model in 2007. It became a topic of research for the Colorado Department of Education in 2008.

Dr. DeStefano reflected that this model is complex.

**THE NEXT MEETING of this group is AUGUST 30th in Springfield, IL. at 10:00 a.m.**
GROWTH MODEL MEETING
AUGUST 30, 2010
10:00 CDT-3:15

Present: Dr. Lizanne DeStefano, Chair; Chris Koch, State Superintendent of Schools; Joyce Zurkowski, ISBE; Joseph Matula; Amy Nowell; Mark Doan; Amy Alsup; Carmen Acevado; Travis McQuire; Mike McKindles, guest speaker, ISBE; Jeremy Shoen, guest speaker, IBM; Nellie Kelley, guest speaker, IBM; Terry Chamberlain, guest speaker, ISBE; Robin Ehrhart; Linda Tomlinson, ISBE; Debbie Meisner-Bertauski, IBHE; Laura Cresap; Jamie Craven; Connie Wise, ISBE; Colleen Legge; Bob Grimm; Ellen Cwick; Wilma VanSyoc, ISBE; Brian Durham; Susy Woods; Beth Yacobi; Mark Mitovich; and Kathlene Shank.

Item #1: Refresher on our Work and Overview of this Meeting

Chair DeStefano reviewed what we have done in our first two meetings. Summaries have been emailed to committee members relative to both of these meetings.

Chair DeStefano stated that today’s meeting would commence with an up-date on Race to the Top; that in this meeting we would spend considerable time on “Illinois Longitudinal Data System project; and presentations would be given on “District assessments of student growth”.

**Dr. Koch joined us as we began the overview of the Illinois Longitudinal Data System and stayed with us for part of McKindles presentation (see Item # 4 below). Dr. DeStefano shared an overview of our work as it meshes with PCAP and PARCC. She asked Dr. Koch about the “Assessments for Learning” RFP and its status due to its relationship to our work. Dr. Koch responded that all this relates to reauthorization of ESEA. He said ISBE continues to look at ways to fund these efforts. Dr. DeStefano said this group’s recommendation would be to move ahead with the formative assessments of learning as districts would really benefit from this work.

Item #2: PCAP Up-Date by Joseph Matula

Joseph Matula is on both committees, PCAP and Growth Models; he shared that PCAP at its August 20 meeting had presentations on three principal and teacher evaluation systems. PCAP has broken down into sub-committees and Joe is on the group looking at “multiple measures of student growth”. PCAP has district people and teachers. This group is moving forward on
developing a state template; this may occur as early as its next meeting. The model developed by this committee will focus on performance.

Dr. DeStefano addressed how this committee and our work go together. Dr. DeStefano noted that there is real correlation between the concerns and thoughts expressed on this committee and on PCAP. Dr. DeStefano asked about the “principal performance review”; she suggested we may want to look at this document. Dr. Tomlinson said this document in on the ISBBE website.

**Item #3 Race to the Top**

Joyce Zurkowski explained there are two aspects of the Race to the Top. The state grant fund competition which we were not awarded; we were #15 and only 10 states were funded.

The other aspect of Race to the Top is the focus on assessment; there are two consortia competing for these funds. Illinois has joined one of these two; we joined PARCC. This consortium provides for states to have a more involved voice. Specific issues are currently being discussed at meetings including consortium members. The next meeting of the consortium will involve higher education relative to the college level common goal component. The consortium will work together in development of the end of high school assessment and when this is developed this test will be the test used as the end of high school assessment. Dr. DeStefano said that states are suppose to have autonomy as a member of PARCC and able to make its own discrete decisions. Amy Alsop asked what subject areas were being addressed; Joyce said the focus is on English, language arts, and Math at this point. Joyce said part of the discussion will be about how the information and efforts of the states which got Race to the Top will be accessed and used by the consortium. There is a total of 350 million available over four years.

In September this committee will further explore PARCC and the work ahead of this consortium.

Dr. DeStefano asked about the formative assessment item development RFP that was included in our Race to the Top proposal as to whether this will still occur. Connie Wise responded this was still under consideration; Chair DeStefano said she felt this work was important to our work.

**Item # 4 Illinois Longitudinal Data System (See handout: Illinois State Board of Education, Illinois Longitudinal Data System on committee web-site)**

Illinois Longitudinal Data System-

Mike McKindles, ISBE, shared an overview of the Project. (**Dr. Koch joined us as we began this presentation. See Item #1 above).
Mr. McKindles said that there are federal and state funds providing financial support of this work. Dr. DeStefano commended ISBE for getting the significant grants that funded this work.

There are five components including a “State Education Data Advisory Committee whose work is in progress. There are twenty eight members from varying stakeholders. This group has met six times and is helping in defining the directions and evaluation criteria. This group will be involved in implementation relative to communication, pilot testing, on-going feedback, and training needs.

A second component involves developing an “enterprise-wide data architecture”. This will result in a blueprint for the Education Enterprise Data Warehouse.

Number 3 component is to “establish a system of data stewards and enhanced data audit procedures”. Goals are to work to reduce redundancy; ensure accuracy; provide staff training in use of data; and provide technical assistance. An important aspect of this is sustainability.

Component #4 is the “data warehouse”. There will be a public side; there will be support of research agendas. Providing federal data is a very important element. This will include data form varying currently existing systems. The data warehouse will provide for linking data.

Component #5 is to “link the ISBE student data with Postsecondary and employment data and use the data for research and analysis”. The question was asked about this linking relative to Community Colleges. Mr. McKindles responded that there is a consortium that includes community colleges and four year institutions and this consortium includes 95% of the institutions providing higher education. This consortium is already looking at data being collected and to be collected in the future.

Mr. McKindles shared the timeline; all five components are underway with some level of activity. The data warehouse piece has just been initiated (Component #4); design should commence in January 2011 with some initial pieces coming out of the warehouse by December 2011.

Dr. Koch again mentioned that ESEA will really affect this work and our work as we do not want to go too far and get ahead of reauthorization.

Joseph Matula asked who the governing body is over the data; Mike McKindles answered that ISBE is the governing body and the Advisory Group for the Illinois Longitudinal Data System. Another question asked was about access to data for research; Mike responded the intent is for this to enhance access to data for researches not interfere with this type of work.

There are three focus groups working with the Advisory Group: Data Advisory Committee, Enterprise-wide Data Architecture (PCG has been selected as a vendor and is doing work relative to the architecture component); and Data Stewardship. Relative to Data Stewardship two
positions of four needed have been filled and analysis is underway. Maine, Arkansas, New Jersey, and Louisiana are also working with PCG on data system architecture.

The RFSP for Data Warehouse to seek a vendor is in progress. Data sharing agreements are in place to enhance the data available and use of the data; high school feedback report data sharing activities are ongoing. Linking of ISBE data with Postsecondary and Employment Data and use of this data for research and analysis is the goal. Eventually ACT will produce the report for us. One of the challenges is being able to match the data. The student ID needs to link from high school to higher education. Dr. DeStefano asked what would be the student identifier; ISBE currently assigns an identifier and this is the identifier that will be used. It is not currently used by public higher education. Publics will link but privates will have an option.

Activities in the SLDS Expansion Grant (ARRA) Project include four components.

#1: Statewide Transcript System for Middle and High School Students

A system that ties teacher to student will be established. It will capture courses, grades, and teachers; it will allow for multiple teachers. The current thinking is that it would be set up so that one teacher could be identified as primary and others as secondary. A question asked was who would have access and McKindles answered this was not his decision but the intent was not at this time to have full public access. To do this courses would have to be defined and this is progress for vocational and technical courses. Connie Wise answered there was work in progress cross-walking state definitions to federal which would be shared with superintendents this fall.

A comment about access to this data and teacher evaluations and the confidential nature of these was made. Confidentiality is important. Mike responded we are developing the system but governance lies in other entities. This is a very sensitive area.

ISBE is collecting data that has as a possibility of being used to assess teachers' effectiveness. Evaluation of teachers is a local district responsibility thus use of data is a local decision and the data feed for this is not yet under consideration. Data feeds being established at this time are specific to the “transcript system for middle and high school students”. Data feeds include to “State Transcript Depository”, ISAC; and postsecondary institutions for student applications and enrollment functions.

A question was asked about students in special education and McKindles responded we currently collect significant data related to special education and the system of data collection for special education will be refined and modified.

Part of the definition of courses in progress includes remedial, honors, and other aspects of who is in courses and differences in courses. Carmen Acevedo brought up issues with grading and
what a grade represents; McKindles said a teacher focus group also expressed this concern and that what grades mean is very subjective. How useful grades truly are is a big question.

Dr. DeStefano asked about the relationship of this work to data related to growth.

The second component is “Integration of Student Level Data with Teacher and Administrator Data”. This will involve “redesign of ISBE teacher and administrator data collection systems”.

The third component is “Improvement of Postsecondary and Workforce Data Collection through Greater Access to Public and Private Higher Education Data and Incorporation of Unemployment Insurance Wage Records into the Data Collection Systems”. The link will be to the final exit assessment at the end of the high school. One goal is to plan and design modifications to High School Feedback Report to expand upon information regarding postsecondary enrollment and performance. Another goal is to plan and design a process to incorporate data from the National Student Clearinghouse to better track transfer and concurrent enrollment of students.

The fourth component is “Expansion of Early Childhood Data Collection System”. Goal is to expand currently collected student information.

This funded set of activities recently funded by ARRA is less advanced than the previous work described. The timeline reflects the four components of this work on-going through 2013. The “Data Quality Campaign” identified 10 essential elements of data collection and we either already do these or it is in progress. “America Competes” has 12 essential elements and we do most of these. The P-20 Act specifies data to be collected and Illinois is aligned to these.

Illinois is collecting data from very collection systems including: Student Information System; Special Education Data System and Special Education Approval and Reimbursement System; Child Nutrition; Educator Certification System; Electronic Grants Management System; Annual Financial Report; Financial Reimbursement Information System; Non-Public Registration; Facilities and Inventory; Federal Career and Technical Education; Bilingual Education Reports; and Regional and Truants Alternative and Optional Education Program reporting.

What don’t we currently collect in Illinois: daily attendance; student health information; grades and course information; library records; AP information; transportation records; formative assessment information.

Two members of the Growth Model group are on the Illinois Longitudinal Data System Advisory Committee also and will be our liaisons to this advisory committee’s work, Amy Alsop and Amy Nowell.
Item #5: Student Information System presentation by representatives of IBM (Jeremy Shoen and Nellie Kelley-See handout on website: Growth Model Presentation)

This project has been in progress since 2005. The project piloted collection of student data in 2005 and additional data collections have been added each year. (IBM is the vendor and the contract goes through 2013.)

Each student is assigned a unique student identifier number (SID). The project collects demographic, enrollment, performance, and program participation data for each student. Dr. DeStefano asked about how districts use this SID. Responses to this query included that some districts use this ID number and others assign an ID and link it to the SID.

This system interfaces with internal ISBE systems. It tracks students from school/facility to school/facility and district to district within the state. It provides schools, districts, and facilities the ability to report timely and accurate data/information through standardized reporting capabilities.

A question was asked as to whether data relative to enrollment was reported differently last year on the school report card and Joyce said she would look at this and get an answer back.

Assessment data received from vendors includes: ACCESS, I, ISAT, and PSAE. Student data received from the schools has approximately 350 data elements, including native language, home language, race/ethnicity, and gender. Program indicators include: migrant, career/tech education, SES, homeless, LEP, IEP, FRL/low income, Century 21, eligibility for immigrant education, Reading 1st, Reading Improvement Block Grant Program and Title 1.

Resource materials are available on the ISBE SIS webpage at www.isbe.net/sis/. A question was asked about entities that have access to the information provided to a district. The response was that the superintendent of the district determined who got access and at what levels and which information.

The question was asked how SIS fits into the data warehouse. The answer was that SIS collects data whereas the data warehouse is how it is housed for use.

Item #6 Reflections on what we have heard thus far today

It was mentioned that SIS does collect information on characteristics of learners that we previously had talked about may be factors we would want to consider in a growth model, such as SES, native language, and IEP.

Another comment related to the differences in districts relative to level of use of data and resources/capacity to support data use and collection was made. Mr. McKindles talked about capability to collect and level of collection does come into play.
Dr. DeStefano commented that levels of access to information are an important consideration. She also commented that with the data warehouse system there had to be a link with districts similar to the one SIS has. Another concern is that the Data Warehouse planned is for middle and high school and does not include high school.

It was brought up that consideration had to be given to classroom teacher access to data if the data is to help teachers to make decisions.

Dr. DeStefano asked about the real time processing of data so it is useable by teachers. Also, collection of formative data is not being addressed as the current data collected is all summative.

A comment made also addressed the affordability of efforts as small districts may not have the capacity.

Dr. DeStefano asked about “data stewards”. Ms. Wise responded by giving an example that currently a data steward looks at data quality issues by examining data from what year to a next.

Another comment made spoke to being sure that we consider carefully and plan for “roll out” of a growth model and training to understand and use data, what it means and what it does not mean.

A clarification was made that PARCC will develop the assessment to be used but not be the determiner of what course the assessment aligns to.

Another comment was made that it is important that the work districts are already engaged in will be considered as we move ahead as some districts are moving ahead relative to measuring student learning. We need to keep in mind what districts have and are doing. A state wide survey was suggested as a way to look at capacity. A question was asked of Ms. Wise if there were efforts to help districts relative to increasing capacity. The response was not really but there were some efforts to help districts with technology.

**Item #7: Sharing by Committee Members of Local Perspectives**

Carmen Acevedo: Carmen shared the “Plainfield Community Consolidated School District 202: The Journey towards Growth”. The district has common assessments for PreK-12. The district partners with the “Curriculum Leadership Institute”. The district has subject area committees. For each subject the curriculum is examined relative to what is taught and it is aligned to learner goals. Gaps are found and curriculum is developed. After curriculum is developed assessment development occurs. Common assessments are validated before data is collected. Data collected is analyzed and recommendations are made for district-wide interventions.
An important aspect is that textbooks are considered a resource and not as constituting the curriculum. The outcome assessments can be customized as to format; this is done electronically. There is time set aside at elementary and middle grade levels for intervention.

Growth in enrollment and diversity has been significant in Plainfield. Plainfield has analyzed data for seven years in Math and Reading; it has looked at data relative to race and ethnicity and compared the district to other districts. AYP data has also been analyzed. Carmen conveyed that the district believes that the focus on development of a district curriculum, having resources that align and support the curriculum, the common assessments, and teacher development and ownership of the assessments have significantly contributed to the gains made by students in reading and math. The district is not using a growth model at this time. A key component is the electronic data analysis which assures data is readily available to the teacher. The reports provide information that is helpful in making classroom decisions. Plainfield has been involved in this work for five years. A question was asked as to whether five years was reasonable; could it be done in a shorter amount of time? Carmen answered that it took time to do this well. Also, Carmen said that teachers have to own the process and system and have to see value in the processes and this takes time.

Bob Grimm: Dr. Grimm presented on District 211’s (Palatine) Teacher Evaluation Pilot Program. This district tied the work of Charlotte Danielson to a high school setting. The domains are: Domain 1- Planning and Preparing for Classes; Domain 2-Classroom Environment; Domain 3- Instruction; and Domain 4- Professional Responsibility.

Common concepts are assessed with the assessment being formative; assessments are developed by teachers. These assessments are designed to “test for understanding”. (These have not been done in all subjects.) The assessments are used by the teachers and data is generated. Teachers meet in professional learning communities to examine the data from these subject matter specific measures of student growth and to make comparisons as to the extent given sets of students have learned given concepts. Discussion occurs as to how to help students that did not get it, get it and how different teachers approached teaching a given concept.

Teachers set their goals and this is used for evaluation. Goals must be agreed to by the evaluator. Currently most teachers use the “professional learning community” (PLC) goals.

A question posed is “To what extent does collective responsibility for student growth come into play.”

Item #8 What Comes Next
At our next meeting, September 27th, in Springfield we will look at PARCC as hopefully it will be funded.

Timing is an issue; the implication of timelines versus the time it takes to do assessment of student learning well is something we need to address.

Development of a test bank was discussed as being necessary to align to common core standards. Selection by teachers and buy in by teachers was also discussed. Part of the value of development of curriculum and common assessments is the process.

The lack of consistency in the preparation of teachers was brought up as an issue. Teachers don’t all come in ready and able to do the same things.

It was shared that according to a recent National Clearinghouse on Teacher Quality report the majority of teachers (70%) are not self-contained or single subject matter teachers for which achievement test scores directly measure what the teacher has taught. It is important we keep in focus that assessment of student learning goes beyond the common core areas (TQ Research & Policy Brief “Challenges in Evaluating Special Education Teachers and English Language Learner Specialists”, July 2010, www.TQSOURCE.ORG).

We are on multiple tracks according to our Chair. Teacher and principal evaluation must include measures of student growth. The state must also select a “Growth Model” to use for district and building accountability. The fourth would be for instructional purposes.

A suggestion was made that at the next meeting we collectively generate a list of ways we perceive student growth can be measured.

Data quality is important; also important is the interconnectiveness of assessments. Linking assessments to varying uses is also a consideration.

We agreed by show of hands to send a letter to Superintendent Koch to encourage development of an item bank specific to the common core that could be used by teachers and districts. We are not suggesting this be like the one described in Race to the Top. Other states have done this and these item banks are not stagnant.

NEXT MEETING IS SEPTEMBER 27, 2010-ISBE, Springfield; 10:00 a.m. CDT.
Growth Model October 23 Meeting Summary

(Next meeting is December 1, 2010)

Present: Dr. DeStefano, Chair; Kathlene Shank, summarizing; Amy Alsop; Ellen Cwick; Susy Woods; Daniel Brown and Jim Palmer, ISBE Assessment; Debbie Meisner-Bertauski; Laura Cresap; Joseph Matula (also, PEAC); Mark Doan; Wilma Vanscyoc, ISBE; Connie Wise, ISBE; Robin Ehrhart; Michael Afolayan (left at noon); Amy Nowell; Carmen Acevedo; Sue Morrison; Mark Mitovich; John Byrne; Robert Grimm; and Elizabeth Freeman, ISBE.

Agenda Item # 1: Charge of the Workgroup

Connie Wise shared with us relative to our charge and a December goal. She also announced that Joyce Zurkowski is leaving Illinois; Daniel Brown and Jim Palmer will be joining us.

Superintendent Koch has asked us to focus on state and district accountability.

Dr. DeStefano suggested we look at short term relative to Illinois’ current assessment system and suggestions to PARCC consortium. She shared she will continue to talk to the Chair of PEAC and that Joseph Matula will continue to keep us informed.

Amy Alsop shared her concern that it was important to be mindful that the work can not be separated with a “wall in between”. She shared that we need to keep in mind the relationship between the two. Dr. DeStefano says she shares Amy’s thoughts.

Kathlene shared her concern with a December timeline given we only have this meeting and the next one is December 1, 2010. Dr. DeStefano sees that we could possibly agree on the characteristics of a growth model and short term recommendations.

Agenda Item #2: PARCC

The purpose of looking at this is to understand purposes and what the Consortium is proceeding to do. Dr. DeStefano sees what PARCC is doing as being “flexible”.

Daniel Brown shared information on PARCC; he shared the web sites and said that Jim Palmer would post these for us. We can GOOGLE PARCC. PARCC stands for “Partnership for Assessment of Readiness for College and Careers”. Dr. DeStefano shared that both consortiums share the same goals because the RFP guidelines for funding were prescriptive. PARCC has
more of the larger states and states with more diversity. Dr. DeStefano is on the national “TAC” which is the doe advisory committee that will advise the secretary of education and see the work of both consortia.

PARCC will look at an assessment system (28 states are in the PARCC consortium) that will use college and career readiness (CCR) as an anchor; measure rigorous content and students’ ability to apply that content; measure learning and provide information throughout the school year; and the plan is to leverage use of technology. SMART is looking at computer based assessments. Dr. DeStefano commented the differences between the two consortia at this time are “subtle”.

Dr. DeStefano reflected on where she perceives Illinois is relative to its Illinois Longitudinal Data System and our state capabilities and current limitations of data management.

PARCC will cover grades 3-11 and will look at three points of assessment (25%, 50%, and 75%); the assessments will be based on the current adopted common core in reading/English/language arts and math. There will be “pilot phase” and the states in the Consortium will be involved. There will contractors hired by next September. The goal is for assessments to be ready by 2014. Schools will need the scope and sequence in Jim Palmer’s opinion as soon as possible.

Dr. DeStefano brought up that teachers will need to be given technical assistance and technology support if this is going to work.

The money at this point is going to development of assessments; very little to states and none to districts. Amy Alsop asked where the districts are to get the funds to do what needs to be done relative to Common Core Standards and district related curriculum scope and sequence work. She also suggested we as a group need to be sure part of our recommendations have to do with funding and support for local districts. Dr. DeStefano said she agreed this should be one of our recommendations.

Kathlene shared her concerns that we have to think through how to assure the system is coherent as PERA (Performance Evaluation Reform Act) also has a timelines across 2014-2016.

Susy Wood asked if learners with significant disabilities are included in the thinking of development of assessments and that answer was “not at this time”.

One goal is that teachers will have an assessment system that provides as much for them a it asks from them. Another goal is for students to know their areas that need improvement as they move through not just at exit. The common assessment system will help leaders and policymakers make the case for support of education. Amy Nowell made the point that this is a “tool” of improvement; it will not in and of itself “improve education”. There will be a summative statement made at the end of 11th grade relative to college and career readiness of each student.
The high school plan is not clearly specified at this time. The tests by 2014 will be in English/reading/language arts; course does not mean “course”.

Illinois has made commitments related to PARCC to fully use the assessments by 2014-2015; to use the assessments in their state accountability systems; and to provide staff to support “partnerships” activities.

Kathlene Shank asked if these developed “assessments” are planned to be “piloted”, and assured to be “reliable” and “valid” by 2014. Mr. Brown said this is the plan and he recognizes this is an aggressive timeline.

The Theory of Action is: “Building collective capacity to dramatically increase the rates at which students graduate from high school prepared for success in college and the workplace.”

Carmen Acevedo brought up issues related to English Language Learners that will need to be considered.

Joe Matula brought up that teacher evaluations for non-tenured teachers occur in March and the timelines for these assessments do not seem consistent. He suggested this needs to be considered.

**Agenda Item #3: State Accountability between now and 2014-Our Recommendations**

The focus of our recommendations in Dr. DeStefano’s thinking is on school accountability not on teacher and principal accountability.

Dr. DeStefano shared the importance of identifying the purposes. CCSSO has summarized the four common purposes for using growth: provide additional information to parents on their student’s progress; enhance school accountability; improve teaching and learning; program evaluation; and evaluate teachers or principals.

Amy Nowell said she thought we were looking at school accountability. Dr. DeStefano responded affirmatively to this statement.

Joe Matula suggested that we look at a model that included same students from one month to the next, e.g. March to March; two years or more in a given district; and students that have attendance rates of 95% or more. Colleen Legge brought up the mobility piece. Kathlene brought up her concerns with what we have learned about the Tennessee and Colorado models and the need for at least 3 years of data. Colleen brought up need for fairness across districts/schools and diversity of learners, including disabilities. Mark Mitovich brought up our lack of sophistication in our current assessments and the lack of a level playing field across districts. Robin Ehrhart asked for clarification about the year to year comparison. Dr. DeStefano said we would use past performance to build the comparisons. Amy Alsop said that “it is critical to use the same measures across the state” and we need to compare “apples with apples”. 
Amy Alsop asked for clarification of “what can ISBE do?” Capacity is an issue. Dr. DeStefano responded that it is important to know what Illinois has. ISBE has four years of data on ISAT, IAA and the prairie state test. Carmen brought up for ELL we have 3 years of data.

Mark Doan asked about direction from ISBE relative to getting ISBE to 2015. Dr. DeStefano is talking about short term recommendation for an interim measure of growth. We currently have a status model. AYP is a status model.

Kathlene brought up the need for caution given that the states that have been using growth models have not been reporting data to districts for multiple years and the issues with ELL, special education, and other aspects of diversity.

Amy asked the relationship between using the model and making decisions relative low performing schools. Colleen asked about the growth model and the tests we are using. Colleen asked would the state be refining a method of using the AYP and growth model data.

Carmen brought up that being able to look at growth versus status could be helpful.

Dr. DeStefano said “it is additional information”.

Kathlene brought up that there is no doubt that AYP, a status model, will go away with reauthorization, and replaced by growth models but we needed to be sure that we address appropriate use.

We talked about the importance of our recommendations including recommendations about use.

Dr. DeStefano explained that one question we need to answer is how much growth is enough. We also need to consider what differences we are going to control for. Another question is “should it be the same for every student” rather than discriminating by group”. Also, are we talking about typical or normal growth versus expected rate of growth?

Statistical controls can be used in value added which can not be accomplished through a growth model that is not value added. For value added there is little evidence it is useful at the teacher level.

Illinois does not have a vertical scale in the test data we have so we must use a model that is compatible (e.g. Delaware). Delaware has a “transition matrix”. This model uses “value tables” so cut scores are set and become very important. In place of “value tables” quartiles” can be used. Kathlene asked who used “quartiles”? Dr. DeStefano says that Colorado uses a “quartile regression growth model”. The test is normed and then uses quartiles and can be used for sub groups.
Dr. DeStefano again said “we must use AYP” (is still in NCLB which will be reauthorized most likely in the year ahead).

Laura Cresap said she would want to see school data that does three things: district, school, and individual.

Attendance could be put into a “regression model”. Mark Mitovich brought up that here is technology out there that will disaggregate data.

John Byrne brought up teachers need for data. Amy Alsop brought up that there is no money to get this down to the teacher level of understanding and use and we need to focus on use at the district level.

A “difference gain score” requires a common scale. This model asks: “is the gain for a group higher or lower than average”? Validity relies on a strong vertical scale in Dr. DeStefano’s professional judgment this does not match the data we have in Illinois. This model does not do well with high and low performing schools; it does better with middle performing schools.

“Growth in relation to performance standards” (a linear equating model) if system is not well aligned this model is not a good one. Illinois has disconnects between ISAT and PSE. Colorado and Maryland have element of this. This one has problems for Illinois due to the tests we are currently using. The question asked is “did students stay at the same quartile”? Dr. DeStefano has not seen this used at the grade 3-8 level. Given we do not have multiple measures at the high school level this may not be a good one for us. This model is the one that helps look at whether that student is making progress.

Amy Nowell brought up that primary goal is school accountability relative to the decisions we are trying to make.

“Residual gain score” model does not require a common scale. It does better at both ends of performing schools versus “difference gain scores. Here we are looking at a group to see if it is higher or lower than average. This model does not require a common scale. These scores are easier to interpret.

Amy Alsop brought up we need to decide if we want something predictive.

Our current tests were not developed with growth in mind, according to Dr. DeStefano. We do not technically have a vertical scale. Question we have to answer is “do we want to go with a growth model that requires a vertical scale?”

John Byrne brought up “Explore”; it is not given to all students across the state or by all districts.
Dr. DeStefano shared information on “confidence intervals”. With some measures the band of error is so great the data is not useable. If we want to be sure that the data has confidence we need to use a “wide band” confidence interval. Given our instruments were not designed to be used we will want to recommend a “wide band”.

Missing data is an issue; for this reason we need to make rules on missing data. Precision is very important when you look at subgroups. We need to make a recommendation on the importance of transparency with missing data especially relative to sub groups.

We make want to make specific recommendations on student mobility. Some models keep students in that stay within a district; an adjustment can be made for number of months student is in the school. Focus can be on inclusion or exclusion depending on the recommendation.

Alternate Assessments are not included in most state growth models. The “Alternate Assessment” students in Illinois are currently included in state results. In Illinois we do not have the “Alternate Assessment” on a common scale. We could recommend they not be included in the “Growth Model” in this interim period. Kathlene brought up that the CCSSO documents say that one significant issue with the pilot state use of growth models was the unresolved issues with inclusion of students being assessed on “alternate assessments”.

Levels of reporting is another area where recommendations needs to be made. We discussed level of use. Opinions were expressed that districts use data at the student level and the model needs to do this. A dissenting opinion was expressed that in the interim this is not necessarily a good idea as the tests we have were not designed to be used for these purposes. We could recommend that local districts can use data in varying ways but the purpose of the state growth model is not to “drill down to student level data”.

Another decision point “Is how to use growth data in accountability decisions?” In addition to status is necessary until ESEA is reauthorized and AYP is no longer required. “Safe harbor” is federally defined so this would also have to continue so growth model would also be in addition to “safe harbor”.

Data analysis is another recommendation point. Average scores by time by grade assesses linearity of growth. False positives and false negatives are something we may want TAC to look at. Stability of scores over time of the ISAT scores is an issue that would effect useability of the data in a growth model.
We are recommending what the State needs to consider in choosing a growth model. TAC will run some models and then that group will recommend which one to use.

Amy Alsop brought up that “Residual Gain Scores” predict and we may want to have TAC look at “Residual” versus “Transition”.

Agenda Item # 4 “Draft”

Dr. DeStefano will do a “draft” that we can read and think through prior to our December 1 meeting. Dr. DeStefano would like to see us have recommendations done to share with TAC which meets December 16 and 17.

Kathlene Shank shared her concern about this date and some of our members needs to share with the groups they represent.

Meeting adjourned at 1:40 p.m.
ISBE Growth Model Work Group (GMWG) Meeting April 7, 2011

Summary

Present: Chris Koch, State Superintendent of Schools, joined the group at 10:15 and left about 40 minutes later; Lizanne DeStefano, Chair; Kathlene Shank, recording; Deborah Larson; Amy Novell; Joseph Matula; Mark Doan; Amy Alsop; Colleen Legge; Bob Grimm, William Sayer, ISBE staff; Robin Ehrhart; Linda Tomlinson, ISBE staff; Connie Wise, ISBE staff; Jim Palmer, ISBE staff; Laura Cresap; Jamie Craven; Carmen Acevedo; Mark Mitovich; Melina Wright; John Byrne; Rense Lange, ISBE staff.

Agenda Item #1: Update on PEAC

Joe Matula reported that the teacher evaluation will be based on Charlotte Danielson. The topic currently being looked at measures of student learning and the principal evaluation is also in progress.

Time line is for principal evaluation is to include student growth as of September 2012.

Training is also being discussed; fidelity is something that must be addressed.

Agenda Item 2: Chris Koch remarks

Dr. Koch said this would be this group’s last meeting. That a “Technical Committee” will pick up where we are. He says reauthorization of ESEA is likely to include the topics we have been dealing with and he perceives subgroups will remain a focus.

Agenda Item #2: Technical Advisory Committee (TAC)

This group is national experts; Dr. DeStefano is included on this committee. Several of the experts also advise other states according to Dr. Koch. This group will meet in mid-April.

Agenda Item #3: PARCC Update

The Center for K-12 Assessment and Performance Management at ETS released in February 2011 the document “Coming Together to Raise Achievement”. Scope and sequence charts for English, Language Arts, and Math are in progress and will be out for public comment in June. In summer computer assessments will be examined and will be piloted next Fall.

PARCC includes 25 states trying to come together and agree on assessments. Math is currently troubling as issues are examined relative to integration across states. These will be interim assessments; a question was asked about timing of giving the assessments. Jim Palmer responded that there are issues as schools start at different times. He said it is likely we will have “continuous windows of assessment.” There are also issues of security and the number of “forms” of tests that will be needed. The plan is to compare scores across states.

The same skills are being taught in a sequence even if not at the same calendar time. Through course assessments including an interim are planned. The thinking now is the first interim would have the least
weight and should have some instructional value. PARCC is backing off of the assessments having instructional value.

How will PARCC compare across SMARTER Balanced Assessments was another question. The answer was that PARCC will be compared with PARCC states and SMARTER will be compared across the states in that consortium. PARCC INCLUDES MORE OF THE LARGE STATES.

The question was asked about ACHIEVE; the plan right now according to Jim Palmer is that these are to be included in the summary of assessments.

Chris Koch says he perceives the common core will hold.

Chris was asked about the relationship between the growth model and PARCC. Without knowing what the assessments look like it can not be said what assessment data would be used in the growth model. At a minimum it would be the “summative score”.

Currently grades 3-11 are being addressed by PARCC. Reauthorization may limit the grade range. Mitch Chester, Superintendent of Massachusetts chairs the PARCC meetings. ACHIEVE is also at the table. There are also higher education voices at the table and other entities in addition to ISBE.

Dr. DeStefano commented that in a meeting with GATES people the Illinois Longitudinal Data Systems was mentioned as being a very good model.

Jim Palmer also said that IBM is looking at applying artificial intelligence technology in the scoring of PARCC.

**Agenda item #4: Work Group Draft Report**

We went through page by page:

The references to SGWG should be GMWG throughout the document.

Page 1 no comments

Page 2 “The timeline initiating the teacher and principal performance evaluation process is September 1, 2012.

It was brought up we would like to have a timeline for PERA, PARCC, Illinois Longitudinal Data System, and the Growth Model.

Page 3 The sentence that data could be matched for teacher and students and coursework by September 30, 2011 was questioned. It was asked what exactly this means. Dr. DeStefano agreed to add a sentence that tells what this data can be used for.
Top of page 6: It was brought up that Tenn. and Ohio should be moved from under Projection Models and there should be a separate “VALUE ADDED” and put CPS, Tenn., and Ohio.

Paragraph right before Conclusions: “The SGWG evaluated these models.... Under 1.add “to begin to implement”...

Conclusions, under first bullet to be reworded: “For lack of a vertical scale that includes IAA and PSEA...

2cd bullet: Weak alignment .... Add words: “may limit the accuracy of projection models”.

4th bullet: Delete the bullet

Suggestion to add to the first full paragraph that starts: “The selection of a student growth model...” a sentence that addresses that depending on the growth there are winners and losers. Dr. DeStefano suggested that it be put it in the next paragraph.

It was requested that in the first sentence the sentence stop at will be used and eliminate the rest of the sentence beginning with including.... We agreed to change the wording “recognizing districts that produce the most growth, allocating resources..., and identifying. It was also suggested we add that it can be in addition to AYP.

It was brought up “permits inclusion” and eliminates the words “easy”. And it will say “permits inclusion”. The question was asked about “residual gain scores”. Dr. DeStefano said it is implied in the next paragraph: “A simple Post-on-Pre. In the sentence “Value added systems... words “residual gain scores” will be added.

Top paragraph page, at the end of the paragraph verbiage will be added as to what CPS is doing. “At grade level excluding high school:

Paragraph that starts...“Much of the discussion...” it was asked if it is worth the use of State resources to facilitate studies specific to EPAS (Educational Planning and Assessment System). Recommendation 2.3 is related to this discussion around EPAS; question was asked about Elementary being included. ACT does not support the use of EPAS as a measure of growth. It was decided to say that “considerable challenges exist” rather than “challenges are expected”. WCER studies should be considered.

Recommendations

2.1 Discussion: There was discussion about the State communicating the reasons for why a “particular model” was chosen. Wording will be changed to say this.

2.2 No comments

2.3 Suggestion to spell out EPAS as Educational Planning and Assessment System (EPAS). The question was asked why we are saying this as it was not part of this work group charge.

Under 3: Changed under 3.1, A representative to “Representatives...”
Page 3 second paragraph: To what extent sentence... it was brought up that we should add the word “reliable” along with valid in this sentence. It was also brought up that relative to PARCC districts will have to “scale up” to update curriculum. Spring 2012 PARCC may have some “stuff” ready for field testing.

Amy Alsop brought up that the political and social contexts are not addressed in our document.

Page 4 Conclusions insert the word “state” before accountability, to read: “student growth for State accountability”.

Re: 1.1, page 4, It was asked if ISAT should be specified. The answer was that ISAT can be used in the growth model BUT NOT as a measure of learner growth in the teacher and principal evaluation process. It was brought up whether data could be used along with or in place of AYP. To use the data the Feds would have to agree and the State Report Card would need to be modified. It was brought up that the 8th Grade ISAT and the PSEA do not equate.

Concerns were expressed relative to what can be accomplished by October 31, 2011. The issues expressed were around what “implement” means. We suggested adding the words “beginning to implement.” The March 2011 data will be the first ISAT data available to use for these purposes.

Also, on 1.1 (or when phrase) needs to be modified by as “or when two or more years of PARCC ...”. We agreed to this change.

It was brought up that IAA is included in the list along with ISAT and PSAE. We discussed adding sentences that says: Technical considerations as to how IAA results would be incorporated into the growth model have not been considered by this work group. It is recommended that the technical committee needs to consider the issues related with the use of the IAA.

1.2, we agreed that ISAT, IAA, PSEA will be listed here also. The last sentence is seen as a communication plan and it was suggested that this should be a separate recommendation. It was brought up that this additional recommendation should address a communication plan, training, roll out, legislature, and sub groups needs to be discussed. We can include this on page 8 in recommendations. topics that should be included are; uses of growth model, growth models can and can not tell you, issues related to IAA timelines for the diverse pieces, the transition to PARCC, issues related to specific populations such as special education and ELL and training and roll out.

The question was raised that it is a mistake to do away with this work group; on page 8 we could add a recommendation on an end user group being needed to work with TAC. We agreed we would add these words: “implications of state and local resources and funding” to sentence that starts out “report regularly...”

Recommendation 2: no changes
We discussed the need for an “end user” group in addition to TAC. It was decided that a recommendation be made and be put into 4. Under 4.1 under the e.g. add “Data Advisory Committee”.

We also agreed to recommend that ISBE appoint an “end user” advisory group to TAC.

Under 1.2, we decided to reword to be specific to a “designated individual to represent this work group.

**Agenda Item #5: Some Illinois Data on Three Value Added Approaches (Rense Lange and Andy Metcalf)**

“Post-on Pre: A finding relative to data analyzed using Post-on Pre (using adjacent grades) two period value added model was that “such a rule would have to vary according to the district level poverty. This means data “would have to be re-estimated for every subgroup.

A problem is that “many schools and certain classrooms are too small to do this reliably. Data had to be analyzed at the district level instead of schools, otherwise all conclusions are imprecise.

“Student Growth Percentiles” data “argues against the use of a single gain score” but the “method is better than Post-on Pre”. Smaller categories may work better.

“Value Tables or Transition Model” also called the “growth to a standard or trajectory model” tracks changes in performance levels from one year to the next for each student. The tables for a district summarize students’ changes in performance levels, rather than scale scores.

**Concluding Thoughts:**

Dr. DeStefano will send out the draft for a last look; we will not be word “smithing” but dealing only with the main thought and capturing changes we agree to.

This is the last meeting of this group; participants were thanked for their commitment to this work.

**Announcements:**

At the IAASE Spring meeting Lynne Holdeheide from NCTQ, National Comprehensive Center for Teacher Quality, Vanderbilt University, Nashville, Tenn. will give a session on Special Education and Principal Evaluation and the research NCTQ has done on this topic. The conference is in Collinsville, May 12 (Thursday) and May 13 (Friday). You can go to the Illinois Alliance of Administrators of Special Education (IAASE) web site for details of registration or call Norma Gerrish, 618-622-8800, for registration information. The registration cost is $125 and registration closes on May 5, 2011.
Attachment D

Summary of Federally-Approved Student Growth Models
Summary of Federally-Approved Student Growth Models
Date: February 7, 2010

Summary: After Secretary Margaret Spellings announced the United States Department of Education (USDE) growth pilot program in 2005, fifteen states have been approved to use student growth in their calculations of Adequate Yearly Progress (AYP). The growth pilot program announced in 2005 was designed to allow up to 10 states to include a growth measure in their federal accountability model. In December 2007, USDE announced that all states meeting eligibility requirements could apply for the growth pilot and that the pilot program was no longer limited to 10 total states. The most recent review of proposals resulted in the approval of four additional states for the growth pilot program, Colorado, Minnesota, Pennsylvania, and Texas.

The fifteen approved growth models can be classified into three general types of growth models, growth to proficiency models, value/transition tables, and regression-based models. Eight of the fifteen states were approved to use growth to proficiency models, three (Delaware, Minnesota, and Michigan) were approved to use value/transition tables, and four (Ohio, Pennsylvania, Tennessee, and Texas) were approved to use regression-based models.

1. Alaska
2. Arizona
3. Arkansas
4. Colorado
5. Delaware
6. Florida
7. Iowa
8. Michigan
9. Minnesota
10. Missouri
11. North Carolina
12. Ohio
13. Pennsylvania
14. Tennessee
15. Texas

The following pages summarize the growth models approved for the fifteen states and ways in which growth is used in these states’ accountability models.

Alaska

Growth Model Description: Alaska administers annual assessments in grades 3-10 in the content areas of reading, writing, and mathematics. Reading and Writing scores are combined to yield a Language Arts score. Alaska also administers annual assessments in grades 4, 8, and 10 in science. In Alaska, a score of 300 or above on the statewide assessment defines proficiency for each content area. Scores on these state assessments are not on vertical scales. For students with scores below 300 in mathematics, growth targets are set. For students with scores below 300 in language arts, growth targets are set. The targets depend on how far a student’s score is below the proficiency cut score, the student’s grade, and the results from the student’s base testing year. The first year that a student is tested is considered the student’s “base year,” and the student’s scaled score on that test is the student’s “base score.”

Using the student’s base score, a student will be assigned a “target score” to be achieved each of the subsequent years the student has to become proficient. If the student’s observed scaled score on the Standards Bases Assessment (SBA) is equal to or higher than the target score, and equal to or higher than the score from the previous grade level, the student will be considered to be “on track to becoming proficient” for that school year. If the observed scaled score on the SBA is less than the target or less than the score from the previous grade level, the student will be considered to not be on track and therefore will not count positively for his/her school. The target score will be calculated by first estimating the student’s true score (using classical measurement theory) for the base year. Making those calculations requires the grand mean for the state and the reliability of the SBA taken in that base year.

A student’s estimated true score (ETS) is calculated as follows:

ETS = Grand Mean + Reliability * (Observed Score – Grand Mean)
If the student’s base year is grades 3-6, the student is given four years to become proficient. If the base year is grade 7 or higher, then the student is given the difference between the base year and 10; for example, if the student’s base year is grade 7, the student is given 3 years to become proficient, and if the student’s base year is grade 9, the student is given one year to become proficient. Students must be proficient by the end of grade 10 to count positively for their school.

As an example, a student who starts in an LEA in grade 3 and scores 200, which is below the grade 3 proficiency score of 300, will have AYP determined by the growth model. Given the state mean for Math is 355, the student’s score of 200 is converted to an estimated true score: 

$$ETS = 355 + .91 \times (200 - 355),$$

or 214.

The increments for the above student are calculated as follows: proficiency score (300) minus the ETS (214) divided by the number of years to proficient (4) = 21.5.

For this student, the following growth targets apply:

- **Grade 4:** 235
- **Grade 5:** 257
- **Grade 6:** 278
- **Grade 7:** the student’s scale score is expected to be at proficiency, or 300.

**Student Growth in AYP Calculations:** The Alaska AYP growth model first evaluates how many subgroups in a campus and district meet minimum n requirements. For those subgroups and schools, participation requirements are then evaluated. Then, subgroups and schools are evaluated on the performance requirement, meaning the percentages of students meeting proficiency or on track to meeting proficiency under the growth model separately in language arts and mathematics are compared with the annual measurable objectives. The comparison includes use of a confidence interval. Then, schools are evaluated based on the other academic indicator. If schools and subgroups do not meet the performance requirement, Alaska implements an improvement or safe harbor provision.

---

**Arizona**

**Growth Model Description:** Arizona tests students annually in reading and mathematics in grades 3-8 and 10. The statewide test in Arizona has a vertical scale in grades 3-8 reading and mathematics, and these grades and subjects are used in the growth model. In Arizona, the growth model includes growth targets and predicted scores. The growth target part of the model requires that students reach proficiency within three years or by the eighth grade, whichever comes first. To calculate student growth targets, Arizona subtracts a student’s current year scale score from the scale score for proficiency three grades later and divides by the number of remaining grades. As an example (note that information was adapted from Arizona’s proposal, which has been posted at [http://www.ed.gov/lead/account/growthmodel/az/azgmpadd.doc](http://www.ed.gov/lead/account/growthmodel/az/azgmpadd.doc)), suppose a student scores 362 on the 3rd grade math test. The proficient cut score on the 6th grade math test is 496. The student’s math score must improve 45 points each year—(496-362)/(6-3) = 45—for her/him to reach proficiency by 6th grade. A regression is used to estimate a predicted score for each student based on previous grade 3 to 4 growth in each school. The predicted score for the above student is 417—this score is adjusted to the lower bound of the 97.5th percentile using the standard error of the prediction and equals 409. In order to count as proficient this student’s lower bound of the predicted score in 4th grade must be at least (362+45) 407. Given the lower bound of the predicted score is 409 and the target score is 407, this student made AYP.

**Student Growth in AYP Calculations:** For schools and subgroups to meet AYP proficiency targets, the number of students meeting growth targets is added to the number of students who are proficient but who did not meet growth targets. Then, this sum is divided by all students in the analysis. This proportion is
compared with the proficiency target for the year. If the value meets or exceeds the target, that school or subgroup meets the AYP performance requirements.

**Arkansas**

**Growth Model Description:** Arkansas administers language arts and mathematics exams in grades 3-8, and scores on these exams are on vertical scales. Arkansas also administers a literacy exam in grade 11 and an algebra I and geometry end-of-course exam. Arkansas implements nonlinear growth trajectories for students in grades 4-8 with the expectation that students will reach proficiency by eighth grade. Growth increments required to reach proficiency vary across the years as shown by the graph below (graph taken from Arkansas proposal at http://www.ed.gov/programs/lead/account/growthmodel/ar/argmp.doc).

![Graph of Growth Model](image)

**Student Growth in AYP Calculations:** In AYP calculations, a school or district is expected to meet the proficiency target that year based on the status model, the safe harbor model provision, or the target for individual growth. A district, school, or subgroup can meet safe harbor if the proportion of students in that school or subgroup scoring below proficient decreases by at least 10% from the prior year. The Arkansas growth model uses the proficiency targets from its status model to determine AYP. For example, under the Arkansas status model, the proficiency target for grades K-5 mathematics is that 64.08% of the students should be proficient in 2007-2008 in each school and subgroup within a school. Under the proposed growth model, 64.08% of the students in these grades in a school and in each subgroup in the school must make assessment gains in mathematics for the school to be making AYP for 2007-08, or the percentages of such students must meet the safe harbor standard. Required growth is calculated for all students, including those currently below proficient and for those at proficient or above.

**Colorado**

**Growth Model Description:** Colorado tests students by using the Colorado Student Assessment Program (CSAP) in grades 3-10. The content areas include reading/writing and mathematics, which are given in grades 3-10, and science which is given in grades 5, 8 and 10. The scores of reading/writing and mathematics of CSAP are on a vertical scale. Colorado also tests student in grades 11 by using the Colorado ACT, which
includes reading/writing, mathematics, and science, and in grades 4, 8, and 12 by using the National Assessment of Educational Progress (NAEP), which includes mathematics, reading/writing, science and pilots in Social Studies. Only the CSAP assessments (reading/writing and mathematics) will be used in the growth model analysis. For students with cognitive disabilities, they are given the Colorado Student Assessment Program Alternate (CSAPa). However, they are not included in the growth model analysis. Colorado will implement a quantile regression growth model with expectation that all students will reach proficiency within 3 years or by 10th grade. The quantile regression growth model is combined with all available prior test scores to determine whether each student is on track to be proficient or on track to maintain proficient within 3 years or by 10th grade, whichever comes first. In order to do so, the model initially calculates a growth percentile for each student relative to all other students in the state with the same prior academic history (academic peers). Then the growth percentile is compared to the percentile needed for that student to reach proficient within 3 years or by 10th grade, whichever comes first. If the growth percentile is larger than the percentile needed, the student is considered as "on track to reach proficiency" or "on track to maintain proficiency". If the growth percentile is smaller than the percentile needed, the student is not considered "on track."

**Student Growth in AYP Calculations:** Colorado will calculate the growth percentile separately for reading and mathematics. The AYP growth calculation adds the number of unsatisfactory students on track to be proficient to the number of proficient students on track to maintain proficiency. This numerator is divided by the total number of full academic year students in grades 4-10 with at least two valid CSAP scores in the districts/schools/disaggregated groups. The resulting percentage is then compared with the growth AMO to determine whether or not the districts/schools/disaggregated groups made AYP.

**Delaware**

**Growth Model Description:** Delaware tests students in reading and mathematics in grades 2-10, writing in grades 3 – 10, and science and social studies in four grades each (grades 4, 6, 8, and 11). Scores on these assessments are not on vertical scales.

Delaware implements a value tables approach to growth-based accountability using their reading and math scores from grades 2 - 10. Each student in a subgroup earns points for moving across proficiency levels. The number of points increases as students move from levels below proficiency to levels at proficiency. Students who regress contribute zero points.

Delaware has five proficiency levels. In the growth model, the two lowest levels are subdivided into two groups each. For example, if a student scored at Level 1A in Year 1 and Level 1B in Year 2, that result would earn the student's school a certain number of points (e.g., 150 according to the table below). The points in the value tables were set by committee.

An example of a value table used in Delaware (taken from a slide show at link http://www.ccsso.org/content/PDFs/DF_Model_2006_032706.ppt#392,7,Slide 7) is shown below. As shown in the table, the maximum average score for a group of students is 300, which is equivalent to 100% proficient. An average score of 300 would indicate that all students are meeting the standards.
**Student Growth in AYP Calculations:** In Delaware, a school or subgroup meets AYP if that school or subgroup meets three conditions:
- 1. proficiency targets in reading and mathematics (meets growth targets)
- 2. meets participation target
- 3. meets other academic indicator requirements

The growth target for a school or subgroup in any one year is calculated as the proficiency target times 300. For example in 2007 the proficiency target for English language arts was 68%. The growth target was then calculated as 68% of 300 or 204. A school or subgroup needed to have an average growth value of at least 204 to meet growth expectations.

---

**Florida**

**Growth Model Description:** Florida administers its state assessment in grades 3-11. The test contains criterion-referenced tests (measuring selected benchmarks in mathematics, reading, science, and writing) and norm-referenced tests in reading and mathematics (measuring individual student performance against national norms). Florida’s state assessment is reported on a vertical scale from grades 3 to 10 with scale values ranging from 0 to 3000. Florida expects students who are not proficient to reach proficiency in three years. Growth targets are set by taking the difference between the proficiency score three years later and the student’s initial test score. Students are expected to score at least a third of that difference greater each year. For example, the amount of improvement in terms of decreasing the score difference between the initial score and the proficiency point three years in the future is shown in the table (table information taken from Florida’s proposal at link http://www.ed.gov/admins/lead/account/growthmodel/fl/firevisions2006.doc) below.

<table>
<thead>
<tr>
<th>Year In State-Tested Grade</th>
<th>Decrease From Baseline Assessment In Performance Discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33% of original gap</td>
</tr>
<tr>
<td>2</td>
<td>66% of original gap</td>
</tr>
<tr>
<td>3</td>
<td>Student must be proficient</td>
</tr>
</tbody>
</table>

**Student Growth in AYP Calculations:** Under Florida’s growth proposal, each subgroup will have AYP calculated using the status model, safe harbor criteria and a growth model calculation. A school or subgroup makes AYP if it meets participation rate requirements; meets the writing requirement and the graduation rate requirement (as applicable); and meets the proficiency target that year based on the status model, satisfies the safe harbor model provision, or meets the target for individual growth.
**Iowa**

**Growth Model Description:** Iowa districts test all grade 3-8 students using the Iowa Tests of Basic Skills (ITBS) which has a vertical scale. Districts also test all grade 11 students using the Iowa Tests of Educational Development (ITED). Iowa’s growth model was approved for grades 3-8 only. To set growth targets for non-proficient students, Iowa divides the below proficient scale score range for grades 3-8 into three categories and has established category boundaries on the scale score system for non-proficient students across grades. A student’s growth trajectory must cross a category boundary in order to be considered to have met growth expectations. In their US ED growth proposal, Iowa decided to label the meeting of growth expectations as Adequate Yearly Growth. Adequate Yearly Growth is defined as the score improvement that non-proficient students are expected to make from one year to the next. The figure below (taken from Iowa’s proposal at link [http://www.ed.gov/lead/account/growthmodel/ia/iajmp07.doc](http://www.ed.gov/lead/account/growthmodel/ia/iajmp07.doc)) shows the category boundaries for non-proficient students across grades. A student’s growth trajectory must cross a category boundary in order to be considered for Adequate Yearly Growth.

![Iowa's Categorical Growth Model](image)

**Student Growth in AYP Calculations:** In AYP calculations, Iowa adds the number of students who meet proficiency and those who meet growth targets and divides by all students in the analysis. The proportion is compared with the proficiency target for that year. When safe harbor is applied, students who meet growth targets are included in the analyses. For example, when safe harbor is examined, the number of students meeting AYP is combined with the number of students meeting growth for each year. This combined percentage is subtracted from 100% to determine the percent of non-proficient students, in the prior year and current year. If the percent of non-proficient students is reduced by 10% or more, from the prior year to the current year, the group meets safe harbor.

**Michigan**

**Growth Model Description:** Michigan administers the Michigan Educational Assessment Program (MEAP) in grades 3-8, the Michigan Merit Examination to all students in grade 11, and the MI-Access for students with disabilities. Beginning in 2005-2006, the MEAP and MI-Access were expanded to
grades 3-8. The Michigan growth model divides each of the four MEAP performance levels (not proficient, partially proficient, proficient, and advanced) into three sub-levels (low, middle, and high). A similar process divided the MI-Access levels into sub-levels, yet fewer sub-levels were used with this alternate assessment. Then, the model sets expectations and tracks student transitions from one year to the next. Based on the numbers of transitions a student must make and the number of years to achieve proficiency, each student is given an improvement target. For example, a student in the low Not Proficient category who needs to make 6 transitions in 3 years to reach proficiency would need to make 2 transitions each year. The improvement target for this student would be 2. See table below (note that this table was taken from Michigan’s proposal) for student improvement targets.

The tracking mechanism is called a transition value table. The transition tables compare student performance to increasing expectations and indicate whether students are declining, whether they are exhibiting no change, or whether they are improving their standing. Students’ change in performance level is classified into five categories (significant decline, decline, no change, improvement, significant improvement) with accompanying abbreviations (SD, D, N, I, SI, respectively).

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Previous Performance</th>
<th>Number of Sub-Levels Improvement Needed to Achieve Proficiency</th>
<th>Number of Years to Achieve Proficiency</th>
<th>Improvement Target</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level</td>
<td>Sub-Level</td>
<td>Unrounded</td>
<td>Rounded</td>
</tr>
<tr>
<td>MEAP</td>
<td>Not Proficient</td>
<td>Low</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Partially</td>
<td>Low</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Proficient</td>
<td>Mid</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>MI-Access</td>
<td>Emerging</td>
<td>Low</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Functional</td>
<td></td>
<td>Mid</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Independence</td>
<td></td>
<td>High</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>MI-Access</td>
<td>Emerging</td>
<td>No Sub Divisions</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Participation &amp; Supported</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Student Growth in AYP Calculations:** Michigan determines whether districts, schools, and subgroups within districts and schools meet the AYP proficiency targets using the following:

\[
100 \times \frac{(n_{\text{Proficient}} + n_{\text{On Trajectory}} + n_{\text{Provisional}})}{n_{\text{Valid Assessments}}}
\]

where \(n_{\text{Proficient}}\) is the number of proficient students, \(n_{\text{On Trajectory}}\) is the number of students on trajectory to proficiency within the next three years, \(n_{\text{Provisional}}\) is the number of provisionally proficient students (i.e., students who meet proficiency when a confidence interval is applied), and \(n_{\text{Valid Assessments}}\) is the number of students receiving valid scores on the assessment (non-valid scores cannot count toward participation rates and are not used in proficiency rate calculations).

**Minnesota**

**Growth Model Description:** Minnesota administered Minnesota comprehensive Assessments (MCAs) as well as Minnesota Test of Academic Skills (MTAS) for students with disabilities in grade 3-8, 10, 11 and high school. The tests assess reading, mathematics and science. Reading is tested in grades 3-8 and 10. Mathematics is tested in grades 3-8 and 11. Science tests are given in grade 5 and 8, and once in high school, depending on when students complete their life sciences curriculum. For English learners, the Mathematics Test for English Language Learners (MTEL.L) is given in grades 3-8 and 11, and Test of Emerging Academic
English (TEAE) is given in grades 3-12 for testing reading and writing. Only reading and mathematics tests are used to determine whether schools and districts have made adequate yearly progress (AYP) toward all students being proficient in 2014. Scores on these two subjects are on vertical scales in grades 3-8, but the vertical scale does not extend to the high school. Minnesota will implement a value table growth model into the current federal accountability system. There are 4 achievement levels in the MN value table (see the table below, which is taken from the proposal of Minnesota), and the two lowest levels (i.e., “Does Not Meet Low” and “Partially Meets”) is subdivided into two groups (i.e., “low” and “high”). Each student in a subgroup earns points for moving from a lower achievement level or sublevel to a higher achievement level or sublevel during one year. The points increase as the discrepancy between the two levels increases and the highest point in this table is 100. If the student makes two consecutive years of growth, then the students is eligible for additional points, called compounding points, which is one-half the difference in the points for the next highest performance range. At a campus level, students’ average points are used in the accountability system. These points in the value table were determined based on the same points values used in the status model for scale compatibility with the AMOs\(^1\) and observations of actual student performance from prior to current year on the assessments.

<table>
<thead>
<tr>
<th>Points Awarded</th>
<th>Does Not Meet LOW</th>
<th>Does Not Meet HIGH</th>
<th>Partially Meets LOW</th>
<th>Partially Meets HIGH</th>
<th>Meets</th>
<th>Exceeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does Not Meet LOW</td>
<td>0</td>
<td>50</td>
<td>65</td>
<td>80</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Does Not Meet HIGH</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>75</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Partially Meets LOW</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>65</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Partially Meets HIGH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Meets</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Exceeds</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

**Student Growth in AYP Calculations:** In Minnesota, a school or subgroup makes AYP if the school or subgroup meets the following criteria:
1. The school in total and each subgroup test at least 95 percent of students;
2. The school demonstrates a 90 percent attendance rate or 0.1 percent improvement in the percentage of students in attendance;
3. The school demonstrates an 80 percent graduation rate or a 0.1 percent improvement in graduation rate;
4. The school in total and each subgroup meet the reading and mathematics proportion proficient target or Safe Harbor provisions or the Growth Model provisions.

Minnesota will use the annual measurable objectives (AMOs) established in Minnesota’s approved Accountability Workbook as the growth targets (see Evidence 1.1.1.1, pages 25-27). By using the growth model, the school MUST meet the criterion 1-3. Then the school’s growth score will be compared to the AMO target for a subject. If the growth score exceeds the AMO, the school will make AYP. The school’s growth score is calculated as:

\[
Growth\_score = \left( \frac{Total\_points\_earned\_for\_all\_students}{Number\_of\_all\_students} \right) \times 0.01
\]

\(^1\) Minnesota stakeholders have supported that the same set of AMOs be used for growth and for status determinations. As a result, Minnesota had to align points awarded in the value table to points currently being awarded for the performance index. For example, under the status model, “Partially Meets Standards” students earn 0.5 points; so, in the growth model, students who remain Partially Meets Standards must also earn 0.5 points, i.e., 50 (as shown in the value table) × .01, to have a comparable scale.
Besides using growth in AYP, Minnesota will continue to hold schools and districts accountable for universal proficiency by 2013-14 using the other two methods (i.e., a combination of status and safe harbor) in determining AYP for schools, districts, and the state.

**Missouri**

**Growth Model Description:** The Missouri Assessment Program (MAP) includes assessments in grades 3-8, and in one grade at the high school level annually in communication arts, mathematics, and science. Beginning in 2008-2009, high school assessments will be replaced by end-of-course assessments in English II, Algebra I, and Biology. Students whose significant cognitive disabilities prevent them from participating in MAP subject-area assessments are assessed with the Missouri Assessment Program-Alternate (MAP-A).

The first year a student tests in a Missouri public school is the baseline year. Student growth targets are established using that baseline score to determine the scores that the student should achieve in each subsequent year of testing to be proficient at the end of four years, or by the end of grade 8, whichever occurs first. Growth targets are calculated as the numeric difference between the student’s scale score in the baseline year (grade 3 for the majority of students) and the scale score defining proficiency at the end of the target grade level.

Growth targets represent the amount of improvement (in terms of scale score changes) the student should show each year in order to reach proficiency by the target grade level (the earlier of grade 8, or four years from the baseline score). The table below (taken from Missouri’s proposal at link http://www.ed.gov/admins/lead/account/growthmodel/mo/mogmp.doc) illustrates growth targets for baseline scores determined in each grade 3-7.

<table>
<thead>
<tr>
<th>Baseline Grade (Status)</th>
<th>Year 1 Benchmark</th>
<th>Year 2 Benchmark</th>
<th>Year 3 Benchmark</th>
<th>Target Grade (Status)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Grade 3 ¼ distance from baseline to grade 7</td>
<td>Grade 5 ¼ distance from baseline to grade 7</td>
<td>Grade 6 ¼ distance from baseline to grade 7</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>Grade 5 ¼ distance from baseline to grade 8</td>
<td>Grade 6 ½ distance from baseline to grade 8</td>
<td>Grade 7 ¼ distance from baseline to grade 8</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>Grade 6 1/3 distance from baseline to grade 8</td>
<td>Grade 7 2/3 distance from baseline to grade 8</td>
<td>N/A</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>Grade 7 ½ distance from baseline to grade 8</td>
<td>N/A</td>
<td>N/A</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>Status for grade 7</td>
<td>N/A</td>
<td>N/A</td>
<td>8</td>
</tr>
</tbody>
</table>

**Student Growth in AYP Calculations:** Missouri incorporates a growth model calculation into its accountability system at grades 3-8, establishing unique growth trajectories that will ensure that, by 2014, all students will either be proficient or “on-track to be proficient” by the end of grade 8, or within four years of the baseline score, whichever is reached first. For buildings that do not make AYP based on status (as defined in Missouri’s current approved Accountability Workbook), assessment data will be analyzed at the student level to determine which students are “on-track to be proficient.” For each student, a growth trajectory will be calculated that will ensure that the student is “on-track to be proficient” in each content area within four years, or by the end of grade 8, whichever comes first, depending upon the grade level in which the student’s baseline score is determined. The number of students that are “on-track to be proficient” will be added to the numerator of the “Percent Proficient” calculation to determine AYP (based on the state’s
established AMOs identified in the approved Accountability Workbook) for each subgroup, school, district, and the state. All students that have been enrolled in the district for at least one full academic year (as defined in Missouri's current approved Accountability Workbook, Reference 1, p. 14) will be included in the denominator of this calculation.

**North Carolina**

**Growth Model Description:** North Carolina incorporates student test results from grades 3-8 in reading and mathematics. Both end of grade (EOG) and end of course (EOC) assessments are used in the growth model. North Carolina has a vertical scale in the grades 3-8 assessments, but it does not use that vertical scale in its growth model. To set student growth targets, North Carolina transforms students’ scores on to a common scale (using standard deviation units). Then, students who are not proficient are expected to lower by a minimum percent (e.g., 25%) each year the difference between the first test and the proficiency standard, typically four years later. North Carolina also gives a third grade pretest and uses that pretest in growth calculations. For example, a student in grade 3 who scores below proficiency on the pretest is expected to score, by the end of grade 3, 25% closer to the proficiency score by the end of grade 6. The North Carolina grades and tests used to define growth and the percent of score difference expected to be closed each year is presented in the table below (taken from North Carolina’s proposal at link http://www.ed.gov/lead/account/growthmodel/nc/ncgmp.doc).

<table>
<thead>
<tr>
<th>Grade Of First Enrollment</th>
<th>Test Used As The Basis For Prediction</th>
<th>Test Used As Target For Proficiency</th>
<th>Years In Trajectory</th>
<th>Percent Of Difference Closed Per Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3rd grade pretest</td>
<td>6th grade EOG</td>
<td>4</td>
<td>25%</td>
</tr>
<tr>
<td>4</td>
<td>3rd grade EOG</td>
<td>7th grade EOG</td>
<td>4</td>
<td>25%</td>
</tr>
<tr>
<td>5</td>
<td>4th grade EOG</td>
<td>8th grade EOG</td>
<td>4</td>
<td>25%</td>
</tr>
<tr>
<td>6</td>
<td>5th grade EOG</td>
<td>Algebra I or English I EOC</td>
<td>4</td>
<td>25%</td>
</tr>
<tr>
<td>7</td>
<td>6th grade EOG</td>
<td>Algebra I or English I EOC</td>
<td>4</td>
<td>25%</td>
</tr>
<tr>
<td>8</td>
<td>7th grade EOG</td>
<td>Algebra I or English I EOC</td>
<td>3</td>
<td>33%</td>
</tr>
</tbody>
</table>

**Student Growth in AYP Calculations:** In its AYP calculations, North Carolina adds the number of students who score at or above proficiency to those students (below proficiency) meeting growth targets and divides by the number of students in the analysis. The proportion is compared with the proficiency target (i.e., annual measurable objective) for that year. The state runs the AYP growth calculations only after a school or subgroup misses an AYP target using safe harbor or through the use of the confidence interval.

**Ohio**

**Growth Model Description:** Ohio tests lower grade level students in reading and mathematics in grades 3-8, writing grades 4 and 7, science in grades 5 and 8, and social studies in grades 5 and 8. Ohio also tests grade 10 students in reading, mathematics, writing, science, and social studies. Ohio works with Dr. Bill Sanders to calculate student growth. Ohio implements a multivariate longitudinal statistical model which projects student
growth based on up to five years of prior assessment data. For example, a student’s projected score in mathematics will be computed from that student’s prior mathematics, reading, science, and social studies scores. Ohio’s model counts as proficient (for AYP purposes) students who are on a path to reach proficiency within two years (with the exception of 7th grade, where students must reach proficiency by 8th grade).

**Student Growth in AYP Calculations:** Ohio evaluates projected scores separately for reading/language arts and mathematics. It should be noted that projection scores only are used in the determination of AYP for subgroups that were not proficient based on Ohio’s other methods of determining proficiency (meeting or exceeding the target, 2-Year Combined Results and Safe Harbor). When determining whether a subgroup has met the annual proficiency target in reading/language arts, Ohio determines if the subgroup has the sufficient percentage of students who are projected to be proficient within two years. The resulting percentage is compared with the proficiency target for the current year, and if the target is met, the subgroup is considered as having met AYP in the current year.

**Pennsylvania**

**Growth Model Description:** Pennsylvania administers the Pennsylvania System of School Assessment (PSSA) as well as the Pennsylvania Alternate System of Assessment (PASA) for students with disabilities in grades 3-8 and 11. The tests include reading, math, writing, and science. Reading and math is given in grades 3 through 8 and 11, writing is tested in grades 5, 8 and 11, and science is tested in grades 4, 8, and 11. Scores on these subjects are not on a vertical scale. Pennsylvania will implement a projection to proficiency model as an additional AYP method such that students are expected to have projected scores at least as high as the proficiency standard in the projection grade. Projection grades are specified in the table below. For example, a 4th or 5th grade student will be considered proficient if the student is projected to score above the proficiency standard two years into the future. All available achievement data (excluding demographic variables) on the students’ test history and the histories of students with similar performance patterns will be used to yield the projection. Students in grades 4-8 will have projections. Students in other grades will be included in the accountability model using their current-year scores, but students in grades 9-11 will not have projections. Additionally, students with significant cognitive disabilities assessed under PASA will not be included in the growth model either, and their current scores will be used for AYP determination.

<table>
<thead>
<tr>
<th>Present and Projected Grades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present Grade</td>
</tr>
<tr>
<td>Projection to Grade</td>
</tr>
</tbody>
</table>

**Student Growth in AYP Calculations:** Pennsylvania will calculate the projected scores separately for reading and mathematics. Based on the student’s projection calculations, schools will be assigned credit for all students who are projected to be proficient in a specified grade in the future, whether they are currently below proficient or are currently proficient. Then the percent of students projected to be proficient on a future PSSA examination will then be used to determine AYP status based on the presently accepted Annual Measurable Objectives (AMO) status targets.

**Tennessee**

**Growth Model Description:** Tennessee tests students in grades 3-8 in reading/language arts, mathematics, science, and social studies. In high school grades, Tennessee assesses students in end-of-course assessments, including Algebra I, Mathematics Foundations II, English I, English II, and Biology I, Physics, and US History. Students must pass Algebra I, English II, and Biology I in order to earn a high school diploma. Tennessee works with Dr. Bill Sanders and implements a statistically complex projection model such that
students projections, based on average growth in future years, for the targeted grade are compared to the proficiency standard (three years in the future). The growth model applies to students in grades 4-8. For example, a fourth grade student must be projected to be at or above proficiency in grade 7 to meet the growth target in Tennessee's AYP calculations. For example, a third grade student must be projected to be at or above proficiency in grade 6 to meet the growth target in Tennessee’s AYP calculations. Tennessee’s criterion referenced assessments in grades 3-8 are on a vertical scale, though the vertical scale is not used in the projection calculations. To compare scores for purposes of growth, Tennessee transforms all assessment scores to a normal curve equivalent scale.

**Student Growth in AYP Calculations:** Schools and subgroups in Tennessee have two options for meeting elementary and middle school AYP proficiency targets. The first way is to have the percent of students scoring at proficiency or higher at least as great as the proficiency target for Tennessee that year. The second option is to have the percent of students with scores in both reading and mathematics projected three years later as proficient or higher at least as high as the proficiency target for that year. One feature of the second option is that a student who scores at least proficient in the current year but has a projected score three years later that is below proficient will not add to the numerator of the percent of students compared with the proficiency target for that year.

**Impact on Schools and Districts**
Table 5 summarizes the impact of incorporating growth into states’ AYP calculations. The number and percent of districts and campuses who met AYP because of growth in the noted assessment year are those districts and campuses who would not have met AYP had growth not been included. Results suggest that the impact of adding the growth model to state’s AYP calculations varies. For some states, such as Alaska and North Carolina, the impact has been small. For other states, the impact has been much greater. The Ohio projection model, for example, has made a large impact on the numbers of districts and campuses meeting AYP. The reasons for the variability in impact are many and complex, including the conditions under which the model was reviewed, the way the model is calculated, the number of years for which students are able to meet growth targets yet not meet proficiency expectations, and the ways in which growth is incorporated into the AYP calculations. The differential impact relates as much to the type of model implemented as to the way in which growth information is integrated into the AYP calculations.

**Texas**

**Growth Model Description:** The Texas student assessment program assesses students using the Texas Assessment of Knowledge and Skills (TAKS) and TAKS (Accommodated), where these assessments are administered in English in grades 3–11 mathematics; grades 3–9 reading; grades 10–11 English language arts; grades 4 and 7 writing; grades 5, 8, 10, and 11 science; and grades 8, 10, and 11 social studies. Spanish TAKS and TAKS (Accommodated) is administered in grades 3–6 reading and mathematics, grade 4 writing, and grade 5 science. Linguistically accommodated versions of English and Spanish TAKS are administered in grades 3-8 and 10 for reading/English language arts and mathematics and in science in grades 5, 8, and 10. The TAKS-Modified and TAKS-Alternate are offered in the same grades and subjects as are offered for English TAKS. The Texas English Language Proficiency Assessment System (TELPAS) is administered to English language learners in grades K-12. In addition, an exit level Texas Assessment of Academic Skills (TAAAS) test is available for high school students and out-of-school examinees for whom this test is a graduation requirement and an optional Algebra I, Biology, Chemistry, Geometry, and US History End-of-Course assessment are available for students who have completed these respective courses.

Texas will implement a student projection measure, called the Texas Projection Measure (TPM), in calculations of Adequate Yearly Progress (AYP) starting in 2009. The TPM is a multi-level regression method for projecting future student scores in the next high-stakes grade (defined by Texas legislation as grades 5, 8, and 11) using students' current year scale scores in reading/English language arts and mathematics and
average campus scale score in the projection subject (i.e., reading campus mean for reading projections and mathematics campus mean for mathematics projections). Projection equations are developed the year before they are applied, so that the formulas can be established and shared across the state before they are used in Adequate Yearly Progress (AYP) calculations. For example, projection equations developed in 2008 will be applied in 2009 to project student performance. The projection equations will be updated each year after operational testing and will be published before their use the next spring. Projection equations are transparent. Projections are made separately for reading/English language arts and for mathematics. The TPM will be used in 2009 for students taking the Texas Assessment of Knowledge and Skills (TAKS), TAKS (Accommodated), and linguistically accommodated versions of TAKS. Starting in 2010, when academic achievement standards are available for the TAKS—Modified (TAKS—M) and TAKS—Alternate (TAKS—Alt) assessments, Texas proposes to expand the TPM for use with students taking the TAKS—M assessments and implement a transition table approach to growth for students participating in TAKS—Alt.

**Student Growth in AYP Calculations:** Adding the TPM to AYP calculations will result in one change to the Texas AYP determination process. To meet AYP in Texas under the current process, for all districts and campuses, all students and each student group (African American, Hispanic, white, economically disadvantaged, special education, and limited English proficient) meeting minimum size requirements must meet (1) either the performance standard for percent proficient or performance gains criteria, and (2) the standard for participation in the assessment program. The inclusion of the TPM would impact the way the performance standard for percent proficient is calculated. Students who are projected to meet proficiency with the TPM will be counted in the numerator of the AYP percent proficiency calculation along with students meeting the standard, and this new percentage would be compared with the AYP targets to determine if the performance standard for percent proficient is met. For all students and each student group, AYP performance standard requirements would be met if the percent proficient or projected to be proficient, for grades 3–8 and 10 summed across grades by subject for reading/English language arts and mathematics, meets or exceeds the AYP targets. The inclusion of the TPM in AYP calculations will not change the way the performance gains criteria are applied.
Attachment E

Meyer and Dokumaci White Paper on Value-added Models
VALUE-ADDED MODELS AND THE NEXT GENERATION OF ASSESSMENTS

Dr. Robert H. Meyer and Dr. Emin Dokumaci
Value-Added Research Center, University of Wisconsin-Madison

This paper discusses some of the fundamental features of value-added models, with particular focus on the interaction between the design and interpretation of value-added models and the design and properties of student assessments. We present a case study using actual state and district data from Wisconsin.

A value-added model is a quasi-experimental statistical model that yields estimates of the contribution of schools, classrooms, teachers, or other educational units, to student achievement (or other student outcomes), controlling for other (non-school) sources of student achievement growth, including prior student achievement and student and family characteristics. The model produces estimates of school productivity – value-added indicators – under the counterfactual assumption that all schools serve the same group of students. This facilitates “apples and apples” school comparisons rather than “apples and oranges comparisons.” The objective is to facilitate valid and fair comparisons of student outcomes across schools given that the schools may serve very different student populations.

A useful (high quality) value-added model produces indicators of educational productivity at multiple levels of the educational system that are valid and reliable (in the sense of accurately measuring educational productivity). The degree to which a particular value-added system produces high-quality value-added indicators depends directly on five major factors:

1. The quality and appropriateness of the student outcomes (for example, mathematics and reading achievement) used to measure value-added productivity.
2. The availability and quality of longitudinal data on students, teachers, and schools, particularly the degree to which students, classrooms/courses, and teachers are correctly linked.
3. The design of the value-added model (or models) used to produce measures of value-added productivity (and associated measures of the statistical precision of productivity).
4. The volume of data available to estimate the model.

1 Although we are responsible for the views and analyses presented in this paper, we particularly thank our colleagues at the Value-Added Research Center and Deb Lindsey, Director of Research and Assessment in the Milwaukee Public Schools. Our work has benefited from extensive interactions with our district and state research partners, including the public school systems in Chicago, Madison, Milwaukee, Minneapolis, New York City, and Racine, the state of Wisconsin, and the Wisconsin CESAs. This paper draws on research funded by the Wisconsin Department of Public Instruction, the Milwaukee Public Schools, the Joyce Foundation, the Walton Family Foundation, and the Institute for Education Sciences.
2 Although value-added systems have generally measured educational productivity using student test scores, the Value-Added Research Center has recently worked with the Milwaukee Public Schools to develop value-added models for non-assessment outcomes, such as student attendance.
3 At the micro (unit) level, the precision of value-added estimates depends directly on the number of student observations available for a given educational unit. The amount of usable data can be increased by pooling data for given educational units over time. At the macro level, the interpretative utility of value-added indicators depends on
5. The degree to which the student outcomes (and other variables included in a value-added model) are resistant to manipulation or distorted measurement.\textsuperscript{4}

We believe that it is possible to produce valid, reliable, and useful measures of educational productivity if sufficient attention is paid to addressing the five factors listed above. Our focus in this paper is primarily on factors one and three, the design and properties of assessments and the design and assessment requirements of value-added models. Our objective is to identify issues and priorities for enhancing the quality of value-added models and indicators with respect to these two factors.

**Value-Added Analysis in Context**

As a prelude to addressing the specific objectives of this paper, we briefly discuss some of the larger issues related to the appropriate use of value-added indicators in an education system.

In order for a value-added system to be a powerful “engine” of school improvement, it should be systemically aligned with the fundamental needs and operations of schools, districts, and states, including at least the following:\textsuperscript{5}

- Use to evaluate effectiveness of instructional practices, programs, and policies
- Embed within a framework of data-informed decision-making
- Align with school, district, and state policies, practices, and governance procedures
  - Vertical alignment: alignment across all levels of the system (including state, district, cascade of district management levels, school (multiple grades), grade-level team, classroom, teacher, student subgroup, student)
  - Horizontal alignment: alignment across departments and divisions at each level (for example, teaching and learning, human resources, and accountability)
- Provide extensive professional development to support understanding and application of value-added information

A well-developed and aligned value-added system can be used to stimulate school improvement via several different channels, including:

\textsuperscript{4} Accountability systems (including value-added systems or attainment/proficiency-based systems) based on assessments that are open to manipulation could distort the incentives (and thus behavior) of educational stakeholders. For example, if students are tested using the same test form year after year, narrow teaching to the test form (and not the content domain that underlies the test) could be effective in raising test scores without actually increasing student achievement. This problem can in principle be addressed by changing test forms each year. As discussed later in the paper, this requires test developers to horizontally equate test forms if it is desired to produce test scores that can be validly compared over time.

\textsuperscript{5} In our work at the Value-Added Research Center (VARC), we have explored the connections between value-added systems and school district needs and operations in the following reports: Carl, Cheng, and Meyer (2009); Carl, Cheng, and Meyer (2009); Lander, Keltz, Pautsch, and Meyer (2009); Jones, Hisar, and Meyer (2010).
• Provide evidence that all students can learn
• Facilitate triage: Identify and provide assistance to low performing schools or teachers
• Contribute to district knowledge about “what works” (including professional development)
• Incorporate within a performance management system
• Hold educational stakeholders accountable for high performance
• Provide bonuses to high-performing teachers, teams of teachers, and schools
• Provide information to teacher preparation institutions on the value-added performance of the teachers that they have trained

In most applications it is essential to use value-added information in conjunction with other sources of information, such as observational data (based on well-defined rubrics) or value-added information based on multiple student outcomes. In addition, it is typically sensible to use information from multiple years in order to dampen variability due to statistical noise and authentic variation in educational outcomes. Reliance on a single value-added indicator, as opposed to multiple indicators, could provide educators with an incentive to narrowly focus their efforts to improve measured student performance.

Finally, it is important to contrast the appropriateness of using value-added indicators as measures of educational productivity versus attainment/proficiency indicators of the type currently required under the No Child Left Behind legislation (NCLB). We address this issue later in the paper after providing a fuller description of a value-added model.

Description of a Simple Value-Added Model

Since our objective is to discuss conceptual issues in the design and interpretation of value-added models and assessments we intentionally focus on a relatively simple statewide (multi-district) value-added model of school productivity at a given grade level. We then discuss options for making the model more complex. Most, if not all, value-added models (including classroom and teacher value-added models) produce value-added parameters of the type included in this model.

The key features of the model are:

• Two years of (consecutive grade) longitudinal assessment data for each student
  (measured annually at the end or beginning of the school year).

---

6 Although it is beyond the scope of this paper to address the mechanisms by which value-added indicators can be used to stimulate school improvement, we will later discuss how value-added systems support the premise that all students can learn.
7 VARC is currently participating in a 10-year project in Minnesota, North Dakota, and South Dakota, funded by the Archibald Bush Foundation, to produce and provide value-added measures of teacher performance to teachers, teacher preparation institutions, and districts. For additional information on the Teacher Effectiveness Initiative see: http://www.bushfoundation.org/education/teinitiative.asp.
8 As an example, schools participating in the TAP system ("The System for Teacher and Student Advancement") use a blend of teacher/classroom-level value-added ratings, school-level value-added ratings, and observational ratings (based on a well-defined rubric) to rate teacher performance.
9 Note that since statewide testing begins in third grade in many states, only two years of (up-to-date) attainment data are typically available to estimate value-added models of achievement growth from third to fourth grade. In
- School/district value-added productivity effects $\eta_{kt}$ (for school $k$ in district $l$ in year $t$, at a given grade).
- Statewide value-added productivity effects $\pi_t$.
- A posttest on pretest “link” parameter $\lambda$, (which may vary across grades and over time). This parameter allows for the possibility that achievement growth may differ for students with high and low prior achievement and for situations where the distribution of the posttest and pretest variables may be non-uniform over time (more on this below).
- Demographic variables $X_u$ to capture differences across students (within-classrooms) in achievement growth.

Figure 1 provides a schematic diagram of this two-period value-added model.

Figure 1: Diagram of Two-Period State Value-Added Model

The model indicates that achievement at the end of a period (posttest $Y_2$) is the sum of:

1. Student achievement at the beginning of the period (pretest $Y_1$) times a posttest-on-pretest “link” parameter ($\lambda$).\(^{10}\)

---

\(^{10}\) As discussed later in the paper, the value-added model could be extended to include measures of prior achievement in multiple subject areas, as well as measures of non-cognitive outcomes. A model of mathematics achievement, for example, could include prior reading achievement, as well as prior mathematics achievement. We have found that including multiple measures of prior student outcomes typically increases the predictive power of a model and improves the accuracy of estimated value-added effects.
2. Student growth that is correlated with student characteristics such as income, English Language Learner (ELL) status, and race/ethnicity (\( \beta \)).
3. Statewide productivity (\( \pi \)) (see further explanation below).
4. School and district productivity (\( \eta \)) (see further explanation below).
5. Student growth that is due to unknown student characteristics and random test measurement error (\( \varepsilon \)).

The value-added productivity parameters produced by this model are defined in greater detail in Table 1.\(^{11}\)

<table>
<thead>
<tr>
<th>( \pi_t )</th>
<th>Statewide productivity in year ( t ) (for a given grade). Note that this parameter can only be interpreted as a genuine statewide productivity effect only if test scores are accurately horizontally equated over time so that changes in test score growth do not reflect test form effects. (The issue of horizontal equating is discussed later in the paper.) This parameter is typically estimated as a “contrast” effect relative to a baseline year. In this case, statewide productivity is equal to zero in the baseline year and productivity in other years is measured relative to productivity in the baseline year.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \eta_{klt} )</td>
<td>Relative school productivity for school ( k ) in district ( l ) in year ( t ) (for a given grade). This parameter is referred to as a “relative” value-added parameter because it is centered around zero in each year so that the average school in the district has a value-added rating equal to zero and school productivity is measured relative to the average school. Changes in statewide productivity are thus absorbed by the parameter ( \pi_t ).</td>
</tr>
<tr>
<td>( \eta_{klt}^{\text{ABSOLUTE}} )</td>
<td>Absolute (total) school, district, and state productivity. This indicator incorporates relative school productivity plus overall changes in statewide productivity, provided (as mentioned above) that test scores are accurately horizontally equated.</td>
</tr>
<tr>
<td>( \pi_t + \eta_{klt} )</td>
<td></td>
</tr>
</tbody>
</table>

The school/district productivity parameter defined above (either \( \eta_{klt} \) or \( \eta_{klt}^{\text{ABSOLUTE}} \)) is referred to as the “beat the average” (BTA) rating in the Milwaukee value-added system and the “beat the odds” rating in the Minneapolis value-added system because the value of the indicator equals the amount by which it exceeds or falls short of average district productivity in each year (in the case of \( \eta_{klt} \)) or in the baseline year (in the case of \( \eta_{klt}^{\text{ABSOLUTE}} \)).

The simple value-added model presented above provides a framework for discussing issues related to model design and assessment design and their interaction.

---

\(^{11}\) Appendix A presents the model using formal statistical notation and defines the model parameters and variables.
The Design of Value-Added Models

We begin by discussing two basic features of the two-period model: (1) the provision for connecting post and prior achievement via a post-on-pre linking parameter and (2) the inclusion of student-level demographic variables in the model.

*Post-on-Pre Link: The Coefficient on Prior Achievement*¹²

One of the important features of the value-added model considered above is that it allows for the possibility that the coefficient on prior achievement (λ) could differ across grades and years and might not equal one, a parameter restriction that is imposed in some value-added models. The model would be simpler to estimate if it was appropriate to impose the parameter restriction λ = 1, but there are at least four factors that could make this restriction invalid. First, λ could be less than one if the “stock” of knowledge, skill, and achievement captured by student assessments is not totally “durable,” but rather is subject to decay. Second, λ could differ from one if school resources are allocated differentially to students as a function of prior achievement. If resources are tilted relatively toward low achieving students – a remediation strategy – then λ would be reduced. The opposite would be true if resources were tilted toward high achieving students. Third, λ if posttest and pretest scores are measured on different scales, perhaps because the assessments administered in different grades are from different vendors and scored on different test scales or due to instability in the variability of test scores across grades and years. In this case, the coefficient on prior achievement partially reflects the difference in scale units between the pretest and posttest. Fourth, the different methods used to scale assessments could, in effect, transform posttest and pretest scores so that the relationship between post and prior achievement is nonlinear. In this case, a linear value-added model might still provide a reasonably accurate approximation of the achievement growth process, but the coefficient on prior achievement (as in the case of the third point) would be affected by the test scaling. See Meyer and Dokumaci (2009) for discussion of these issues.

In summary, there are four factors that could make it problematic to impose the parameter restriction that the coefficient in prior achievement are identical in all grades and years and equal to particular value (such as λ = 1): (1) durability/decay in achievement, (2) differential resource allocation, (3) differences in the pretest and posttest test scales, and (4) nonlinearity in the test scaling algorithm.

---

¹² Although our focus in this paper is primarily on model design issues rather than statistical estimation techniques it is important to point out that in order to properly estimate the value-added model presented in the text, it is necessary to account for measurement error in test scores (using methods from structural equation modeling (SEM)) and possible endogeneity due to correlation of prior achievement with the equation error term. We discuss the issue of test measurement error later in the paper. Fuller (1987) and Meyer (1992, 1999) discuss methods for correcting for measurement error. VARC has used these methods to control for test measurement error in all of its value-added systems. Methods for addressing the endogeneity of prior achievement have been developed by numerous researchers, including Anderson and Hsaio (1982). Arellano and Honore (2001) provide a recent survey of this literature.
In the case study presented later in the paper we explore whether the standard deviations of prior and post achievement test scores exhibit instabilities across grades and over time (as discussed above).

*Does Achievement Growth Differ for Students with Different Student Characteristics?*

An important feature of the value-added model presented above is that it includes explicit measures of student characteristics. Since most district or state value-added systems are based on administrative data bases (as opposed to special purpose data collections), value-added models generally include a limited number of student measures, for example, poverty status (participation in free or reduced-price lunch), participation in special education, participation in an English Language Learner (ELL) program, gender, or race/ethnicity.

Including measures of student characteristics in a value-added model serves two purposes. First, including these measures in a value-added model makes it possible to measure district or statewide differences in achievement growth by student subgroups (e.g., low vs. high poverty). We refer to these differences as “value-added growth gaps. They are analogous to attainment gaps. Growth gaps are in some ways more fundamental than attainment gaps, because attainment gaps arise via year-to-year accumulation of growth gaps. These statistics are important for policy purposes. Over time a district can monitor changes in growth gaps to evaluate the success of policies and programs designed to reduce inequality in student attainment and student growth.

The second purpose for including characteristics in a value-added model is to “control” for differences across schools in the student composition of schools so that estimates of educational performance reflect differences in school productivity, rather than differences in school composition. In other words, control variables (including prior achievement) are included in the model to achieve, to the extent possible, “apples and apples” school comparisons rather than “apples and oranges comparisons.” Models that fail to include student-level variables will yield results that are systematically biased against schools and educators that disproportionately serve students who, on average, exhibit relatively low within-classroom and within-school achievement growth (for example, low income students). In an era where public policy is focused on providing high quality teachers for all students, it seems particularly unwise to build educational productivity indicators that are biased against exactly the types of students that society is most eager to help.

---

13 It is common practice to include student demographic variables in statistical models of student test scores. For an early reference, see the well known Coleman study (Coleman, 1966).

14 The value-added model presented above allows for state or district-level growth gaps (by student subgroups) and changes over time in these growth gaps. We have developed a generalized value-added model (which we refer to as a differential effects value-added model) that captures differences in growth gaps (by student subgroups) across schools, classrooms, and teachers (and over time). Working with our district partners, we have applied this model in Chicago, Milwaukee, and New York. For additional information see Dokumaci and Meyer (2010) and Christian and Meyer (2010).

15 In the model considered in this paper we have included student-level measures of student characteristics, but have not included school-level (or classroom-level) measures of these variables or other school-level variables, for example, the proportion of students in poverty (by school). We discuss the option of including school-variables in a value-added model in Appendix C. See Meyer (1996) and Willms and Raudenbush (1989) for further discussion of this issue.
Some analysts contend that despite the arguments in favor of including student-level demographic variables in a value-added model, including these variables (or even prior student achievement) could lead to reduced achievement expectations for sub-groups with relatively low achievement or achievement growth.\textsuperscript{16} We strongly reject this contention (although we believe that it needs to be fully and appropriately addressed). We suggest that it arises from a failure to recognize that measuring the productivity of schools, classrooms, and teachers is different from setting student achievement expectations (or standards) and measuring whether students have met those expectations. There are two dimensions to this issue, not one. Attainment information can appropriately be used to identify students who do not satisfy standards and thus are in need of additional resources. Value-added information can appropriately be used to measure the productivity of schools attended by both low and high-achieving students. There is nothing conceptually or practically difficult about addressing both dimensions simultaneously. Later in the paper we illustrate this point using data from the Milwaukee Public Schools.

\textit{Value-Added Productivity, Student Achievement Growth, and Student Attainment}

In this section we highlight the differences and connections between value-added productivity, average student achievement growth (gain), and average student achievement (measured prior to and at the end of the school year). Note that average achievement is a measure that is quite similar to percent proficient in that both indicators measure some feature of the level and distribution of student attainment. The connection between these indicators, given the value-added model presented above, is reported in Figure 2.

\begin{figure}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
 \textbf{(1)} & \textbf{(2)} & \textbf{(3)} & \textbf{(4)} \\
\hline
\textbf{Gain:}
\textbf{Average Growth} & ($\lambda - 1$) \textit{(Average Prior Achievement)} & $+$ & \textit{Average Growth Effect of Student Characteristics} & $+$ & \textbf{Value-Added Productivity} \\
\hline
\textbf{Attainment:}
\textbf{Average Posttest} & $\lambda$ \textit{(Average Prior Achievement)} & $+$ & \textit{Average Growth Effect of Student Characteristics} & $+$ & \textbf{Value-Added Productivity} \\
\hline
\end{tabular}
\caption{The Connections Between Value-Added, Gain, and Attainment Indicators}
\end{figure}

The gain (or growth) indicator differs from value-added productivity in two ways. One, it absorbs growth differences across schools due to differences in student characteristics, if any (column (3)). Two, differences in average prior achievement across schools “leak” into average gain, if the coefficient on prior achievement ($\lambda$) does not equal one (column (2)). These are conditions that can empirically be checked (as illustrated later in the paper). In our experience, estimates of the post-on-pre link parameter ($\lambda$) are generally less than one.\textsuperscript{17} In this case average

\textsuperscript{16} See Appendix C for a discussion of the merits of including classroom and school-level variables in value-added models.

\textsuperscript{17} As discussed in a previous footnote, we assume that the parameters of the value-added model have been estimated using an estimation strategy that controls for measurement error in prior achievement. Failure to control for test
gain absorbs a negative fraction of average prior achievement (since the multiplier \((\lambda - 1)\) in the above equation is negative if \((\lambda < 1)\)). The bottom line is that if there are empirically large differences between average gain and value-added indicators, then it is problematic to rely on the gain indicator as a valid measure of school productivity.

Similarly, average post achievement differs from value-added productivity in two ways. As in the case of the gain indicator, average post achievement absorbs growth differences across schools due to differences in student characteristics, if any. Secondly, average post achievement, as expected, absorbs differences across schools in average prior achievement as long as student achievement is a cumulative growth process, in which case: \(\lambda > 0\). Unless average prior achievement and average student characteristics are identical across schools or are perfectly correlated with value-added productivity (all very unlikely circumstances), average post achievement and other attainment indicators are highly inaccurate measures of school performance. Meyer (1996) presents additional evidence on why attainment indicators generally fail as measures of school performance.

**Value-Added Information and Value-Added Reports**

In this section we discuss approaches for reporting value-added information. We begin by considering reports for a single school district and then consider reports that feature comparisons of multiple districts.

**District Value-Added Reports**

Figure 3 is an example of a school report card from the Milwaukee Public Schools (MPS) that provides information on a school’s value-added rating in reading and mathematics defined using two different metrics. The “beat the average” (BTA) metric is equivalent to “relative school productivity,” as defined in Table 1. This indicator is centered around zero so that the average school in the district has a relative value-added rating equal to zero. The indicator is expressed in units that are identical to the units of student achievement measured at the end of the school year (typically scale score units). The “performance tier” metric is a standardized measure of value-added that generally ranges from 0 to 6, with a district mean equal to 3. The tier rating is equal to: \(\text{tier} = 3 + \frac{\text{BTA}}{\text{SD}}\), where \(\text{SD}\) is equal to the standard deviation (corrected for estimation error) of BTA value-added. In other words, the tier metric is equal to a “z-statistic”, with 3 added so that typical values are positive. As an example, a school with a tier rating equal to 4 has a relative value-added rating that is one standard deviation greater than the average school. The tier metric is relatively easy to use and interpret because it does not

---

1. The information reported in this figure applies to all students in grades 3-5, the elementary school grades covered by the Wisconsin state assessment. The value-added ratings are thus the average of grade-level value-added indicators for grades 3-5.

2. If value-added indicators for a given sample of schools (for example, a district or state) are approximately normally distributed, then the distribution of schools would follow a bell-shaped curve. Roughly 68\% of schools would have a value-added tier rating between 2 and 4. Another 27\% would have tier ratings between 1 to 2 and 4 to 5. Finally, 5\% of the schools would have tier ratings less than 1 or greater than 5.
require knowledge of the units in which student achievement is measured, as in the case of the BTA metric.

Note that the BTA and tier ratings can be used to report relative or absolute value-added productivity. As discussed previously, the difference between the two types of indicators is that relative value-added is centered around the same mean year after year (zero in the case of BTA, 3 in the case of tier ratings). Absolute value-added, in contrast, is centered around mean productivity for a given baseline year. Absolute value-added ratings are not restricted to a pre-specified range, but could shift to the right (in the case of an overall increase in educational productivity) or to the left (in the case of an overall decrease in educational productivity). The MPS school report card reports relative value-added ratings for the most recent three school years in both table and chart formats. (We will discuss the decision to report relative, as opposed to absolute, value-added ratings later in this paper.) The precision of the ratings (as captured by conventional confidence intervals) is also reported in the charts. Attainment information – the percentage of students who are proficient or advanced – is reported in the bottom of the table.  

Figure 3. Example of School Report Card with Value-Added and Attainment Information

---

MPS School Report Card

**School Elementary**

**Value-Added Growth Analysis - Elementary School Grades**

<table>
<thead>
<tr>
<th>Mathematics</th>
<th>Year</th>
<th>Best the Average Performance (Mathematics Scale)</th>
<th>Performance Tier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year/Year</td>
<td>2006-07 to 2007-08</td>
<td>5.5</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>2007-08 to 2008-09</td>
<td>5.6</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>2008-09 to 2009-10</td>
<td>5.5</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>2009-10 to 2010-11</td>
<td>5.4</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>2010-11 to 2011-12</td>
<td>5.3</td>
<td>A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reading</th>
<th>Year</th>
<th>Best the Average Performance (Reading Scale)</th>
<th>Performance Tier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2006-07 to 2007-08</td>
<td>5.5</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>2007-08 to 2008-09</td>
<td>5.7</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>2008-09 to 2009-10</td>
<td>5.5</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>2009-10 to 2010-11</td>
<td>5.4</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>2010-11 to 2011-12</td>
<td>5.3</td>
<td>A</td>
</tr>
</tbody>
</table>

**Value-Added and Attainment Data Status Over Seven Years**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Class</th>
<th>2005-06</th>
<th>2006-07</th>
<th>2007-08</th>
<th>2008-09</th>
<th>2009-10</th>
<th>2010-11</th>
<th>2011-12</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>School</td>
<td>2.4</td>
<td>2.7</td>
<td>2.5</td>
<td>2.6</td>
<td>3.1</td>
<td>3.6</td>
<td>4.6</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td>MPS</td>
<td>3.3</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>41%</td>
</tr>
<tr>
<td>Reading</td>
<td>School</td>
<td>2.5</td>
<td>2.6</td>
<td>2.2</td>
<td>2.3</td>
<td>2.1</td>
<td>1.7</td>
<td>2.7</td>
<td>43%</td>
</tr>
<tr>
<td></td>
<td>MPS</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>54%</td>
</tr>
</tbody>
</table>

20 Note that value-added/post-attainment quadrant information is reported in the lower right part of the report card. Schools are assigned to one of four quadrants depending on whether they have value-added ratings and post-attainment ratings that are above or below average for the district. For example, a quadrant I school has value-added and post-attainment ratings that are both above average. Figure 5 provides information on quadrant status for all MPS elementary schools.
The value-added information in Figure 3 describes a school that exhibited substantial growth in mathematics productivity over the past three years, from tier ratings initially close to the average level (or below) to a high of 4.6. In contrast, value-added ratings in reading were consistently below average. The value-added ratings were estimated with sufficient precision that it is possible to reject the null hypothesis that mathematics productivity in the most recent year was equal to the district average. Despite the fact that mathematics value-added was quite high in the most recent year, the percent of students who met the proficiency standard at the end of the school year was only 53%, somewhat greater than the district average of 49% for that year. These two pieces of information tell an important story: this is a school that serves students with low incoming student achievement (relative to established proficiency standards), but has managed to dramatically improved its productivity in mathematics over the past three years. With respect to mathematics achievement, the attainment information signals that it would be appropriate to provide students in this school with additional resources to spur growth in mathematics achievement (for example, summer school or after school instruction or tutoring). On the other hand, the value-added information indicates that low productivity in mathematics is not the source of the achievement deficit. Indeed, this is a school in which policies should be directed at sustaining excellence with respect to mathematics instruction and pushing the quality of mathematics instruction to the next level (a tier level greater than 5). With respect to reading achievement, the evidence suggests a need to substantially improve productivity. The priorities for the school and the district are to replicate the evident turn-around in mathematics instruction to reading.

One of the key characteristics of the school report card depicted in Figure 3 is that it presents data focused on a single school (although the information is reported so that it is clear where a school stands relative to the district average). Below, we consider a second method for reporting value-added and attainment information that provides a more holistic view of productivity across all schools in the district. In Figures 4 and 5 we present two-dimensional value-added/attainment graphs based on information from the Milwaukee Public Schools for mathematics and reading for a single year. Each data point represents value-added and attainment data for a single school. Figure 4 presents information on value-added productivity and incoming (prior) achievement (measured in terms of a proficiency rate). Figure 5 presents information on value-added productivity and end-of-year achievement (again, measured in terms of a proficiency rate).

---

21 The larger dots in Figures 4 and 5 denote the schools that are located in the Bay View cluster of schools in Milwaukee. The smaller dots denote Milwaukee elementary schools that are not in this cluster. The two types of schools are distinguished, in part, to permit district staff to more efficiently utilize this data. Charts of this kind can be included in district and state performance management systems. Additional data features can be incorporated into the graphs, such as the capacity to drill down to obtain additional information on the students enrolled in each school.
The data in Figure 4 convey a powerful message: all students can learn! As indicated in the figure, there are high (and low) value-added schools in Milwaukee serving schools that range widely in average incoming achievement (arrayed from left to right on the graph). The data on the left side of the figure should serve as a “wake-up call” to schools that have given up on the premise that all students can learn. The proof of the proposition is clearly indicated on the graph: schools in the Milwaukee system can generate very high value-added growth in student achievement despite serving students with very low incoming achievement. These “high value-added/low prior attainment” schools are located in the upper left part of the graph. The graph also reveals that some schools that serve high achieving students (the schools on the right side of the graph) have failed to generate high value-added growth. These schools are effectively “coasting,” and may not even be aware of it. Graphs of this type can be used by school, district, and state staff to: (1) identify low and high-performing schools (that is, schools that need to turn around their performance or sustain excellent performance, respectively) and (2) provide concrete evidence that high performance is an attainable, realistic option for all schools. NCLB, with its exclusive focus on student attainment, is poorly equipped to transmit this message and the strong school performance expectations that go with it.
Figure 5 provides a policy-relevant complement to the preceding graph. Whereas Figure 4 displays information on value-added productivity and incoming achievement, Figure 5 displays information on value-added productivity and end-of-year achievement. This graph is useful for guiding resource allocation decisions designed to ensure that all students meet high achievement standards, as required by NCLB. Student achievement in high value-added schools may be unacceptably low if students enter these schools with very low achievement. Student growth could be accelerated for these students by providing additional resources, such as summer school and after school instruction and tutoring. On the other hand, transferring a low-achieving student from a high-performing school to another school in the district (or reconstituting that school) would probably not improve the learning environment for that student, but, rather, worsen it.

In summary, two-dimensional value-added/attainment graphs can provide school and district staff with information that can be credibly used to set high school performance expectations (or standards) and guide efficient allocation of resources to at-risk students.

Value-Added Reports for Multiple Districts: The Power of Statewide Comparisons

All of the reports discussed in the previous section can, of course, be produced using value-added and attainment information derived from a statewide system. There are several advantages to a statewide system. First, a system based on all of the districts in a state includes many more schools than a single-district system and thus many more opportunities to establish
concrete examples of high value-added productivity. In general, the “observable frontier” (that is, the maximum observed level) of value-added productivity tends to increase as the size of a reference group increases. This is important from a practical policy perspective: it is undoubtedly much easier to establish ambitious productivity standards if policy makers can show that the standard has actually been realized by a school, classroom, or teacher in a given reference group.

Second, in a statewide system it is possible to compare districts with respect to two dimensions: average value-added productivity (across all schools and classrooms in the district) and the consistency of value-added productivity across these entities. Consistency could be measured by the standard deviation of value-added productivity or the percentage of entities with value-added productivity greater than a specified standard. Districts with high average value-added productivity and high consistency could be said to have high “quality control.” This kind of information can easily (and fruitfully) be incorporated into district and state performance management systems.

Third, the overall productivity of a large reference group (such as a state system) is likely to change much more slowly over time than the overall productivity of a small reference group. This suggests that comparisons of performance indicators that are not horizontally equated\textsuperscript{22} (more on this below) may be reasonably be interpreted as providing (approximate) evidence of absolute changes over time.\textsuperscript{23} Alternatively, it may be very policy relevant to measure relative performance within a large reference group, even when it is much less informative to do so within a small reference group.

We should note that the arguments that support statewide, as opposed to single district, reference groups, could equally be used to support the utility of multi-state or national reference groups. Comparisons at this scale would presumably be facilitated by the development of state assessment consortia, if a common scale is used to measure achievement for all members of the consortium.

Below we present the estimates that illustrate the important of cross-district comparisons of district mean productivity and the consistency of productivity (as captured by the district standard deviation). To simplify the analysis, we focus on two of the largest districts in Wisconsin, Milwaukee and Madison. We report estimates of relative value-added productivity (in the beat-the-average metric), rather than absolute value-added productivity (as defined previously in the paper). (In a later section, we consider whether it is feasible to report valid absolute value-added measures, given the assessment data that is available in Wisconsin.)

Estimates from the mathematics and reading value-added models are presented in Table 2. The table reports district mean value-added productivity, the standard error of that estimate, and the standard deviation of school productivity within each district. Figures 6 and 7 report the district mean and standard deviation, respectively, of value-added productivity for mathematics. (The results for reading are quite similar.) As indicated in Table 2 and Figure 6, district average

\textsuperscript{22} Horizontally equated test scores are measured on the same scale over time.

\textsuperscript{23} This follows from the fact that relative changes in productivity are identical to absolute changes in productivity if there is no change in average absolute productivity.
productivity is higher in Madison than in Milwaukee at all grades and in both growth years except for Grade 3-4 in Growth Year 1 for mathematics and Grade 7-8 in Growth Year 1 for reading. For example, an average school in Madison contributes 2.77 more points to a student’s mathematics scale score than an average school in the state in Growth Year 1 from Grade 4-5. On the other hand, an average school in Milwaukee contributes 4.79 fewer points to a student’s mathematics scale score than an average school in the state in Growth Year 1 from Grade 4-5. In contrast, the standard deviation of school productivity is generally much lower in Madison than in Milwaukee. In short, Madison is a more consistent provider of school productivity than Milwaukee.

Table 2. District Value-Added Effects: Madison and Milwaukee

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>District Average</td>
<td>Standard Error</td>
<td>District Standard Deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.48</td>
<td>0.59</td>
<td>3.56</td>
</tr>
<tr>
<td>3</td>
<td>Madison</td>
<td>2.77</td>
<td>0.59</td>
<td>5.60</td>
</tr>
<tr>
<td>4</td>
<td>Madison</td>
<td>-0.95</td>
<td>0.60</td>
<td>5.85</td>
</tr>
<tr>
<td>5</td>
<td>Madison</td>
<td>0.62</td>
<td>0.50</td>
<td>5.37</td>
</tr>
<tr>
<td>6</td>
<td>Madison</td>
<td>2.53</td>
<td>0.58</td>
<td>3.05</td>
</tr>
<tr>
<td>7</td>
<td>Milwaukee</td>
<td>-0.66</td>
<td>0.40</td>
<td>11.39</td>
</tr>
<tr>
<td>3</td>
<td>Milwaukee</td>
<td>-4.79</td>
<td>0.37</td>
<td>8.53</td>
</tr>
<tr>
<td>4</td>
<td>Milwaukee</td>
<td>-6.22</td>
<td>0.39</td>
<td>10.22</td>
</tr>
<tr>
<td>5</td>
<td>Milwaukee</td>
<td>-2.64</td>
<td>0.29</td>
<td>6.38</td>
</tr>
<tr>
<td>6</td>
<td>Milwaukee</td>
<td>-0.14</td>
<td>0.39</td>
<td>6.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade</th>
<th>District</th>
<th>Reading</th>
<th>Growth Year 1: 2005-2006</th>
<th>Growth Year 2: 2006-2007</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>District Average</td>
<td>Standard Error</td>
<td>District Standard Deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.52</td>
<td>0.61</td>
<td>4.71</td>
</tr>
<tr>
<td>3</td>
<td>Madison</td>
<td>3.36</td>
<td>0.61</td>
<td>3.45</td>
</tr>
<tr>
<td>4</td>
<td>Madison</td>
<td>0.90</td>
<td>0.65</td>
<td>5.33</td>
</tr>
<tr>
<td>5</td>
<td>Madison</td>
<td>1.01</td>
<td>0.63</td>
<td>5.75</td>
</tr>
<tr>
<td>6</td>
<td>Madison</td>
<td>1.35</td>
<td>0.64</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>Milwaukee</td>
<td>-1.89</td>
<td>0.41</td>
<td>7.69</td>
</tr>
<tr>
<td>3</td>
<td>Milwaukee</td>
<td>-1.99</td>
<td>0.39</td>
<td>8.23</td>
</tr>
<tr>
<td>4</td>
<td>Milwaukee</td>
<td>-4.17</td>
<td>0.42</td>
<td>8.63</td>
</tr>
<tr>
<td>5</td>
<td>Milwaukee</td>
<td>-2.34</td>
<td>0.39</td>
<td>6.09</td>
</tr>
<tr>
<td>6</td>
<td>Milwaukee</td>
<td>1.92</td>
<td>0.42</td>
<td>6.43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>District Average</td>
<td>Standard Error</td>
<td>District Standard Deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.48</td>
<td>0.59</td>
<td>3.56</td>
</tr>
<tr>
<td>3</td>
<td>Madison</td>
<td>2.77</td>
<td>0.59</td>
<td>5.60</td>
</tr>
<tr>
<td>4</td>
<td>Milwaukee</td>
<td>-0.95</td>
<td>0.60</td>
<td>5.85</td>
</tr>
<tr>
<td>5</td>
<td>Madison</td>
<td>0.62</td>
<td>0.50</td>
<td>5.37</td>
</tr>
<tr>
<td>6</td>
<td>Madison</td>
<td>2.53</td>
<td>0.58</td>
<td>3.05</td>
</tr>
<tr>
<td>7</td>
<td>Milwaukee</td>
<td>-0.66</td>
<td>0.40</td>
<td>11.39</td>
</tr>
<tr>
<td>3</td>
<td>Milwaukee</td>
<td>-4.79</td>
<td>0.37</td>
<td>8.53</td>
</tr>
<tr>
<td>4</td>
<td>Milwaukee</td>
<td>-6.22</td>
<td>0.39</td>
<td>10.22</td>
</tr>
<tr>
<td>5</td>
<td>Milwaukee</td>
<td>-2.64</td>
<td>0.29</td>
<td>6.38</td>
</tr>
<tr>
<td>6</td>
<td>Milwaukee</td>
<td>-0.14</td>
<td>0.39</td>
<td>6.53</td>
</tr>
</tbody>
</table>
Figure 6. Average District Value-Added Productivity in Mathematics for Madison and Milwaukee by Grade and Year

Figure 7. District Standard Deviation of Value-Added Productivity in Mathematics for Madison and Milwaukee by Grade and Year
In order to understand the interplay of differences between the two districts in the mean and consistency of productivity, it is useful to directly examine the distributions of estimated school productivity for each district. Figures 8 reports these distributions for school productivity in mathematics for Madison (lower data plot) and Milwaukee (upper data plot) for the 2006-2007 school year. A separate figure is presented for each grade. The left and right edges of the boxes in the graphs represent the 25th and 75th percentiles of the school productivity distribution in each district. The white dots represent the district averages and the black dots represent individual school productivity effects.

The results are quite striking: the range of value-added productivity is much wider for Milwaukee than Madison (consistent with the results reported in Table 2 that the standard deviation of productivity is higher in Milwaukee than in Madison). Given the wider range of productivity in Milwaukee, it is apparent that the district has schools that are very low performing and schools that are very high performing. In fact, despite the fact that average productivity is higher in Madison than in Milwaukee, the highest performing Milwaukee schools tend to have somewhat higher productivity than the highest performing Madison schools. These data suggest that the Milwaukee Public Schools needs to aggressively work to improve the performance of its lowest performing schools. This district has a significant quality control problem.

Figure 8. Distribution of Value-Added School Productivity in Mathematics in Madison and Milwaukee, 2006-2007 School Year

3rd to 4th Grade (Nov 06 – Nov 07) Mathematics – State VA Model School Effects

School/District VA Productivity Parameters in WKCE Scale Score Units (Relative to State)
4th to 5th Grade (Nov 06 – Nov 07) Mathematics – State VA Model School Effects

MPS School Effects

MMSD School Effects

School/District VA Productivity Parameters in WKCE Scale Score Units (Relative to State)

5th to 6th Grade (Nov 06 – Nov 07) Mathematics – State VA Model School Effects

MPS School Effects

MMSD School Effects

School/District VA Productivity Parameters in WKCE Scale Score Units (Relative to State)
We conclude this part of the paper with several observations. First, the current practice of using only status and/or simple growth models to analyze student growth limits a school or district’s ability to attribute changes in student performance to specific programs or to the instruction provided by teachers. From a public policy perspective, value-added models provide the necessary statistical groundings to more accurately determine the causal reasons behind low and high performance and thus can provide the information needed to identify the most effective school improvement strategies. Second, two-dimensional value-added/attainment graphs can provide school and district staff with information that can be credibly used to set high performance expectations (or standards) and guide efficient allocation of resources to at-risk students. Third, value-added information based on statewide or multi-state data facilitates setting performance expectations, derived from actual data, at high levels. Finally, consistency is an important attribute of school district productivity and should be monitored along with average productivity.

In the remainder of this paper we discuss issues related to the design of value-added models and the design and properties of student assessments.

**How Complex Should a Value-Added Model Be?**

In our work at the Value-Added Research Center, we have followed the following rule in designing and implementing value-added models: “simpler is better, unless it is wrong.” This rule implies that designers of value-added models need to be vigilant in protecting against possible threats to validity and reliability. In practice, this means considering possible model extensions and generalizations and engaging in rigorous diagnostic evaluation and testing of these model enhancements using actual (and possibly simulated) data. Some model enhancements (in some district and state contexts, but perhaps not others) may improve model validity and/or reliability. Others may turn out to have a limited impact on results and thus can either be dropped from the model or retained to demonstrate that the model is robust to inclusion/exclusion of these enhancements. In our work with Chicago, Milwaukee, and New York City, educational stakeholders have been very helpful in identifying possible threats to model validity. For example, a principal in the Chicago Public Schools stimulated a line of inquiry that ultimately led to collecting data on whether a student was homeless and including this student characteristic in the Chicago value-added model.

Although it is beyond the scope of this paper to provide an extensive discussion of important value-added model features, beyond those features included in the two-period model presented above, we list below three features that we have found to be important in our work with districts and states.

*Multiple-year longitudinal test data.* In a two-period value-added model differences across schools in student growth trajectories are captured directly by the student characteristics that are included in the model. Systematic differences in student growth trajectories that are not captured by student characteristics included in the model are absorbed by the estimated value-added effects (thus resulting in bias). One of the key advantages of including three or more

---

24 A model enhancement that does not significantly improve the validity of model estimates may actually reduce statistical precision. In this case, it is generally optimal to drop the enhancement.
achievement outcomes for each student (when that data is available) is that it is possible to better control for differences across schools in the student-level determinants of achievement growth than in a model based on two achievement outcomes. The advantage of having multiple years of longitudinal test data is much greater in classroom and teacher value-added models than in grade-level/school value-added models because the benefits of having multiple years of data depend heavily on the degree of student mobility from unit to unit. (Classroom-to-classroom mobility tends to be much higher than school-to-school mobility.) See the following for discussions of alternative value-added models: Ballou et al (2004); Boardman and Murnane (1979); Hanushek (2005); McCaffrey et al (2004); Meyer (1996, 2004); Rothstein (2007); Sanders and Horn (1994); and Wills and Raudenbush (1989).

Student mobility. Thus far we have not discussed how to measure value-added productivity in real-world situations where some (perhaps many) students change schools during the school year. It is customary when calculating NCLB proficiency rates to exclude students who have been enrolled in a school for a full academic year. Gain (growth) indicators typically either exclude mobile students or pretend that growth can be fully attributed to the school that a student attended at a given point in time (for example, the date of the posttest). We strongly believe that it is problematic from a policy perspective to systematically exclude students from a measurement system that serves an evaluation and accountability function. Systematic exclusion of mobile students (or any other student group) from an accountability system creates an incentive for “agents” to allocate fewer resources to this group. Creating an incentive of this type is a bad idea even if we believe (as we do) that few educators would respond to this incentive.

It turns out that student mobility can be introduced into a value-added model with only a slight tweak in the traditional definition of the school variables included in a standard multilevel (hierarchical) model. In such a model a school variable is set to one if a student ($i$) attended a given school ($k$) during the school year, zero otherwise. In order to accommodate students who changed schools during the school year, the school variable is redefined so that it measures the fraction of time that student attended a given school during the school year. We refer to this variant of the value-added model as the “dose model.” Although this model requires more extensive student attendance data than a model that ignores student mobility, we have found that most districts have data warehouses that can support implementation of this model.

Differential effect value-added. The conventional value-added model (including the model discussed thus far) imposes the restriction that a high-performing classroom or school (at a given grade level at a given point in time) is identically high-performing for all types of students, including, for example, students with low and high prior achievement, and low and high income status. If this assumption is approximately true, schools can be validly compared on the basis of a single performance indicator. However, this assumption might be incorrect: a given school could be very effective for students with low prior achievement, for example, but less so with talented and gifted (TAG) students. These differences in effectiveness could stem from differences in the effectiveness of the multiple programs and courses offered by schools. For example, schools that provide tutoring, after school, and summer school programs for low-performing students, but no additional programs for TAG students, might be relatively more effective with low achieving students than high achieving students.
We have developed a generalized value-added model (which we refer to as a differential effects value-added model) that captures differences in value-added productivity (by student subgroups) across schools, classrooms, and teachers (and over time). Working with our district partners, we have applied this model in Chicago, Milwaukee, and New York. For additional information see Dokumaci and Meyer (2010).

The Intersection of Assessment Design and the Design and Interpretation of Value-Added Models

In this section we consider several aspects of the design and properties of student assessments, with particular focus on the interaction between the characteristics of assessments and the design and interpretation of value-added models. The analysis presented below draws on student assessment data from the state of Wisconsin.

Assessment Scales

Student achievement can be expressed using many different test scales, including development scale scores, percentile scores, and normal curve equivalents.\(^{25}\) Many test developers use a developmental scale as the fundamental scale for building student test scores and then construct other scales and statistics from it. In this paper we report results using development scale scores.

Conventional value-added indicators, as well as average gain and average achievement indicators, exploit the full range of information that is contained within student scale scores. As a result, it is important that scale scores are well measured along the entire distribution of scores—from the lowest to the highest scores. As is well known, assessments that fail to include a sufficient number of relatively easy or relatively difficult test items generally fail to measure with high precision low and high levels of achievement. In fact, excessively easy or difficult assessments may exhibit “floors” and “ceilings,” that is, minimum and maximum test scores that are well within the true distribution of scores. More generally, in order to obtain precise measurements along all parts of the scale, it is best to construct tests so that all parts of the scale are covered by test items of a given difficulty.\(^{26}\)

In sharp contrast to value-added indicators, proficiency rates (the indicators required by NCLB) require accurate measurement only at the “cut” scores that defined the boundaries between proficiency categories (for example, basic, minimal, proficient, and advanced). We speculate that many state assessments have been designed to support accurate measurement at these cut points rather than accurate measurement along the entire achievement spectrum. It is important that the next round of assessments be developed to fully support value-added and growth analyses of the type discussed in this paper.

\(^{25}\) For an extensive review of assessment design and test scaling procedures, see CITATION. See also Wilson (20xx).

\(^{26}\) It is challenging to precisely measure achievement over a wide range of scores using only a single test form. Such a test would need to include a relatively large number of items. Alternatively, some vendors, such as the Northwest Evaluation Association (NWEA), developers of the MAP assessment, use computer adaptive testing (CAT) to select an optimal set of items for every student, thereby increasing the precision of measured achievement and dramatically reducing the number of test items needed to assess a given individual. Add CAT CITATION.
Many, but not all, states have developed test scales that are intended to be comparable from one grade to the next. Assessments of this type are said to be vertically scaled. Scale scores based on different forms of an assessment (designed for a given grade or achievement level) are said to be horizontally scaled if they are scored on the same developmental scale. All state tests are required to be horizontally equated so that it is possible to determine whether school and state proficiency rates have changed over time.

Since it is costly to build assessment systems in which test scores are horizontally and possibly vertically scaled, it is important to be clear about the benefits of building assessments that satisfy these properties. In order to measure student gain (growth) it is, of course, essential that prior and post achievement (more generally, achievement at multiple points in time) be measured on the same (vertical) scale. However, as discussed in the next section, it is not clear that data users should unquestioningly accept test developer claims that test scores from different grades are successfully vertically equated. An important strength of value-added models is that they do not require assessments to be vertically scaled if they include a post-on-pre link parameter, as discussed above.\(^{27,28}\)

The value of horizontal equating is potentially much higher and is, in fact, a requirement of NCLB-required assessments. In short, if assessments are properly horizontally equated (and built to accurately measure achievement along the entire continuum of scores), then it is feasible to measure changes in student attainment and value-added over time. In the context of the value-added model presented earlier, the overall year-specific productivity parameter (\(\pi\)) and absolute value-added (\(\eta^{\text{ABSOLUTE}}\)) can be estimated only if assessments are properly horizontally equated. If assessments are not horizontally equated, then the best that can be accomplished is to compare value-added productivity relative to average productivity (given by the parameter \(\eta\)) over all of the teachers, classrooms, and schools included in the system – the reference group. Since the

\(^{27}\) Another alternative, sometimes used in practice, is to essentially discard the developmental scale produced by a test vendor and follow one of two options: (1) Linearly transform all scale scores, separately by grade and possibly test administration date, so that the transformed test scores have a pre-specified mean and standard deviation (often zero and one, respectively, as in a z-statistic); (2) Nonlinearly transform all scale scores, separately by grade and possibly test administration date, so that the transformed scores conform (approximately) to a normal distribution with pre-specified mean and variance. The normal curve equivalent (NCE) scale implements the second strategy. Unfortunately, neither approach produces test scales that are vertically equated. In some subject/skill areas, an accurately vertically equated assessment could naturally exhibit expanding or declining test score variability across grade levels. Forcing these tests to have a uniform variance would clearly not produce vertically scaled test scores. On the other hand, either of these two approaches could be helpful to provide some standardized meaning to test scores that are clearly not measured on same scale (perhaps produced by different vendors). Note that transforming test scores at a given level separately by years (or test administration dates) should generally be done only if there is reason to suspect that test scores from different years have not been successfully equated. The bottom line is that test scores produced using these methods should not be treated as vertically equated scores and thus should probably not be used to measure gain (growth). They can be used in a value-added model that includes a post-on-pre link parameter.

\(^{28}\) There is an important exception to this rule in the case of short-cycle assessments – that is, assessments that are given multiple times during the school year (for example, in September, December, March, and May). If different students and different schools take/administer short-cycle assessments at substantially different times during the school year, then it is important that the assessments administered on different days be correctly equated and reported on the same scale. Short-cycle assessments such as the NWEA MAP meet this criterion.
overall productivity of a large reference group (such as a state system) is likely to change much more slowly over time than the overall productivity for a small reference group, relative productivity indicators may be reasonably be interpreted as providing (approximate) evidence of absolute changes over time. Hence, relative indicators based on large reference groups may be quite useful for schools and districts (since they represent only a small part of an entire reference group). They obviously are useless for tracking the overall performance of the reference group (since the overall productivity average for the reference group always equal to zero). In our experience, local stakeholders and policy makers universally prefer absolute, rather than relative, comparisons of productivity. They prefer the idea of a fixed performance standard, rather than one that moves up or down depending on the performance of other entities. Moreover, it is important from a national perspective to be able to track the absolute productivity of states and the nation as a whole.

Despite the obvious value of ensuring that assessments used for accountability purposes are horizontally equated, in our experience many assessments do not appear to be properly horizontally equated, particularly when put to the demanding use of supporting measurement of growth or value-added. We present evidence on this later in the paper. One of the problems is that state assessments systems may currently be constructed so that horizontal equating errors for a single test are relatively small, since a single test is all that is required to compute a proficiency rate. In contrast, four different tests are required to determine whether a state’s value-added productivity increased or decreased (a pretest and posttest for two different cohorts). As a result, value-added indicators could, in some cases, be subject to unacceptably large horizontal equating errors. We strongly recommend that states (and state assessment consortia) require test developers to ensure that assessments satisfy clearly specified tolerances for horizontal equating error.

Below, we investigate assessment data from the state of Wisconsin to determine whether the state tests have been properly horizontally equated. We also consider a variety of statistics to determine whether the distributions of tests scores are comparable over time. These are statistics that could routinely be used to evaluate the degree to which assessments have been successfully horizontally equated and designed to accurately measure achievement over a wide spectrum.

**Stability in the Distribution of Test Scores**

The analysis presented below draws on student assessment data from the state of Wisconsin for the schools years 2005-2006 and 2006-2007, and 2007-2008. The state assessment, the Wisconsin Knowledge and Concepts Examination (WKCE), is administered to all students in grades 3-8 and 10 in November of each year. We used this data to estimate value-added models of growth in reading and mathematics achievement for grades 3 to 4, 4 to 5, 5 to 6, 6 to 7, and 7 to 8. There were approximately 55,000 to 60,000 students in each grade and a total of 425 school districts.

---

29 An accountability/reward system that is entirely based on relative performance is akin to a tournament, where the outcomes depend on the performance of all participants. See Lazear and Rosen (1981) and Lazear (1995).

30 Although this case study draws on data from a particular state to illuminate issues on building and interpreting value-added models, we should emphasize that the results presented in this paper are very similar to results we have obtained working with other district and state data bases.
We begin our analysis by examining the means and standard deviations of Wisconsin test scores by grade for four points in time, November 2005, November 2006, and November 2007, and November 2008. Students were included in the analysis if they had both a pretest score and posttest score in two consecutive years. We refer to these samples as matched samples. Similar results were obtained for unmatched samples (samples that included all students.) Table 3 includes the number of students (N), the state means of pretest and posttest scale scores (MEAN), and the standard deviations of pretest and posttest scale scores (STD) for each year, grade and subject.

Table 3. Summary Statistics for Wisconsin State Assessment Data, Matched Samples

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Nov 05 - Nov 06</th>
<th>Nov 06 - Nov 07</th>
<th>Nov 07 - Nov 08</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td>Math</td>
<td>3</td>
<td>4</td>
<td>54079</td>
<td>432.47</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>55879</td>
<td>463.90</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
<td>56099</td>
<td>485.09</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>58971</td>
<td>508.63</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>8</td>
<td>60779</td>
<td>529.05</td>
</tr>
<tr>
<td>Reading</td>
<td>3</td>
<td>4</td>
<td>54079</td>
<td>458.83</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>55879</td>
<td>477.66</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
<td>56099</td>
<td>485.80</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>58971</td>
<td>501.99</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>8</td>
<td>60779</td>
<td>511.95</td>
</tr>
</tbody>
</table>
Figure 9. Growth in Average Mathematics Scale Scores, Wisconsin

Mean Scale Score by Grade (Mathematics)

Figure 10. Growth in Average Reading Scale Scores, Wisconsin

Mean Scale Score by Grade (Reading)
The information in Table 3 is presented graphically in Figures 9-12. In each figure a line represents a different student cohort. For example, in Figure 9, the line for grades 3-4-5-6 tracks the average mathematics scale score of students in grade 3 in 2005, grade 4 in 2006, grade 5 in 2007, and grade 6 in 2008 for the same student cohort. As indicated in Figure 9, the average mathematics score increases with grade, although with a declining rate; the increase in the average mathematics score is less for higher grades. In addition, there is substantial variation in mathematics scores at a given grade level. In contrast, average reading scores exhibit relatively rapid growth in the even-numbered grades and relatively slow growth in odd-numbered grades. In addition, there is very little variation in reading scores at a given grade level.

Figures 11 and 12 display the standard deviations of test scores across grades and years. As in the previous two graphs, each line represents a different cohort. Note that the standard deviations of mathematics scores vary substantially across grades (about 0 to 6 scale score units) and cohorts (about 2 to 6 scale score units). In contrast, the standard deviations of reading scores vary substantially across grades (about 1 to 8 scale score units) but not much across cohorts (about 1 to 2 scale score units).

The bottom line is that there is substantial inconsistency across grades, years, and cohorts in the means and standard deviations of Wisconsin test scores. Taking the data at face value, it would appear that students in Wisconsin learn substantially less in reading in 4th and 6th grade than they do in other grades. The range of mathematics scores, as measured by the standard deviation, floats up and down over time for the same cohort of students. This calls into question whether it is appropriate to view Wisconsin test scores, measured at different grades and different years, as scale scores that are properly vertically and horizontally scaled.
Figure 11. Standard Deviations of Mathematics Scores across Grades, Wisconsin

**Standard Deviation of Scale Scores by Grade (Mathematics)**

Figure 12. Standard Deviations of Reading Scores across Grades, Wisconsin

**Standard Deviation of Scale Scores by Grade (Reading)**
Stability in Achievement Growth

In the first part of the paper we concluded that instability in the variability of test scores, as documented in the previous section, could substantially affect the relationship between prior achievement and achievement at the end of the school year, as captured by the post-on-pre link parameter (λ). This section reports on estimates of the post-on-pre link parameter for two sets of models, models based on the original scale scores and models based on scale scores that have been transformed to have constant variance at all grade levels and in all years. Figures 13 and 14 report estimates of this parameter for both models in mathematics and reading for two growth years.

As indicated in Figure 13, the estimates for mathematics based on the original test scales vary widely across grades (from a low of 0.81 to a high of 1.08) and vary widely across the two growth years for the fourth grade model. The estimates for mathematics based on the transformed (standardized) test scores lie within a narrow band ranging from 0.85 to 0.91. The estimates tend to be smaller in the early grades (approximately 0.86) and larger in the later grades (approximately 0.90). The estimates for reading based on the original test scales also vary widely across grades, with exceptional high values (close to 1 or greater) in grades 3 and 7 and consistently lower in grades 4 to 6. As in the case for mathematics, the estimates for reading based on the transformed (standardized) test scores lie within a narrow band and are very similar at all grade levels.

This analysis in this section reinforces the concerns raised above that the units of the Wisconsin scale scores are unstable from grade to grade and year to year. We note that a value-added model with a post-on-pre link parameter provides built-in protection against instability in the variability of prior and end-of-year test scores. Alternatively, the vendor-developed test scales can be essentially discarded and replaced by scale scores that have been transformed to have constant variance at all grade levels and in all years. We find consistent evidence that the post-on-pre link parameters in the models with variance-standardized test scores are substantially less than one, a parameter restriction that is imposed in some value-added models.
Figure 13. Post-on-Pre Link Coefficients on Mathematics

Figure 14. Post-on-Pre Link Coefficients on Reading
Comparability of Value-Added Estimates over Time: A Test of Horizontal Equating

In this section we consider whether it is appropriate to compare value-added estimates from different cohorts, given the pattern of results presented in the previous section. It is important to assess whether this data supports valid comparisons of growth over time. Unreasonably large increases or decreases in average statewide productivity might indicate problems in maintaining the comparability of test scores over time (that is, the accuracy of horizontal equating across test forms). Errors in horizontal equating of test scores over time would make it illegitimate to interpret a change in an estimated state productivity effect as a genuine change in productivity. In such a case, school value-added indicators could only be used to compare the productivity of schools relative to other schools in the same year. Relative value-added indicators may reasonably be compared over time if we believe that the true change in state average value-added is relatively small. In that case a reported change in the relative value added would be approximately equal to the change in absolute value-added.

Since test vendors generally do not provide direct evidence on the degree to which test scores are successfully equated, we believe that analysts should exercise caution in accepting unreasonably large increases or decreases in productivity as evidence of true changes in productivity. We propose the following subjective rules for detecting possibly erroneous changes in productivity for given schools and for the state as a whole.

- School-level rule: Consider a change in productivity of greater than one standard deviation (also referred to as a tier unit) in value-added productivity (corrected for estimation error) as possible evidence of faulty horizontal equating.
- State-level rule: Consider a change in statewide productivity of greater than 0.5 standard deviation units (also referred to as tier units) as possible evidence of faulty horizontal equating.

Based on our experiences, we suspect that horizontal equating errors are common for statewide tests not originally designed to support growth and value-added analysis.

Below we examine changes in statewide test scores over time to detect whether the changes are consistent with the suggested statewide rule. As indicated in Appendix B, an approximate estimate of the difference in productivity between two growth years is given by the change in average gain between the two years at a given grade level. These numbers can be computed quite readily.

The average gain in Wisconsin scale scores for each growth year is reported in the first three columns of Table 4. These numbers are taken directly from the attainment statistics presented earlier in Table 3. The change in growth, reported in columns 4-6, is computed from the data in columns 1-3. Column 7 reports the value-added tier unit (the standard deviation in value-added school effects) for the baseline year: 2005-2006. (The value-added standard deviations computed in other years are nearly identical to the reported baseline standard deviations.) The change over time in average gain, reported in tier units, is presented in the bottom panel of the table for each grade. These numbers are equal to the change in growth divided by the value-added tier unit.

30
As indicated in the bottom panel of Table 4, the change in average gain in mathematics achievement exceeded the specified cutoff of 0.5 tier units in almost all grades. Indeed, most of the changes in mathematics were substantially greater than a full tier unit. The change in average mathematics gain was especially large in grades 6 and 7, substantially in excess of a full tier unit. In fact, in seventh grade, the average gain in achievement declined from approximately 15.1 scale score points in Growth Year 1, to 6.5 points in Growth Year 2, and then back up to 12.9 points in Growth Year 3. In short, this data implies that it would be unwise to view the absolute value-added productivity estimates in mathematics as comparable based on the current Wisconsin state assessment. As a result, we conclude that the Wisconsin mathematics assessment data only supports comparisons based on relative value-added indicators.

In sharp contrast to the results for mathematics, most (but not all) of the changes in gain in reading are less than the cutoff of 0.5 tier units. As a result, it may be reasonable to view these numbers as being comparable between the three growth years.

We should caution the reader that the rule used in this section to identify failed horizontal equating is designed to pick up only the most egregious violations of successful horizontal equating. Horizontal equating errors that are less than 0.5 tier points will not, of course, be detected, but could still be much larger than is acceptable. This is an important area for further analysis.
### Table 4. Comparison of Average Gain in Achievement Over Time

#### Average Gain, Change in Average Gain

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Average Gain in Pre</th>
<th>Growth Year 1</th>
<th>Growth Year 2</th>
<th>Growth Year 3</th>
<th>Change in Avg Gain from G1 to G2</th>
<th>G2 to G3</th>
<th>G1 to G3</th>
<th>VA Tier Unit in G. Year 1</th>
<th>VA Tier Unit in G. Year 2</th>
<th>VA Tier Unit in G. Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>3</td>
<td>35.79</td>
<td>32.07</td>
<td>40.23</td>
<td>-3.72</td>
<td>8.17</td>
<td>4.45</td>
<td>6.75</td>
<td>7.10</td>
<td>7.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>26.95</td>
<td>27.08</td>
<td>29.20</td>
<td>0.13</td>
<td>2.12</td>
<td>2.25</td>
<td>6.93</td>
<td>7.66</td>
<td>6.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>29.58</td>
<td>24.45</td>
<td>21.02</td>
<td>-5.22</td>
<td>-3.43</td>
<td>-8.56</td>
<td>7.20</td>
<td>7.06</td>
<td>6.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>27.64</td>
<td>19.97</td>
<td>21.80</td>
<td>-7.66</td>
<td>1.82</td>
<td>-5.84</td>
<td>5.42</td>
<td>4.83</td>
<td>5.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>15.10</td>
<td>12.84</td>
<td>12.84</td>
<td>-8.57</td>
<td>6.42</td>
<td>-2.15</td>
<td>5.85</td>
<td>5.06</td>
<td>5.25</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>3</td>
<td>21.20</td>
<td>18.02</td>
<td>19.81</td>
<td>-2.20</td>
<td>1.78</td>
<td>-0.41</td>
<td>4.71</td>
<td>5.19</td>
<td>5.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8.56</td>
<td>7.72</td>
<td>5.50</td>
<td>-0.84</td>
<td>-2.21</td>
<td>-3.06</td>
<td>4.72</td>
<td>4.72</td>
<td>4.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>16.81</td>
<td>18.48</td>
<td>18.82</td>
<td>-0.34</td>
<td>0.34</td>
<td>0.00</td>
<td>5.25</td>
<td>5.34</td>
<td>5.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>12.21</td>
<td>10.62</td>
<td>13.53</td>
<td>-1.59</td>
<td>2.91</td>
<td>1.32</td>
<td>4.32</td>
<td>3.98</td>
<td>3.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>16.29</td>
<td>14.81</td>
<td>13.94</td>
<td>-1.48</td>
<td>-0.97</td>
<td>-2.45</td>
<td>4.35</td>
<td>3.97</td>
<td>4.55</td>
<td></td>
</tr>
</tbody>
</table>

#### Average Gain Changes in Tier Units

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>G1toG2 Avg Gain Pre</th>
<th>G1toG2 Avg Gain Post</th>
<th>G2toG3 Avg Gain Pre</th>
<th>G2toG3 Avg Gain Post</th>
<th>G1toG3 Avg Gain Pre</th>
<th>G1toG3 Avg Gain Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>3</td>
<td>-0.55</td>
<td>-0.52</td>
<td>1.15</td>
<td>1.17</td>
<td>0.66</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-0.71</td>
<td>-0.73</td>
<td>-0.49</td>
<td>-0.52</td>
<td>-1.19</td>
<td>-1.30</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-1.41</td>
<td>-1.59</td>
<td>0.38</td>
<td>0.34</td>
<td>-1.08</td>
<td>-1.09</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>-1.46</td>
<td>-1.69</td>
<td>1.27</td>
<td>1.22</td>
<td>-0.37</td>
<td>-0.41</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>-0.47</td>
<td>-0.42</td>
<td>0.34</td>
<td>0.33</td>
<td>-0.09</td>
<td>-0.08</td>
</tr>
<tr>
<td>Reading</td>
<td>3</td>
<td>-0.18</td>
<td>-0.18</td>
<td>-0.47</td>
<td>-0.46</td>
<td>-0.05</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-0.06</td>
<td>-0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-0.37</td>
<td>-0.40</td>
<td>0.73</td>
<td>0.78</td>
<td>0.30</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>-0.34</td>
<td>-0.37</td>
<td>-0.24</td>
<td>-0.21</td>
<td>-0.56</td>
<td>-0.54</td>
</tr>
</tbody>
</table>

Note: As a rough rule of thumb, year-to-year changes in average gain that exceed 0.5 tier (VA standard deviation) units indicate possible test form effects and thus may not represent genuine changes in average state productivity. Grades in which average gain exceeds this threshold are shaded in the tables.

Note to publisher: We may not be able to use color in the print version. How about in the web version?

**Test Measurement Error**

Substantial test measurement error exists in virtually all assessments that are not adaptive. In our experience the average reliabilities of NCLB assessments tend to be around 85%, although the magnitude of error (as typically measured by the standard error of measurement) tends to vary widely across individuals. Errors are lower for individuals with achievement levels that are closely matched to the difficulty of the items included on the test. Adaptive assessments tend to have much lower levels of measurement error, with average test
reliabilities often in excess of 95%.\textsuperscript{31} This stems from the fact that in adaptive tests, test-takers are directed toward test questions that closely match the achievement levels of the individuals. For more information on computer adaptive testing see CITATION (ETS book).

Test measurement error is problematic in our context for the following reasons. First, in order to properly estimate value-added models of the type presented in this paper, it is necessary to account for measurement error in prior achievement (using methods from structural equation modeling (SEM)). Fuller (1987) and Meyer (1992, 1999) discuss techniques for correcting for measurement error in linear models. These techniques are straightforward to use and have been used to control for test measurement error in all of the value-added systems developed at VARC.\textsuperscript{32}

Second, although measurement error correction techniques are a standard statistical tool, using this tool in the value-added context makes the model more complex and difficult to understand. Indeed, it is baffling to most people that a simple graph of student gain on prior student achievement yields substantially biased results. The existence of substantial test measurement error typically produces the following: students with low prior achievement have large positive test score gains and students with high prior achievement have large negative test score gains. Simple graphical evidence of this type appears to support the conclusion that a school, classroom, or teacher, has been very successful at remediation – raising the achievement of low achieving students, but quite unsuccessful at the raising the achievement of gifted and talented students. In general, results of this type are entirely due to test measurement error.

Finally, test measurement error is a primary culprit for producing low precision value-added estimates when sample sizes are small, for example, data for a teacher that has only taught for only one year. Given the increasing interest in producing value-added estimates at the most disaggregate level (where sample sizes are small), it would be advantageous to favor assessment designs that produce tests with the least amount of error.

Since test measurement error is problematic from the standpoint of conducting formal and informal value-added and growth analyses, we strongly recommend that the next generation of assessments exploit adaptive testing methods to the extent feasible and appropriate.\textsuperscript{33}

Conclusions

The design of value-added models and the design of assessments to support these models is a new area of research that is technically challenging and very important from a policy perspective. In this paper, we have presented a simple value-added framework that both illustrates what can be learned from a carefully design value-added model of educational productivity and provides a framework for discussing the interaction between value-added model

\textsuperscript{31} We have found that test reliabilities for the NWEA MAP assessment, a computer adaptive test, generally exceed 95% or higher.

\textsuperscript{32} Measurement error correction techniques are much more complicated to implement in nonlinear models. VARC is currently working with several districts to develop and implement these techniques.

\textsuperscript{33} Many of the papers included in this volume discuss alternative assessment delivery systems, including systems that employ adaptive testing.
design and the design and properties of student assessments. Our overall message is that value-added models place new demands on the quality and robustness of student assessments. The next generation of student assessments needs to be designed so as to fully support the development of valid and reliable value-added models of student achievement and educational productivity.

One important new area for research and development is the development of new value-added models in subjects and grades other than those covered by NCLB. To make progress in this area, it may be necessary to develop new assessments that are tightly connected to the curriculum being taught in these courses.
Appendix A. Technical Description of Two-Period Value-Added Model

In this appendix we provide a technical description of the simple statewide (multi-district) value-added model of school productivity discussed in the text. Most, if not all, value-added models (including classroom and teacher value-added models) produce value-added parameters of the type included in this model. In the text we discuss options for generalizing the model to allow for multiple longitudinal observations of student test scores, measurement error in test scores, and other factors.

The key features of the model are:

- Two years of (consecutive grade) longitudinal assessment data for each student (measured annually at the end or beginning of the school year).\(^{34}\)
- School/district value-added productivity effects \(\eta_{kl}\) (for school \(k\) in district \(l\) in year \(t\), at a given grade).
- Statewide value-added productivity effects \(\pi_t\).
- A posttest on pretest “link” parameter \(\lambda_t\) (which may vary across grades and over time).

This parameter allows for the possibility that achievement growth may differ for students with high and low prior achievement and for situations where the distribution of the posttest and pretest variables may be non-uniform over time (more on this below).

- Demographic variables \(X_{it}\) to capture differences across students (within-classrooms) in achievement growth.

The two-period value-added model is defined by the following equation:

\[
Y_{2it} = \xi + \lambda_{it} Y_{1it-1} + \pi_t + \beta_t' X_{it} + \sum_k \sum_l \eta_{kl} S_{klt} + \epsilon_{it}
\]

where the variables, parameters, and indices in the model are defined in Appendix Table A.1 and the grade descriptors are omitted, for simplicity.

\(^{34}\) Note that since statewide testing begins in third grade in many states, only two years of (up-to-date) attainment data are typically available to estimate value-added models of achievement growth from third to fourth grade. In later grades, where additional years of longitudinal data are available (except for students with missing test data), it is possible to expand the two-period model to include multiple grades, for example, a model of achievement growth from third to fourth to fifth grade. In a two-period model differences across schools in student growth trajectories are captured directly by the student characteristics that are included in the model. In this model, systematic differences in student growth trajectories that are not captured by student characteristics included in the model are absorbed by the estimated value-added effects. One of the key advantages of including three or more achievement outcomes for each student (when that data is available) is that it is possible to better control for differences across schools in the student-level determinants of achievement growth than in a model based on two achievement outcomes. See the following for discussions of alternative value-added models: Ballou et al (2004); Boardman and Murnane (1979); Hanushek (2005); McCaffrey et al (2004); Meyer (1996, 2006, 2007); Sanders and Horn (1994); and Willms and Raudenbush (1989).
Table A.1. Variables and Parameters in Value-Added Model

<table>
<thead>
<tr>
<th>$i$</th>
<th>Student identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>Within-district school identifier</td>
</tr>
<tr>
<td>$l$</td>
<td>District identifier</td>
</tr>
<tr>
<td>$t$</td>
<td>Year of posttest score</td>
</tr>
<tr>
<td>$g$</td>
<td>Grade (not explicitly included in above model)</td>
</tr>
<tr>
<td>$Y_{2lt}$</td>
<td>Posttest score in year $t$</td>
</tr>
<tr>
<td>$Y_{1lt-1}$</td>
<td>Pretest score in year $(t-1)$ (prior year)</td>
</tr>
<tr>
<td>$X_{il}$</td>
<td>Student demographic characteristics (vector)</td>
</tr>
<tr>
<td>$S_{klt}$</td>
<td>Student indicator, or fractional measure of enrollment, in school $k$, in district $l$, in year $t$</td>
</tr>
<tr>
<td>$\lambda_t$</td>
<td>Coefficient on pretest score: posttest-on-pretest link</td>
</tr>
<tr>
<td>$\beta_t$</td>
<td>Coefficient (vector) for demographic characteristics</td>
</tr>
<tr>
<td>$\xi$</td>
<td>Intercept</td>
</tr>
<tr>
<td>$\varepsilon_{ilt}$</td>
<td>Student level error component</td>
</tr>
</tbody>
</table>

Value-Added Effect Parameters
(Reproduced from text)

| $\pi_t$ | Statewide productivity in year $t$ (for a given grade). Note that this parameter can only be interpreted as a genuine statewide productivity effect if test scores are accurately horizontally equated over time so that changes in test score growth do not reflect test form effects. |
| $\eta_{klt}$ | Relative school and district productivity (hereafter called relative school productivity) for school $k$ in district $l$ in year $t$ (for a given grade). This parameter is referred to as a “relative” value-added parameter because it is centered around zero in each year so that the average school in the district has a value-added rating equal to zero and school productivity is measured relative to the average school. Changes in statewide productivity are thus absorbed by the parameter $\pi_t$. |
| $\eta_{klt}^{\text{ABSOLUTE}}$ | $= \pi_t + \eta_{klt}$ Absolute (total) school, district, and state productivity. This indicator incorporates relative school productivity plus overall changes in statewide productivity, provided (as mentioned above) that test scores are accurately horizontally equated. |

Note that all parameters are allowed to vary by year, including the slope parameters $\lambda_t$ and $\beta_t$. 

36
Appendix B. The Average Change in Statewide Gain is Approximately Equal to the Change in Statewide Value-Added Productivity

Using the value-added model defined in Appendix A, the average statewide gain from year \((t-1)\) to \(t\) at a given grade level is given by:

\[
G_t = \bar{Y}_{2,t} - \bar{Y}_{1,t-1} = \xi + \pi_t + (\lambda_t - 1)\bar{Y}_{1,t-1} + \beta_t'\bar{X}_{t} 
\]  \hspace{1cm} (1)

where the bar over each variable (and the dot replacing the \(i\) index) signifies that the variable is a state mean. The change in statewide gain from posttest year \(s\) to \(t\) is similarly given by:

\[
\Delta_{st} = G_t - G_s = (\pi_t - \pi_s) + (C_t - C_s) 
\]  \hspace{1cm} (2)

where \(C_t\), a cohort variable, is defined as:

\[
C_t = (\lambda_t - 1)\bar{Y}_{1,t-1} + \beta_t'\bar{X}_{t} 
\]  \hspace{1cm} (3)

The cohort variables will typically not change much from year to year so that the change in statewide gain approximately equals the change in statewide productivity, as asserted:

\[
\Delta_{st} = G_t - G_s \approx (\pi_t - \pi_s) 
\]  \hspace{1cm} (4)

The change in statewide gain in tier units is obtained by dividing gain \(\Delta_{st}\) by the standard deviation of school productivity in the baseline year \(\omega\). 

37
Appendix C

Student and School-Level Variables in Value-Added Models

Meyer (1996) and Willms and Raudenbush (1989) discuss some of the conceptual and empirical issues involved in including student and school-level control variables in value-added models, for example, average poverty status. The primary concern with including school-level control variables is that the estimated coefficients on these variables could be substantially biased if school resources and “intrinsic” school productivity are not “assigned” to schools such that school-level control variables are uncorrelated with unobserved school productivity. This condition would be violated, for example, if high-performance teachers and administrators prefer to work in schools with low poverty. In this case, the coefficient on average poverty status would absorb this negative correlation, yielding an estimated coefficient biased in the negative direction.

Estimated value-added indicators would, of course, be biased if the coefficients on school-level control variables are biased. For example, if the coefficient on average poverty status is biased in the negative direction, the estimated value-added performance of schools that disproportionately serve high poverty students would be biased upward. This is problematic if one of the important purposes of a value-added system is to identify schools that are in need of improvement. Due to this concern, model designers need to be cautious about including school-level control variables in value-added models. This is an important area for further research.

Note that the statistical concerns discussed above do not apply to student-level control variables. These coefficients are estimated off of variation within schools or classrooms, for example, the contrast in achievement growth between low and high-poverty students within schools or classrooms. If resources are allocated within schools or classrooms in a way that is systematically related to student characteristics, then the coefficients on student-level control variables will capture these systematic patterns. The other role for student-level control variables is to proxy for the differences in resources provided to students by their families.
Resources on Value-Added Models and Related Literature


Christian, M., and Meyer, R.H.


Harris, D. and Sass, T. "Value-added models and the measurement of teacher quality", Unpublished manuscript, Florida State University, 2006.


Meyer, R. "Value-added indicators of school performance", in Hanushek, E. and Jorgenson, W., ed.,'Improving America's schools: The role of incentives', National Academy Press,


Meyer, R. H.; Dokumaci, E.; Morgan, E.; and Geraghty, E. (February 2009) *Demonstration of a State Value-Added System for Wisconsin*, Report to the Wisconsin Department of Public Instruction, Wisconsin Center for Education Research. Value-Added Research Center, Madison, WI.


Attachment F

Betebenner White Paper on Student Growth Percentiles
Growth, Standards and Accountability

Damian W. Betebenner
The Center for Assessment
DBetebenner@nciea.org
April 6, 2009

Abstract

Over the last decade, large scale annual testing has provided states with unprecedented access to longitudinal student data. Current use of this data focuses primarily upon analyses of growth most often directed toward accountability decisions. Analyses using this longitudinal data source for other purposes have gone largely untapped. This paper introduces analysis techniques and results showing how student growth percentiles, a normative growth analysis technique, can be used to examine the illuminate the relationship between standards based accountability systems and the performance standards on which they are based.

Introduction

The impact of NCLB upon research connecting large scale assessment outcomes with school quality has been profound leading to a proliferation of initiatives centered on standards based accountability. The rapid development of all aspects of standards based accountability systems including the establishment of performance standards, the development of status based accountability systems, and research on growth analysis techniques has produced vast amounts of specialized knowledge promoting a sense that these areas are largely distinct. To the contrary, this paper promotes the view that these disparate topics are actually closely interconnected, needing only a lens to assist stakeholders in viewing the interconnections between the sometimes disparate pieces. This paper supplies that lens.

†Special thanks to the Colorado Department of Education for their generous support for work on the Colorado Growth Model on which this work is based. All opinions are those of the author and are not necessarily those of the Colorado Department of Education. Permission is granted to copy, distribute and/or modify this document and its contents under the terms of the Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License. No commercial or for-profit use of this work is permitted without the express written consent of Damian W. Betebenner. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/us/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.
Current state accountability systems rely heavily upon performance standards to make judgments about the quality of education. Specifically, accountability systems constructed according to federal adequate yearly progress (AYP) requirements use annual “snap-shots” of student achievement relative to state performance standards to make judgments about education quality. Since their adoption, such standards based status measures have been the focus of persistent criticism (Linn, 2003; Linn, Baker, & Betebenner, 2002). Though appropriate for making judgments about the achievement level of students, they are inappropriate for judgments about educational effectiveness. In this regard, status measures are blind to the possibility of low achieving students attending effective schools. It is this possibility that has led some critics of No Child Left Behind (NCLB) to label its accountability provisions as unfair and misguided and to demand the use of growth analyses as a better means of auditing school quality and increasing the validity of accountability systems (Braun, 2008).

A fundamental premise associated with using student growth for school accountability is that “good” schools bring about student growth in excess of that found at “bad” schools. Students attending such schools—commonly referred to as highly effective/ineffective schools—tend to demonstrate extraordinary growth that is causally attributed to the school or teachers instructing the students. The inherent believability of this premise is at the heart of current enthusiasm to incorporate growth into accountability systems. It is not surprising that the November 2005 announcement by Secretary of Education Spellings for the Growth Model Pilot Program (GMPP) permitting states to use growth model results as a means for compliance with NCLB achievement mandates was met with great enthusiasm by states. (Spellings, 2005).

Consistent with current accountability systems that hold schools responsible for the assessment outcomes of their students, the primary thrust of growth analyses over the last decade has been to determine, using sophisticated statistical techniques, the amount of student progress/growth attributable to the school or teacher (Braun, 2005; Rubin, Stuart, & Zanutto, 2004; Ballou, Sanders, & Wright, 2004; Raudenbush, 2004). Such analyses, often called value-added analyses, purport to estimate the teacher/school contribution to student achievement. This contribution, called the school or teacher effect, quantifies the impact on achievement that this school or teacher would have, on average, upon similar students assigned to them for instruction. Clearly, such analyses lend themselves to accountability systems that hold schools or teachers responsible for student achievement.

A weakness of value-added analyses is the difficulty anchoring the normative results with the performance standard criteria on which the accountability system rests. In essence, universal proficiency does not necessarily follow from above average growth (No Child Left Behind versus No Child Left Behind on Average). The gap between value-added analyses and standards based accountability systems is simply a reflection of a deeper schism between underlying norms and standards. This paper puts establishes that reconciling this gap involves looking at both normative and standards based interpretations simultaneously within a unifying normative framework. Doing so illuminates the interconnections between growth, standards, and accountability:

**Growth and Standards**

- What rates of student growth are necessary to fulfill universal proficiency mandates
embedded in accountability systems?

- Are rates of student growth necessary to fulfill universal proficiency mandates realistic?

**Standards and Accountability**

- Given current performance standards, are universal proficiency mandates realistic?

- How should accountability balance the desire for ambitious yet reasonable performance targets for all students?

**Growth and Accountability**

- Given current rates of student progress is failure to meet universal proficiency mandates the fault of the education system?

- What levels of effectiveness are required of the education system (and the education system in general) to fulfill universal proficiency mandates?

This paper argues that the barriers currently present between growth, standards, and accountability are largely the result of incomplete information: normative data lacking criterion referenced standards or criterion referenced standards lacking a normative basis. This view is consistent with an aphorism attributed to Angoff (1974): “scratch a criterion and you’ll find a norm”. A complete understanding of phenomena associated with large scale assessment results requires *both* a normative and criterion referenced component. The normative framework proposed to unify growth, standards and accountability are *student growth percentiles*.

In contrast to the majority of longitudinal analysis techniques using student assessment data that seek to explain the variability of student scores vis-à-vis teacher or school effects, the primary concern with calculating student growth percentiles is to describe this variability and, following similar descriptions used in pediatrics, give stakeholders a sense of what the current range of student growth is.\(^1\) With this normative vocabulary in place, informed and productive discussions of student growth regarding what is?, what should be?, and what is reasonable? can follow.

An intention of providing this normative backdrop to student progress is to promote a move toward descriptive (Linn, 2008) or regulatory (Edley, 2006) approaches to accountability and away from the current high stakes accountability systems geared toward assigning blame for success/failure (i.e., establishing the cause). In the introductory chapter to *The Future of Test Based Educational Accountability* Linn describes such a descriptive approach:

Accountability system results can have value without making causal inferences about school quality, solely from the results of student achievement measures and demographic characteristics. Treating the results as descriptive information

---

\(^1\)See [http://www.nutropin.com/patient/3.5.4.growth.velocity.jsp](http://www.nutropin.com/patient/3.5.4.growth.velocity.jsp) for an online implementation of pediatric growth percentiles for height.
and for identification of schools that require more intensive investigation of organizational and instructional process characteristics are potentially of considerable value. Rather than using the results of the accountability system as the sole determinant of sanctions for schools, they could be used to flag schools that need more intensive investigation to reach sound conclusions about needed improvements or judgments about quality (Linn, 2008, p. 21).

Christopher Edley (2006), in his invited presidential address at the 2006 AERA conference expresses similar sentiments:

This is the difference between a retrospective question of identifying fault as opposed to a prospective strategy to engineer some corrective measure, almost independent of considering whether there was blame-worthiness. And to move away from the blame-worthiness paradigm toward something that is more regulatory in nature where one might seize upon disparities or circumstances that are for some reason deemed unacceptable and engineer the interventions needed to bring about the necessary change. ... It’s the no-fault gap closing strategy in which the effort is to build a consensus about a vision of an improved society rather than figure out where’s the person ... we want to pillory.

As Linn (2008) notes, such an accountability system would represent a profound change from current systems. An essential first step toward such a change is the creation of appropriate and compelling descriptive measures on which to base the system.

**Student Growth Percentiles**

It is a common misconception that to measure student growth in education, the subject matter and grades over which growth is examined must be on the same scale—referred to as a vertical scale. Not only is a vertical scale not necessary, but its existence obscures fundamental concepts necessary to understand growth. Growth, fundamentally, requires change to be examined for a single construct, like math achievement, over time—*growth in what?* A single scale for the construct is necessary to measure the *magnitude* of growth, but not growth in general (Betebenner, 2008; Yen, 2007).

Consider the familiar situation from pediatrics where the interest is on measuring the height and weight of children over time. The scales on which height and weight are measured possess properties that educational assessment scales aspire towards but can never meet.

An infant male toddler is measured at 2 and 3 years of age and is shown to have grown 4 inches. The magnitude of increase—4 inches—is a well understood quantity that any parent can grasp and calculate at home using a simple yardstick. However, parents leaving their pediatrician’s office knowing only how much their child has grown would likely be wanting for more information: Parents are not interested in an absolute magnitude of growth, but instead in a normative criterion locating that 4 inch increase alongside the height increases of
similar children. Examining this height increase relative to the increases of similar children permits a diagnosis of how (ab)normal such an increase is.

Given this reality in the examination of change where scales of measurement are perfect, it is absurd to think that in education, where scales are, at best, quasi-interval, one can/should examine growth differently.

Supposing scales did exist in education similar to height/weight scales that permitted the calculation of absolute measures of annual academic growth for students, the response parents receive to questions such as, “How much did my child progress?”, would come as a number of scale score points—an answer likely to leave most parents bewildered wondering whether the number of points is good or bad. With regard to measures used in education, the search for a description regarding change in achievement over time (i.e., growth) is best conducted in two steps: First by considering a normative quantification of student growth—a student growth percentile. And second by establishing growth adequacy criteria—a percentile growth projection.

A student’s growth percentile describes how (ab)normal a student’s growth is by examining their current achievement relative to their academic peers—those students with identical prior achievement. That is, a student growth percentile examines the current achievement of a student relative to other students who have, in the past, “walked the same achievement path”. Heuristically, if the state assessment data set were extremely large (in fact, infinite) in size, one could examine the data set and select out those students with the exact same prior scores and compare how the selected student’s current year score compares to the current year scores of those students with the same prior year’s scores—their academic peers. If the student’s current year score exceeded the scores of most of their academic peers, in a normative sense they have done well. If the student’s current year score was less than the scores of their academic peers, in a normative sense they have not done well.

The four panels of Figure 1 depict what a student growth percentile represents in a situation considering students having only two consecutive achievement test scores.

**Upper Left Panel** Considering all pairs of scores for all students in the state yields a bivariate (two variable) distribution.

**Upper Right Panel** Conditioning upon prior achievement fixes a the value of the 2005 scale score (in this case at 600) and is represented by the red slice taken out of the bivariate distribution.

**Lower Left Panel** Conditioning upon prior achievement defines a conditional distribution representing the distribution of outcomes on the 2006 test assuming a 2005 score of 600. This distribution is indicating as the solid red curve.

**Lower Right Panel** The conditional distribution provides the context within which a student’s 2006 achievement can be understood normatively. Students with achievement in the upper tail of the conditional distribution have demonstrated high rates of growth relative to their academic peers whereas those students with achievement in the lower tail of the distribution have demonstrated low rates of growth. Students with current
achievement in the middle of the distribution could be described as demonstrating “average” or “typical” growth. In the figure provided the student scores approximately 650 on the 2006 test. Within the conditional distribution, the value of 650 lies at approximately the 70th percentile. Thus the student’s growth from 600 in 2005 to 650 in 2006 met or exceeded that of approximately 70 percent of students starting from the same place. This 50 point increase is above average. It is important to note that qualifying a student growth percentile as “adequate”, “good”, or “enough” is a standard setting procedure requiring stakeholders to examine a student’s growth relative to external criteria such as performance standards/levels.

Figure 1 also illustrates the relationship between a vertical scale and student growth percentiles. Using the vertical scale implied by Figure 1, the student grew 50 points (from 600 to 650) between 2005 and 2006. This 50 points represents the magnitude of change. Quantifying the magnitude of change is scale dependent. However, relative to other students, the achievement growth of the student has not changed—their growth percentile is invariant to scale transformations common in educational assessment. Student growth percentiles normatively situate achievement change bypassing questions associated with the magnitude of change, and directing attention toward relative standing which is likely to interest stakeholders most.
The percentile of a student’s current score within their corresponding conditional distribution translates to a probability statement of a student obtaining that score taking account of prior achievement. That is:

\[ \text{Student Growth Percentile} \equiv \Pr(\text{Current Achievement}|\text{Past Achievement}) \cdot 100. \]

Whereas unconditional percentiles normatively quantify achievement, conditional percentiles normatively quantify growth. Because past scores are used solely for conditioning purposes, one of the major advantages of using growth percentiles to measure change is that estimation does not require a vertical scale.

Student Growth Percentile Calculation

Calculation of a student’s growth percentile is based upon the estimation of the conditional density associated with a student’s score at time \( t \) using the student’s prior scores at times \( 1, 2, \ldots, t - 1 \) as the conditioning variables. Given the conditional density for the student’s score at time \( t \), the student’s growth percentile is defined as the percentile of the score within the time \( t \) conditional density. By examining a student’s current achievement with regard to the conditional density, the student’s growth percentile normatively situates the student’s outcome at time \( t \) taking account of past student performance. The percentile result reflects the likelihood of such an outcome given the student’s prior achievement. In the sense that the student growth percentile translates to the probability of such an outcome occurring (i.e., rarity), it is possible to compare the progress of individuals not beginning at the same starting point. However, occurrences being equally rare does not necessarily imply that they are equally “good”. Qualifying student growth percentiles as “(in)adequate”, “good”, or as satisfying “a year’s growth” is a standard setting procedure requiring external criteria (e.g., growth relative to state performance standards) combined with the wisdom and judgments of stakeholders.

Estimation of the conditional density is performed using quantile regression (Koenker, 2005). Whereas linear regression models the conditional mean of a response variable \( Y \), quantile regression is more generally concerned with the estimation of the family of conditional quantiles of \( Y \). Quantile regression provides a more complete picture of the conditional distribution associated with the response variable(s). The techniques are well suited for estimation of the family of conditional quantile functions (i.e., reference percentile curves). Using quantile regression, the conditional density associated with each student’s prior scores is derived and used to situate the student’s most recent score. Position of the student’s most recent score within this density can then be used to qualify deficient/sufficient/excellent growth. Though many state assessments possess a vertical scale, such a scale is not necessary to produce student growth percentiles.

In analogous fashion to the least squares regression line representing the solution to a minimization problem involving squared deviations, quantile regression functions represent

\(^2\)Since \( \Pr(\text{Current Achievement}|\text{Past Achievement}) \cdot 100 \) is not always an integer between 1 and 100 the expression denotes a student growth quantile. To simplify, the result is rounded down and termed a percentile.
the solution to the optimization of a loss function (Koenker, 2005, p. 5). Formally, given a class of suitably smooth functions, \( G \), one wishes to solve

\[
\arg\min_{g \in G} \sum_{i=1}^{n} \rho_{\tau}(Y(t_i) - g(t_i)),
\]

(1)

where \( t_i \) indexes time, \( Y \) are the time dependent measurements, and \( \rho_{\tau} \) denotes the piecewise linear loss function defined by

\[
\rho_{\tau}(u) = u \cdot (\tau - I(u < 0)) = \begin{cases} u \cdot \tau & u \geq 0 \\ u \cdot (\tau - 1) & u < 0. \end{cases}
\]

The elegance of the quantile regression Expression 1 can be seen by considering the more familiar least squares estimators. For example, calculation of \( \arg\min \sum_{i=1}^{n} (Y_i - \mu)^2 \) over \( \mu \in \mathbb{R} \) yields the sample mean. Similarly, if \( \mu(x) = x'\beta \) is the conditional mean represented as a linear combination of the components of \( x \), calculation of \( \arg\min \sum_{i=1}^{n} (Y_i - x_i'\beta)^2 \) over \( \beta \in \mathbb{R}^p \) gives the familiar least squares regression line. Analogously, when the class of candidate functions \( G \) consists solely of constant functions, the estimation of Expression 1 gives the \( \tau \)-th sample quantile associated with \( Y \). By conditioning on a covariate \( x \), the \( \tau \)-th conditional quantile function, \( Q_{y,\tau}(x) \), is given by

\[
Q_{y,\tau}(x) = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_{\tau}(y_i - x_i'\beta).
\]

In particular, if \( \tau = 0.5 \), then the estimated conditional quantile line is the median regression line.\footnote{For a detailed treatment of the procedures involved in solving the optimization problem associated with Expression 1, see Koenker (2005), particularly Chapter 6.}

Following Wei & He (2006), we parameterize the conditional quantile functions as a linear combination of B-spline cubic basis functions. B-splines are employed to accommodate non-linearity, heteroscedasticity and skewness of the conditional densities associated with values of the independent variable(s). B-splines are attractive both theoretically and computationally in that they provide excellent data fit, seldom lead to estimation problems (Harrell, 2001, p. 20), and are simple to implement in available software.

Figure 2 gives a bivariate representation of linear and B-splines parameterization of decile conditional achievement curves. Note that the assumption of linearity imposes conditions upon the heteroscedasticity of the conditional densities. Close examination of the linear deciles indicates slightly greater variability for higher grade 5 scale scores than for lower scores. By contrast, the B-spline based decile functions better capture the greater variability at both ends of the scale score range together with a slight, non-linear trend to the data.

Calculation of student growth percentiles is performed using R (R Development Core Team, 2009), a software language and environment for statistical computing, with the SGP package (Betebenner, 2009). Other possible software (untested with regard to student growth percentiles) with quantile regression capability include SAS and Stata. Estimation of student growth percentiles is conducted using all available prior data, subject to certain suitability
Figure 2: Linear and B-spline conditional deciles based upon bivariate math data, grades 5 and 6 conditions. Given assessment scores for \( t \) occasions, \((t \geq 2)\), the \( \tau \)-th conditional quantile for \( Y_t \) based upon \( Y_{t-1}, Y_{t-2}, \ldots, Y_1 \) is given by

\[
Q_{Y_t}(\tau|Y_{t-1}, \ldots, Y_1) = \sum_{j=1}^{t-1} \sum_{i=1}^{3} \phi_{ij}(Y_j)\beta_{ij}(\tau),
\]

where \( \phi_{i,j}, i = 1, 2, 3 \) and \( j = 1, \ldots, t - 1 \) denote the B-spline basis functions. Currently, bases consisting of 7 cubic polynomials are used to “smooth” irregularities found in the multivariate assessment data. A bivariate rendering of this is found is Figure 2 where linear and B-spline conditional deciles are presented. The B-spline basis models the heteroscedasticity and non-linearity of the data to a greater extent than is possible using a linear parameterization.

Percentile Growth Projections/Trajectories

Operational work calculating student growth percentiles with state assessment data yields a large number of coefficient matrices derived from estimating Equation 2. These matrices, similar to a lookup table, “encode” the relationship between prior and current achievement scores for students in the norming group (usually an entire grade cohort of students for the state) across all percentiles and can be used both to qualify a student’s current level growth as well as predict, based upon current levels of student progress, what different rates of growth (quantified in the percentile metric) will yield for students statewide.

When rates of growth necessary to reach performance standards are investigated, such calculations are often referred to as “growth-to-standard”. These analyses serve a dual purpose in that they provide the growth rates necessary to reach these standards and also shed light on the standard setting procedure as it plays out across grades. To establish growth per-
centiles necessary to reach different performance/achievement levels, it is necessary to investigate what growth percentile is necessary to reach the desired performance level thresholds based upon the student’s achievement history.

Establishing criterion referenced growth thresholds requires consideration of multiple future growth/achievement scenarios. Instead of inferring that prior student growth is indicative of future student growth (e.g., linearly projecting student achievement into the future based upon past rates of change), predictions of future student achievement are contingent upon initial student status (where the student starts) and subsequent rates of growth (the rate at which the student grows). This avoids fatalistic statements such as, "Student X is projected to be (not) proficient in three years" and instead promotes discussions about the different rates of growth necessary to reach future achievement targets: "In order that Student X reach/maintain proficiency within three years, she will have to demonstrate nth percentile growth consecutively for the next three years." The change is phraseology is minor but significant. Stakeholder conversations turn from "where will (s)he be" to "what will it take?"

**Percentile Growth Projection/Trajectory Calculation**

Parallel growth/achievement scenarios are more easily understood with a picture. Using the results of a statewide assessment growth percentile analyses, Figures 3 and 4 depict future growth scenarios in math and reading, respectively, for a student starting in third grade and tracking that student’s achievement timeline based upon different rates of annual growth expressed in the growth percentile metric. The figures depict the four state achievement levels across grades 3 to 10 in color together with the 2007 achievement percentiles (inner most vertical axis) superimposed in white. Beginning with the student’s achievement starting point at grade 3 a grade 4 achievement projection is made based upon the the most recent growth percentile analyses derived using prior 3rd to 4th grade student progress. More specifically, using the coefficient matrices derived in the quantile regression of grade 4 on grade 3 (see Equation 2), predictions of what 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth lead to are calculated. Next, using these six projected 4th grade scores combined with the student actual 3rd grade score, 5th grade achievement projections are calculated using the most recent quantile regression of grade 5 on grades 3 and 4. Similarly, using these six projected 5th grade scores, the 6 projected 4th grade scores with the students actual third grade score, achievement projections to the 6th grade are calculated using the most recent quantile regression of grade 6 on grades 3, 4, and 5. The analysis extends recursively for grades 6 to 10 yielding the percentile growth trajectories in Figures 3 and 4. The figures allow stakeholders to consider what consecutive consecutive rates of growth, expressed in growth percentiles, yield for students starting at different points.

Figure 3 depicts percentile growth trajectories in mathematics for a student beginning at the unsatisfactory/partially proficient threshold. Based upon the achievement percentiles depicted (the white contour lines), approximately 7 percent of the population of 3rd graders rate as unsatisfactory. Moving toward grade 10, the percentage of unsatisfactory students increases dramatically to near 35 percent. The black lines in the figure represent six different growth scenarios for the student based upon consecutive growth at a given growth per-
Figure 3: Growth chart depicting future mathematics achievement conditional upon consecutive 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the unsatisfactory/partially proficient cutpoint.
Reading Growth Percentile Trajectories
Growth Trajectories for 10th, 25th, 40th, 50th, 60th, 75th, & 90th Percentiles

Figure 4: Growth chart depicting future reading achievement conditional upon consecutive 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the partially proficient/proficient cutpoint
centile, denoted by the right axis. At the lower end, for example, consecutive 25th percentile growth leaves the student, unsurprisingly, mired in the unsatisfactory category. Consecutive 40th, 50th and 60th percentile growth also leave the child in the unsatisfactory category. This demonstrates how difficult (probabilistically, based upon current rates of progress) it is for students to move up in performance level in math statewide. With the green region representing proficient, a student would need to demonstrate growth percentiles consecutively in excess of 75 to reach proficiency showing how unlikely such a event currently is. In light of NCLB universal proficiency mandates, the growth necessary for non-proficient students to reach proficiency, absent radical changes to growth rates of students statewide, is likely unattainable for a large percentage of non-proficient students.

Figure 4 depicts percentile growth trajectories in reading for a student beginning at the partially proficient/proficient threshold in grade 3. In a normative sense, the performance standards in reading are less demanding than those in mathematics (particularly in the higher grades) with approximately 30 percent of students below proficient in grades 3 to 10. The black lines in the figure represent seven growth scenarios for the hypothetical student based upon consecutive growth at a the given growth percentile. Compared with the growth required in mathematics, more modest growth is required to maintain proficiency. Typical growth (50th percentile growth) appears adequate for such a student to move up slightly into the proficiency category.

Discussion

Having established a normative framework for understanding student growth and anchored that framework to the performance standards used in current accountability systems, addressing issues/questions related to growth, accountability and standards is straightforward:

Growth and Standards

- What rates of student growth are necessary to fulfill universal proficiency mandates embedded in accountability systems?

- Are rates of student growth necessary to fulfill universal proficiency mandates realistic?

Clearly, the rates of student growth necessary to reach or maintain proficiency differ by student based upon the current level of achievement for the student. As the achievement level of the student decreases, the rate of growth necessary for them to “catch-up” increases. Figures 3 and 4 indicate the rates of growth necessary in mathematics and reading for third graders beginning 3rd grade at two different levels of achievement to reach various targets.

Similar figures are possible for every student in the education system. Along these lines, the state of Colorado currently produces student reports (see Figures 5, 6 and 7 on Pages 18 to 20) for parents and teachers showing individualized historical rates of growth as well as anticipated rates of growth necessary to reach future achievement goals.

One of the strengths of quantifying student growth normatively is that the growth percentile targets that are calculated of what it will take to reach a level of achievement quickly translate into the likelihood of such an event occurring. This dimension of student progress
as it relates to accountability is absent from most growth-to-standard discussions. Today, achievement mandates are stipulated based upon the moral imperative of high standards for all children with little concern regarding the likelihood the students reaching these goals. Given current progress of students, it is unlikely that the sustained levels of growth necessary to reach these standards will occur will occur without substantial changes to both the education system and society as a whole.

Standards and Accountability

- Are universal proficiency mandates realistic in terms of current performance standards?

- How should accountability balance the desire for ambitious yet reasonable performance targets for all students?

The difference in growth required to maintain or progress in mathematics and reading in Figures 3 and 3 is not surprising once one examines the normative stringency of the performance standards across grades. In mathematics, for example, with decreasing percentages of students reaching partially proficient, proficient, and advanced in higher grades, it is a tautology that high normative rates of growth will be required to maintain or move up. Similar, with near constant percentages of students at or above proficient in reading, typical rates of growth will be sufficient to maintain. With increasing percentages of students reach performance standards in higher grades, below typical rates of growth would be needed for students to “catch-up”.

Anchoring growth to achievement goals normatively adds the prospect of making adequacy judgments to the normative measures that are used. If standards based accountability policy demands universal proficiency, then for each student it is possible to determine what growth percentile is necessary to reach proficiency in 1 or more years. Accountability policy mandates can be immediately scrutinized in terms of reasonableness. If the growth percentiles indicate an unreasonable expectation then realization of these goals will likely require either policy expectation or the standards on which they are based to change. The metric allows for clear discussions regarding the demands being made.

Growth and Accountability

- Given current student growth rates is failure to meet universal proficiency mandates the fault of the education system?

- What levels of effectiveness are required of the education system (and the education system in general) to fulfill universal proficiency mandates?

The question of responsibility for (lack of) student growth is extremely difficult. Viewed normatively, this is the question value-added models address. When anchored to standards, the question becomes as much philosophical as statistical. For thousands of students in state education systems, the consecutive growth percentiles necessary for these students to reach proficiency exceed 90. No school is systemically producing these rates of growth for its average student. Is it fair to place responsibility for failure to reach these impossible goals on schools and the educators working within them? A fundamental dictum of moral philosophy ascribed to Kant is that “ought implies can”: If someone ought to do something, they
can do it, in the sense that they have the possibility/capacity to do it. Growth percentiles bring Kant’s dictum to the fore when considering performance standards, the likelihood of students reaching/maintaining these standards, and accountability systems holding stakeholders responsible for student failure to reach/maintain these standards.

The reality of the present needn’t define a blueprint for the future. A primary goal of education reform is to turn today’s exemplary growth into tomorrow’s typical growth. Achieving that would bring us closer to the goal of universal proficiency engendered by NCLB. To that end, this paper introduces student growth percentiles as a quantification of student progress and a descriptive measure of what is?. Criterion referenced questions of what should be? coincide with decisions about whether growth is “enough” or “adequate” to reach or maintain desired levels of achievement. Student growth percentiles and percentile growth projections/trajectories serve to inform the standard setting procedure by communicating what is reasonable?. Only by considering, what is, what should be, and what is reasonable simultaneously, can accountability systems, built upon performance metrics like student growth percentiles, be equitable, just, and truly informed.

References

of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6),
3–16.
R Development Core Team. (2009). R: A language and environment for statistical computing
Individual Growth and Achievement Reports

The individual growth and achievement reports that follow are presented for reading, writing, and mathematics. The reports show an achievement timeline for the student depicting historical achievement and growth data together with projections of what low, typical and high growth will lead to. The reports use the state achievement levels as a backdrop to convey to stakeholders the absolute criteria against which student progress is judged. The figures are intended to assist in better understanding the range of student growth and what different growth rates lead to in terms of student achievement relative to state designated performance levels.
Figure 5: Individual growth and achievement report in reading depicting historical achievement and growth together with growth projections.
Figure 6: Individual growth and achievement report in writing depicting historical achievement and growth together with growth projections.
Figure 7: Individual growth and achievement report in mathematics depicting historical achievement and growth together with growth projections.
Attachment G

Preliminary Analysis of Three Growth Models Using ISAT
Hierarchical Modeling Approach:
Wisconsin group (Meyer & Dokumaci, 2010)

Post-on-Pre (using adjacent grades) two period value added model makes use of two years of consecutive grade longitudinal assessment data and answers the question, “To what extent do schools contribute to student growth, controlling for other sources such as prior achievement and student and family characteristics?” Value added systems facilitate comparisons of student growth across schools serving different student populations. While typically based upon student test scores in a particular content area, VAMs could be expanded to include prior achievement in other areas and other student measures such as attendance and demographics. While there may be some advantage of expanding the model to include three or more periods, an initial investigation of the feasibility and utility of a two period model would represent a significant step forward in providing growth information for accountability purposes. Though some multi-level VAMs are dependent upon vertical scales, this simple model for use at the school level could work within the parameters of the current ISAT. CPS has been using VAM at the school level for the past few years.

Meyer and Dokumaci see two major reasons for including student characteristics in value-added models (p. 6)

1. Usual purpose: To identify differences between subgroups (e.g., low vs. h high poverty)
2. HLM specific: To control for differences across schools/districts in student composition to achieve an “apples to apples” comparison.

Example:

\[
\text{ISAT}_8 \sim \lambda \text{ISAT}_7 + C, \text{ where } C \text{ can be interpreted as the “Value Added”}
\]

Does \( \lambda \) depend on student characteristics?

\[
\lambda \sim w_0 + w_1 \text{FRL}_\text{dist} + w_2 \text{LEP}_\text{dist} + w_3 \text{IEP}_\text{dist}
\]

If so, then demographics change the basic equation, and C (Value Added) is biased.
As an example, I applied the above to Grade 7-8 ISAT data for 2010 using HLM 6, using **aggregated district (not school) demographics as a second level variable.**

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>T-ratio</th>
<th>d.f.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT2, B0</td>
<td>54.247708</td>
<td>0.401502</td>
<td>135.112</td>
<td>733</td>
<td>0.000</td>
</tr>
<tr>
<td>MATH7 slope, B1</td>
<td>0.834957</td>
<td>0.001420</td>
<td>588.049</td>
<td>124667</td>
<td>0.000</td>
</tr>
<tr>
<td>IEP_MEAN, G11</td>
<td>-0.014165</td>
<td>0.007519</td>
<td>-1.884</td>
<td>124667</td>
<td>0.059</td>
</tr>
<tr>
<td>LEP_MEAN, G12</td>
<td>-0.024139</td>
<td>0.010530</td>
<td>-2.292</td>
<td>124667</td>
<td>0.022</td>
</tr>
<tr>
<td>FRL_MEAN, G13</td>
<td>-0.019963</td>
<td>0.002314</td>
<td>-8.626</td>
<td>124667</td>
<td>0.000</td>
</tr>
</tbody>
</table>

So, only FRL matters. Hence, this redo:

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>T-ratio</th>
<th>d.f.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT2, B0</td>
<td>54.381157</td>
<td>0.395024</td>
<td>137.665</td>
<td>733</td>
<td>0.000</td>
</tr>
<tr>
<td>MATH7 slope, B1</td>
<td>0.841588</td>
<td>0.001610</td>
<td>522.664</td>
<td>124669</td>
<td>0.000</td>
</tr>
<tr>
<td>FRL_MEAN, G11</td>
<td>-0.022117</td>
<td>0.002176</td>
<td>-10.165</td>
<td>124669</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Summary**

The upshot is that the same formula cannot be applied across districts as it differs depending on poverty. The FRL weight is -0.02, indicating that greater poverty at the district level pushes anticipated performance “automatically” down. (Centering vs. non-centering does not alter this conclusion).
Student Growth Percentiles
Betebenner (2009)

This approach tries to answer the question, “How much did students progress in relation to their academic peers (i.e. students with identical prior achievement)?” Because past scores are used solely for conditioning purposes (i.e., only ordinal info is needed or used), use of student growth percentiles does not require a vertical scale. They differ from the previous approaches by focusing on a student’s growth relative to other students rather than the magnitude of growth itself. Normative models such as student growth percentiles are helpful for determining “reasonable” growth as compared to the very high rates of growth in trajectory or projection models based on NCLB targets of 100% proficient by 2014.

Student growth percentiles are investigated most comprehensively with “Quantile regression.” This approach is less widely known than Hierarchical Regression type modeling, and currently it can be done only using the R programming language – there is no nice user interface as there is for HLM when doing hierarchical modeling. As there is no commercial software, the following uses Koenker’s 1995 (and later) R software.

In the following figures we see that in the 2009 – 2010 ISAT data the quantile intervals widen as the 2009 ISAT math score increase. Thus, the performance of poor performers is more uniform, and greater increases are made only by high performing students. Again, this argues against the use of a single “gain” score.

Since FRL students are found disproportionately among low performing students, these results agree with the hierarchical regression approach.
Value tables or transition models

Also called the “growth to a standard or trajectory model,” this approach tracks changes in performance levels from one year to the next for each student. The goal is for each student to transition to the higher performance levels. The tables for a school or district summarize students’ changes in performance levels, rather than scale scores. Therefore, vertical scales are not needed. Additionally, alternate assessments performance level data can be combined with the regular assessment data in the summary tables.

Delaware, Iowa, and Michigan incorporate value tables or transition models. These states determine whether students are improving, declining, or maintaining their performance levels. For the non-proficient students, the states determine whether they are on track to becoming proficient over time. Michigan and Delaware subdivided the non-proficient performance levels into sub-levels of low, middle, and high. Iowa also created boundaries for non-proficient students. For each state, a non-proficient student’s growth trajectory towards becoming proficient needs to cross over each sub-performance level to meet growth expectations. All of the states in their AYP calculations add the number of students on track to becoming proficient to those that are proficient to the numerator.

$$100 \times \frac{N \text{ on track to becoming proficient} + N \text{ proficient}}{N \text{ of valid assessments}}$$

After matching SIS numbers of 2009 and 2010 ISAT student data, it was determined that in mathematics, regardless of grade level, three-fourths of the students maintained their performance levels. For example, below standards students remained below standards in the next year, and meets standards students remained meets standards in the next year, etc. For reading, about seventy percent maintained their performance levels. Therefore, changes in performance levels are not sensitive enough to track, so our performance levels should be further subdivided to track growth over time.

Our main finding is that the current four Illinois performance categories are too broad to define a transition model. For instance, up to eighty percent of the students stay in the same category from yea-to-year.

<table>
<thead>
<tr>
<th>Math 2009 to 2010</th>
<th>2010 Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2009 Results</strong></td>
<td>Acad Warn</td>
</tr>
<tr>
<td>Acad Warn</td>
<td>16.1%</td>
</tr>
<tr>
<td>Below</td>
<td>3.7%</td>
</tr>
<tr>
<td>Meets</td>
<td>0.1%</td>
</tr>
<tr>
<td>Exceeds</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

74.3% Stay the Same
12.6% Do Better
13.1% Do Worse
100.0%
### Read 2009 to 2010

<table>
<thead>
<tr>
<th>2009 Results</th>
<th>Acad Warn</th>
<th>Below</th>
<th>Meets</th>
<th>Exceeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acad Warn</td>
<td>11.4%</td>
<td>82.7%</td>
<td>5.7%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Below</td>
<td>1.2%</td>
<td>64.0%</td>
<td>34.3%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Meets</td>
<td>0.0%</td>
<td>10.3%</td>
<td>76.6%</td>
<td>13.0%</td>
</tr>
<tr>
<td>Exceeds</td>
<td>0.0%</td>
<td>0.4%</td>
<td>34.6%</td>
<td>65.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2010 Results</th>
<th>Acad Warn</th>
<th>Below</th>
<th>Meets</th>
<th>Exceeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acad Warn</td>
<td>11.4%</td>
<td>82.7%</td>
<td>5.7%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Below</td>
<td>1.2%</td>
<td>64.0%</td>
<td>34.3%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Meets</td>
<td>0.0%</td>
<td>10.3%</td>
<td>76.6%</td>
<td>13.0%</td>
</tr>
<tr>
<td>Exceeds</td>
<td>0.0%</td>
<td>0.4%</td>
<td>34.6%</td>
<td>65.0%</td>
</tr>
</tbody>
</table>

- 69.9% Stay the Same
- 15.7% Do Better
- 14.5% Do Worse
- 100.0%

To obtain a better picture, ISBE also analyzed data by creating subgroups based on standard errors away from the proficient mark, both above and below. As the tables show, doing this allows us to track changes in student performance at a finer and more useful level.
### m9 * m0 * GR Crosstabulation

<table>
<thead>
<tr>
<th>GR</th>
<th>m9</th>
<th>3+SE below</th>
<th>2 below</th>
<th>1 below</th>
<th>1 above</th>
<th>2 above</th>
<th>3+SE above</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>m9</td>
<td>4807</td>
<td>1192</td>
<td>1160</td>
<td>797</td>
<td>458</td>
<td>502</td>
<td>8916</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>53.9%</td>
<td>13.4%</td>
<td>13.0%</td>
<td>8.9%</td>
<td>5.1%</td>
<td>5.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>1431</td>
<td>762</td>
<td>1055</td>
<td>885</td>
<td>588</td>
<td>731</td>
<td>5452</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>26.2%</td>
<td>14.0%</td>
<td>19.4%</td>
<td>16.2%</td>
<td>10.8%</td>
<td>13.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>778</td>
<td>514</td>
<td>864</td>
<td>887</td>
<td>736</td>
<td>1063</td>
<td>4842</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>16.1%</td>
<td>10.6%</td>
<td>17.8%</td>
<td>18.3%</td>
<td>15.2%</td>
<td>22.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>767</td>
<td>583</td>
<td>1112</td>
<td>1342</td>
<td>1404</td>
<td>2405</td>
<td>7613</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>10.1%</td>
<td>7.7%</td>
<td>14.6%</td>
<td>17.6%</td>
<td>18.4%</td>
<td>31.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>2 above</td>
<td>497</td>
<td>381</td>
<td>911</td>
<td>1360</td>
<td>1686</td>
<td>4045</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>5.6%</td>
<td>4.3%</td>
<td>10.3%</td>
<td>15.3%</td>
<td>19.0%</td>
<td>45.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>3+SE above</td>
<td>673</td>
<td>660</td>
<td>1600</td>
<td>2833</td>
<td>4842</td>
<td>99026</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>.6%</td>
<td>.6%</td>
<td>1.5%</td>
<td>2.6%</td>
<td>4.4%</td>
<td>90.3%</td>
<td>109634</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>Total</td>
<td>8953</td>
<td>4092</td>
<td>6702</td>
<td>8104</td>
<td>9714</td>
<td>107772</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>6.2%</td>
<td>2.8%</td>
<td>4.6%</td>
<td>5.6%</td>
<td>6.7%</td>
<td>74.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>5</td>
<td>m9</td>
<td>4866</td>
<td>1761</td>
<td>838</td>
<td>686</td>
<td>335</td>
<td>376</td>
<td>8862</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>54.9%</td>
<td>19.9%</td>
<td>9.5%</td>
<td>7.7%</td>
<td>3.8%</td>
<td>4.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>1690</td>
<td>1361</td>
<td>874</td>
<td>784</td>
<td>539</td>
<td>618</td>
<td>5866</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>28.8%</td>
<td>23.2%</td>
<td>14.9%</td>
<td>13.4%</td>
<td>9.2%</td>
<td>10.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>1 below</td>
<td>972</td>
<td>1011</td>
<td>782</td>
<td>943</td>
<td>633</td>
<td>834</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>18.8%</td>
<td>19.5%</td>
<td>15.1%</td>
<td>18.2%</td>
<td>12.2%</td>
<td>16.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>1 above</td>
<td>1055</td>
<td>1503</td>
<td>1451</td>
<td>2040</td>
<td>1714</td>
<td>2783</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>10.0%</td>
<td>14.3%</td>
<td>13.8%</td>
<td>19.3%</td>
<td>16.3%</td>
<td>26.4%</td>
<td>10546</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>2 above</td>
<td>465</td>
<td>837</td>
<td>999</td>
<td>1766</td>
<td>1839</td>
<td>4263</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>4.6%</td>
<td>8.2%</td>
<td>9.8%</td>
<td>17.4%</td>
<td>18.1%</td>
<td>41.9%</td>
<td>10169</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>3+SE above</td>
<td>512</td>
<td>881</td>
<td>1421</td>
<td>3255</td>
<td>5414</td>
<td>92098</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>.5%</td>
<td>.9%</td>
<td>1.4%</td>
<td>3.1%</td>
<td>5.2%</td>
<td>88.9%</td>
<td>103581</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>Total</td>
<td>9560</td>
<td>7354</td>
<td>6365</td>
<td>9474</td>
<td>10474</td>
<td>100972</td>
</tr>
<tr>
<td></td>
<td>% within m9</td>
<td>6.6%</td>
<td>5.1%</td>
<td>4.4%</td>
<td>6.6%</td>
<td>7.3%</td>
<td>70.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>GR</td>
<td>m9</td>
<td>3+SE below</td>
<td>2 below</td>
<td>1 below</td>
<td>1 above</td>
<td>2 above</td>
<td>3+SE above</td>
<td>Total</td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>------------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>m9</td>
<td>3+SE below</td>
<td>6063</td>
<td>2180</td>
<td>1274</td>
<td>1198</td>
<td>646</td>
<td>715</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>50.2%</td>
<td>18.1%</td>
<td>10.5%</td>
<td>9.9%</td>
<td>5.3%</td>
<td>5.9%</td>
</tr>
<tr>
<td></td>
<td>2 below</td>
<td>Count</td>
<td>1422</td>
<td>1129</td>
<td>814</td>
<td>1026</td>
<td>687</td>
<td>946</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>23.6%</td>
<td>18.7%</td>
<td>13.5%</td>
<td>17.0%</td>
<td>11.4%</td>
<td>15.7%</td>
</tr>
<tr>
<td></td>
<td>1 below</td>
<td>Count</td>
<td>973</td>
<td>950</td>
<td>918</td>
<td>1313</td>
<td>1093</td>
<td>1770</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>13.9%</td>
<td>13.5%</td>
<td>13.1%</td>
<td>18.7%</td>
<td>15.6%</td>
<td>25.2%</td>
</tr>
<tr>
<td></td>
<td>1 above</td>
<td>Count</td>
<td>756</td>
<td>1045</td>
<td>1084</td>
<td>1843</td>
<td>1940</td>
<td>4198</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>7.0%</td>
<td>9.6%</td>
<td>10.0%</td>
<td>17.0%</td>
<td>17.9%</td>
<td>38.6%</td>
</tr>
<tr>
<td></td>
<td>2 above</td>
<td>Count</td>
<td>338</td>
<td>546</td>
<td>735</td>
<td>1421</td>
<td>1983</td>
<td>6926</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>2.8%</td>
<td>4.6%</td>
<td>6.2%</td>
<td>11.9%</td>
<td>16.6%</td>
<td>58.0%</td>
</tr>
<tr>
<td></td>
<td>3+SE above</td>
<td>Count</td>
<td>383</td>
<td>540</td>
<td>725</td>
<td>1765</td>
<td>3273</td>
<td>90929</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>.4%</td>
<td>.6%</td>
<td>.7%</td>
<td>1.8%</td>
<td>3.4%</td>
<td>93.2%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Count</td>
<td>9935</td>
<td>6390</td>
<td>5550</td>
<td>8566</td>
<td>9622</td>
<td>105484</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>6.8%</td>
<td>4.4%</td>
<td>3.8%</td>
<td>5.9%</td>
<td>6.6%</td>
<td>72.5%</td>
</tr>
<tr>
<td>7</td>
<td>m9</td>
<td>3+SE below</td>
<td>5776</td>
<td>1245</td>
<td>910</td>
<td>782</td>
<td>401</td>
<td>343</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>61.1%</td>
<td>13.2%</td>
<td>9.6%</td>
<td>8.3%</td>
<td>4.2%</td>
<td>3.6%</td>
</tr>
<tr>
<td></td>
<td>2 below</td>
<td>Count</td>
<td>2171</td>
<td>1044</td>
<td>1016</td>
<td>1080</td>
<td>705</td>
<td>643</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>32.6%</td>
<td>15.7%</td>
<td>15.3%</td>
<td>16.2%</td>
<td>10.6%</td>
<td>9.7%</td>
</tr>
<tr>
<td></td>
<td>1 below</td>
<td>Count</td>
<td>1492</td>
<td>980</td>
<td>1181</td>
<td>1571</td>
<td>1264</td>
<td>1450</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>18.8%</td>
<td>12.3%</td>
<td>14.9%</td>
<td>19.8%</td>
<td>15.9%</td>
<td>18.3%</td>
</tr>
<tr>
<td></td>
<td>1 above</td>
<td>Count</td>
<td>814</td>
<td>772</td>
<td>1093</td>
<td>1705</td>
<td>1806</td>
<td>2944</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>8.9%</td>
<td>8.5%</td>
<td>12.0%</td>
<td>18.7%</td>
<td>19.8%</td>
<td>32.2%</td>
</tr>
<tr>
<td></td>
<td>2 above</td>
<td>Count</td>
<td>441</td>
<td>456</td>
<td>788</td>
<td>1629</td>
<td>2085</td>
<td>5102</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>4.2%</td>
<td>4.3%</td>
<td>7.5%</td>
<td>15.5%</td>
<td>19.9%</td>
<td>48.6%</td>
</tr>
<tr>
<td></td>
<td>3+SE above</td>
<td>Count</td>
<td>455</td>
<td>481</td>
<td>846</td>
<td>2385</td>
<td>4493</td>
<td>92944</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>.4%</td>
<td>.5%</td>
<td>.8%</td>
<td>2.3%</td>
<td>4.4%</td>
<td>91.5%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Count</td>
<td>11149</td>
<td>4978</td>
<td>5834</td>
<td>9152</td>
<td>10754</td>
<td>103426</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>7.7%</td>
<td>3.4%</td>
<td>4.0%</td>
<td>6.3%</td>
<td>7.4%</td>
<td>71.2%</td>
</tr>
<tr>
<td>GR</td>
<td>m9</td>
<td>3+SE below</td>
<td>Count</td>
<td>2 below</td>
<td>Count</td>
<td>1 below</td>
<td>Count</td>
<td>1 above</td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>------------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>8</td>
<td>m9</td>
<td>3+SE below</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Count</td>
<td>5023</td>
<td>2873</td>
<td>2051</td>
<td>1107</td>
<td>499</td>
<td>448</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>41.9%</td>
<td>23.9%</td>
<td>17.1%</td>
<td>9.2%</td>
<td>4.2%</td>
<td>3.7%</td>
</tr>
<tr>
<td></td>
<td>2 below</td>
<td>Count</td>
<td>978</td>
<td>1061</td>
<td>1245</td>
<td>1030</td>
<td>627</td>
<td>556</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>17.8%</td>
<td>19.3%</td>
<td>22.6%</td>
<td>18.7%</td>
<td>11.4%</td>
<td>10.1%</td>
</tr>
<tr>
<td></td>
<td>1 below</td>
<td>Count</td>
<td>680</td>
<td>1001</td>
<td>1317</td>
<td>1320</td>
<td>1001</td>
<td>1100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>10.6%</td>
<td>15.6%</td>
<td>20.5%</td>
<td>20.6%</td>
<td>15.6%</td>
<td>17.1%</td>
</tr>
<tr>
<td></td>
<td>1 above</td>
<td>Count</td>
<td>447</td>
<td>976</td>
<td>1658</td>
<td>2123</td>
<td>1902</td>
<td>2829</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>4.5%</td>
<td>9.8%</td>
<td>16.7%</td>
<td>21.4%</td>
<td>19.1%</td>
<td>28.5%</td>
</tr>
<tr>
<td></td>
<td>2 above</td>
<td>Count</td>
<td>222</td>
<td>525</td>
<td>1191</td>
<td>1874</td>
<td>2348</td>
<td>5154</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>2.0%</td>
<td>4.6%</td>
<td>10.5%</td>
<td>16.6%</td>
<td>20.8%</td>
<td>45.6%</td>
</tr>
<tr>
<td></td>
<td>3+SE above</td>
<td>Count</td>
<td>187</td>
<td>447</td>
<td>1129</td>
<td>2436</td>
<td>4367</td>
<td>92569</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>.2%</td>
<td>.4%</td>
<td>1.1%</td>
<td>2.4%</td>
<td>4.3%</td>
<td>91.5%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Count</td>
<td>7537</td>
<td>6883</td>
<td>8591</td>
<td>9890</td>
<td>10744</td>
<td>102656</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within m9</td>
<td>5.2%</td>
<td>4.7%</td>
<td>5.9%</td>
<td>6.8%</td>
<td>7.3%</td>
<td>70.2%</td>
</tr>
<tr>
<td>GR- PREVIOUS YEAR</td>
<td>NEXT YEAR</td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3+SE below</td>
<td>2 below</td>
<td>1 below</td>
<td>1 above</td>
<td>2 above</td>
<td>3+SE above</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3+SE below</td>
<td>15154</td>
<td>3190</td>
<td>1701</td>
<td>1822</td>
<td>1062</td>
<td>917</td>
<td>23846</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>63.5%</td>
<td>13.4%</td>
<td>7.1%</td>
<td>7.6%</td>
<td>4.5%</td>
<td>3.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>2 below</td>
<td>2466</td>
<td>1429</td>
<td>1026</td>
<td>1454</td>
<td>955</td>
<td>954</td>
<td>8284</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>29.8%</td>
<td>17.3%</td>
<td>12.4%</td>
<td>17.6%</td>
<td>11.5%</td>
<td>11.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>1 below</td>
<td>1746</td>
<td>1360</td>
<td>1112</td>
<td>1719</td>
<td>1521</td>
<td>1864</td>
<td>9322</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>18.7%</td>
<td>14.6%</td>
<td>11.9%</td>
<td>18.4%</td>
<td>16.3%</td>
<td>20.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>1 above</td>
<td>1076</td>
<td>1094</td>
<td>1066</td>
<td>1948</td>
<td>2038</td>
<td>3284</td>
<td>10506</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>10.2%</td>
<td>10.4%</td>
<td>10.1%</td>
<td>18.5%</td>
<td>19.4%</td>
<td>31.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>2 above</td>
<td>688</td>
<td>795</td>
<td>841</td>
<td>1842</td>
<td>2368</td>
<td>5242</td>
<td>11776</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>5.8%</td>
<td>6.8%</td>
<td>7.1%</td>
<td>15.6%</td>
<td>20.1%</td>
<td>44.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>3+SE above</td>
<td>764</td>
<td>868</td>
<td>1082</td>
<td>3122</td>
<td>6019</td>
<td>69508</td>
<td>81363</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>9.1%</td>
<td>11.1%</td>
<td>13.8%</td>
<td>3.8%</td>
<td>7.4%</td>
<td>85.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>21894</td>
<td>8736</td>
<td>6828</td>
<td>11907</td>
<td>13963</td>
<td>81769</td>
<td>145097</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>15.1%</td>
<td>6.0%</td>
<td>4.7%</td>
<td>8.2%</td>
<td>9.6%</td>
<td>56.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td>5</td>
<td>3+SE below</td>
<td>13077</td>
<td>3210</td>
<td>2516</td>
<td>1993</td>
<td>740</td>
<td>608</td>
<td>22144</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>59.1%</td>
<td>14.5%</td>
<td>11.4%</td>
<td>9.0%</td>
<td>3.3%</td>
<td>2.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>2 below</td>
<td>1957</td>
<td>1377</td>
<td>1552</td>
<td>1786</td>
<td>871</td>
<td>748</td>
<td>8291</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>23.6%</td>
<td>16.6%</td>
<td>18.7%</td>
<td>21.5%</td>
<td>10.5%</td>
<td>9.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>1 below</td>
<td>998</td>
<td>863</td>
<td>1193</td>
<td>1651</td>
<td>943</td>
<td>924</td>
<td>6572</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>15.2%</td>
<td>13.1%</td>
<td>18.2%</td>
<td>25.1%</td>
<td>14.3%</td>
<td>14.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>1 above</td>
<td>1209</td>
<td>1360</td>
<td>2147</td>
<td>3691</td>
<td>2792</td>
<td>3686</td>
<td>14885</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>8.1%</td>
<td>9.1%</td>
<td>14.4%</td>
<td>24.8%</td>
<td>18.8%</td>
<td>24.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>2 above</td>
<td>403</td>
<td>586</td>
<td>1182</td>
<td>2716</td>
<td>2800</td>
<td>5637</td>
<td>13324</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>3.0%</td>
<td>4.4%</td>
<td>8.9%</td>
<td>20.4%</td>
<td>21.0%</td>
<td>42.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>3+SE above</td>
<td>404</td>
<td>513</td>
<td>1191</td>
<td>3848</td>
<td>6314</td>
<td>66474</td>
<td>78744</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>.5%</td>
<td>.7%</td>
<td>1.5%</td>
<td>4.9%</td>
<td>8.0%</td>
<td>84.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>18048</td>
<td>7909</td>
<td>9781</td>
<td>15685</td>
<td>14460</td>
<td>78077</td>
<td>143960</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>12.5%</td>
<td>5.5%</td>
<td>6.8%</td>
<td>10.9%</td>
<td>10.0%</td>
<td>54.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>GR- PREVIOUS YEAR</td>
<td>NEXT YEAR</td>
<td></td>
<td></td>
<td></td>
<td>3+SE above</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3+SE below</td>
<td>2 below</td>
<td>1 below</td>
<td>1 above</td>
<td>2 above</td>
<td>3+SE above</td>
</tr>
<tr>
<td>6</td>
<td>r9</td>
<td>Count</td>
<td>9382</td>
<td>3296</td>
<td>2245</td>
<td>2132</td>
<td>908</td>
<td>674</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>50.3%</td>
<td>17.7%</td>
<td>12.0%</td>
<td>11.4%</td>
<td>4.9%</td>
<td>3.6%</td>
</tr>
<tr>
<td></td>
<td>2 below</td>
<td>Count</td>
<td>1396</td>
<td>1623</td>
<td>1822</td>
<td>2489</td>
<td>1494</td>
<td>1400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13.7%</td>
<td>15.9%</td>
<td>17.8%</td>
<td>24.3%</td>
<td>14.6%</td>
<td>13.7%</td>
</tr>
<tr>
<td></td>
<td>1 below</td>
<td>Count</td>
<td>533</td>
<td>904</td>
<td>1191</td>
<td>2449</td>
<td>1861</td>
<td>2284</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.8%</td>
<td>9.8%</td>
<td>12.9%</td>
<td>26.6%</td>
<td>20.2%</td>
<td>24.8%</td>
</tr>
<tr>
<td></td>
<td>1 above</td>
<td>Count</td>
<td>397</td>
<td>718</td>
<td>1213</td>
<td>3096</td>
<td>3313</td>
<td>6414</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.6%</td>
<td>4.7%</td>
<td>8.0%</td>
<td>20.4%</td>
<td>21.9%</td>
<td>42.3%</td>
</tr>
<tr>
<td></td>
<td>2 above</td>
<td>Count</td>
<td>182</td>
<td>300</td>
<td>583</td>
<td>1753</td>
<td>2586</td>
<td>8399</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.3%</td>
<td>2.2%</td>
<td>4.2%</td>
<td>12.7%</td>
<td>18.7%</td>
<td>60.8%</td>
</tr>
<tr>
<td></td>
<td>3+SE above</td>
<td>Count</td>
<td>189</td>
<td>268</td>
<td>440</td>
<td>1912</td>
<td>4066</td>
<td>71392</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.2%</td>
<td>.3%</td>
<td>.6%</td>
<td>2.4%</td>
<td>5.2%</td>
<td>91.2%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Count</td>
<td>12079</td>
<td>7109</td>
<td>7494</td>
<td>13831</td>
<td>14228</td>
<td>90563</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8.3%</td>
<td>4.9%</td>
<td>5.2%</td>
<td>9.5%</td>
<td>9.8%</td>
<td>62.3%</td>
</tr>
<tr>
<td>7</td>
<td>r9</td>
<td>Count</td>
<td>8889</td>
<td>1688</td>
<td>1232</td>
<td>732</td>
<td>394</td>
<td>289</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>67.2%</td>
<td>12.8%</td>
<td>9.3%</td>
<td>5.5%</td>
<td>3.0%</td>
<td>2.2%</td>
</tr>
<tr>
<td></td>
<td>2 below</td>
<td>Count</td>
<td>2112</td>
<td>1142</td>
<td>1130</td>
<td>875</td>
<td>588</td>
<td>422</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>33.7%</td>
<td>18.2%</td>
<td>18.0%</td>
<td>14.0%</td>
<td>9.4%</td>
<td>6.7%</td>
</tr>
<tr>
<td></td>
<td>1 below</td>
<td>Count</td>
<td>1708</td>
<td>1361</td>
<td>1682</td>
<td>1567</td>
<td>1154</td>
<td>1065</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20.0%</td>
<td>15.9%</td>
<td>19.7%</td>
<td>18.4%</td>
<td>13.5%</td>
<td>12.5%</td>
</tr>
<tr>
<td></td>
<td>1 above</td>
<td>Count</td>
<td>1566</td>
<td>1557</td>
<td>2425</td>
<td>3077</td>
<td>3125</td>
<td>3926</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10.0%</td>
<td>9.9%</td>
<td>15.5%</td>
<td>19.6%</td>
<td>19.9%</td>
<td>25.0%</td>
</tr>
<tr>
<td></td>
<td>2 above</td>
<td>Count</td>
<td>662</td>
<td>787</td>
<td>1508</td>
<td>2499</td>
<td>3286</td>
<td>6916</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.2%</td>
<td>5.0%</td>
<td>9.6%</td>
<td>16.0%</td>
<td>21.0%</td>
<td>44.2%</td>
</tr>
<tr>
<td></td>
<td>3+SE above</td>
<td>Count</td>
<td>543</td>
<td>620</td>
<td>1434</td>
<td>3342</td>
<td>6471</td>
<td>73690</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6%</td>
<td>.7%</td>
<td>1.7%</td>
<td>3.9%</td>
<td>7.5%</td>
<td>85.6%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Count</td>
<td>15480</td>
<td>7155</td>
<td>9411</td>
<td>12092</td>
<td>15018</td>
<td>86308</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10.6%</td>
<td>4.9%</td>
<td>6.5%</td>
<td>8.3%</td>
<td>10.3%</td>
<td>59.3%</td>
</tr>
<tr>
<td>GR- PREVIOUS YEAR</td>
<td>3+SE below</td>
<td>2 below</td>
<td>1 below</td>
<td>1 above</td>
<td>2 above</td>
<td>3+SE above</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>8 r9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3+SE below</td>
<td>Count</td>
<td>6613</td>
<td>3136</td>
<td>2977</td>
<td>2573</td>
<td>1053</td>
<td>660</td>
<td>17012</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>38.9%</td>
<td>18.4%</td>
<td>17.5%</td>
<td>15.1%</td>
<td>6.2%</td>
<td>3.9%</td>
<td>100.0%</td>
</tr>
<tr>
<td>2 below</td>
<td>Count</td>
<td>726</td>
<td>1038</td>
<td>1542</td>
<td>2394</td>
<td>1370</td>
<td>969</td>
<td>8039</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>9.0%</td>
<td>12.9%</td>
<td>19.2%</td>
<td>29.8%</td>
<td>17.0%</td>
<td>12.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td>1 below</td>
<td>Count</td>
<td>348</td>
<td>497</td>
<td>1044</td>
<td>1967</td>
<td>1493</td>
<td>1315</td>
<td>6664</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>5.2%</td>
<td>7.5%</td>
<td>15.7%</td>
<td>29.5%</td>
<td>22.4%</td>
<td>19.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>1 above</td>
<td>Count</td>
<td>357</td>
<td>673</td>
<td>1645</td>
<td>4052</td>
<td>4424</td>
<td>5745</td>
<td>16896</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>2.1%</td>
<td>4.0%</td>
<td>9.7%</td>
<td>24.0%</td>
<td>26.2%</td>
<td>34.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>2 above</td>
<td>Count</td>
<td>120</td>
<td>222</td>
<td>700</td>
<td>2404</td>
<td>3837</td>
<td>8907</td>
<td>16190</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>.7%</td>
<td>1.4%</td>
<td>4.3%</td>
<td>14.8%</td>
<td>23.7%</td>
<td>55.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>3+SE above</td>
<td>Count</td>
<td>107</td>
<td>150</td>
<td>478</td>
<td>2010</td>
<td>5205</td>
<td>73524</td>
<td>81474</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>.1%</td>
<td>.2%</td>
<td>.6%</td>
<td>2.5%</td>
<td>6.4%</td>
<td>90.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>Count</td>
<td>8271</td>
<td>5716</td>
<td>8386</td>
<td>15400</td>
<td>17382</td>
<td>91120</td>
<td>146275</td>
</tr>
<tr>
<td></td>
<td>% within r9</td>
<td>5.7%</td>
<td>3.9%</td>
<td>5.7%</td>
<td>10.5%</td>
<td>11.9%</td>
<td>62.3%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>