OSSE
DC Office of the
State Superintendent
of Education

October 15, 2008

Kerri L. Briggs, Ph.D.
Assistant Secretary of Elementary and Secondary Education
U.S. Department of Education
400 Maryland Avenue, S.W.
Washington, D.C. 20202-6400

Dear Dr. Briggs:

The District of Columbia is excited to submit for your review the enclosed DC Growth Model Proposal. A growth model will be particularly valuable in the District of Columbia where the majority of students historically have scored well below proficiency. This new approach will better demonstrate the progress that schools, districts, and the state are making toward the goal of 100% proficiency by 2013-2014. Over the past months, the District has received invaluable assistance from several experts in the area of growth and value-added models including U.S. Department of Education peer reviewers.

The District of Columbia student data tracking systems have a long history resulting, in part, from the District's ability to closely monitor enrollment and achievement data. Because of the District's size, on-site internal and external enrollment audits of all public schools have been conducted each October for the last six years. The audits physically track the enrollment of every student enrolled in the DC public schools. In addition, external observers monitor the state assessment administrations in every public school in the District each spring.

As a result, the District is uniquely able to ensure the accuracy of student achievement data across years. Longitudinal achievement data from the District were used in some of the earliest large-scale studies of growth and value-added studies including those by the New American Schools in 2002-2003. In 2007-2008, the District of Columbia was fortunate to receive $5.7 million from an Institute for Education Sciences grant to support the District of Columbia Statewide Longitudinal Data System. The District made the commitment to invest $19 million over five years to further improve the state data systems and create an integrated data warehouse. To match student data over time, the District will rely on the Levenshtein algorithm to validate that students' records were properly merged. The matching algorithm is described in Appendix B of this proposal.

Again, we are excited about the growth model proposal and we look forward to the formal feedback and to working closely with the U.S. Department of Education to continue to improve and refine the model. The District of Columbia remains committed to developing state-of-the-art systems to support high quality data-driven decision- making and improved student achievement.

If you have questions about these submissions, please contact Bill Caritj in the OSSE Division of Assessment and Data Reporting at 202-741-0256 or at bill.caritj@dc.gov.

Sincerely,

Deborah A. Gist
State Superintendent for Education

Attachments

cc: Kimberly A. Statham, Deputy State Superintendent
Susan Rigney, U.S. Department of Education
Patrick Rooney, U.S. Department of Education
William H. Caritj, State Director of Assessment

Submission for the U.S. Department of Education

NCLB Growth Model Pilot Program

Washington DC

October 15, 2008

Table of Contents

Submission for the U.S. Department of Education NCLB Growth Model Pilot Program

District of Columbia
Growth Model Proposal

Introduction

In 2008, the District of Columbia submitted a proposal for implementing a growth model for all schools. The U.S. Department of Education (ED) peer review suggested that the original proposal had two major problems. First, the original model proposed would have required a vertical scale, and at the time the proposal was submitted, that scale was being developed. Second, the peers had concerns regarding the District's ability to merge student records over time.

In the past year, the District has invested significant time in researching these issues. Our research has resulted in a different growth model that has no requirements for a vertical scale and, at the same time, can be used as a model that includes 100% of the tested students—no other state can make this claim. Second, we have identified an empirical method that can be used to reliably merge student records over time. This is a groundbreaking method that we have investigated internally and found to be invaluable as a tool for connecting student scores over time to create a longitudinal data file.

Our proposal is unique in many ways and our plans to include a growth model are based on sound techniques for measuring student growth and for reliably creating the longitudinal data file. The highlights of this proposal include:

- A unique growth toward the standards (GTS) model that forms student projections probabilistically. This prevents us from making claims beyond what the data can support since it is impossible to know if a student is truly "on track" or not. This model differs vastly from the current models implemented for the growth model pilot program.

- A method for merging student records based on the Levenshtein algorithm. This method for merging is a unique quality control procedure that is not in use in any other State education agency.

- Full color, variable information score reports that will be sent to parents communicating the results of the growth model in a clear and transparent way.

These three factors will ensure that a technically rigorous model for growth is implemented and, at the same time, report the results of the growth model in a simple and user-friendly way. The District highly values transparency, and this is evidenced in our proposal and our planned use of the model. That is, we have provided a full technical description of the model for review, technical description of our merging methods, samples of our score reports, and a complete description of how the growth model will be applied for accountability.

However, the idea of transparency differs across audiences. For those who are technically inclined, a complete technical description of the growth model is provided with substantive examples of how it operates. Other audiences, such as parents and teachers, tend to be more interested in how this information can be used to support better classroom practices and student achievement. Therefore, we have made a significant investment in the variable information score reports that portray the results of the growth model in a manner that translates statistical information into information that can be used for instructional planning.

These aspects of our proposal are provided in complete detail in the sections that follow. This proposal is organized as follows. We first provide some background on the State assessment system and our current methods for determining Adequate Yearly Progress (AYP). Subsequently, we indicate how our proposed use of the growth model meets each of the Seven Core Principles.

There are two attachments to this proposal that are heavily relied on and often referenced. The first is a manuscript that brings full transparency to the growth model and how it will be applied in the District. Experiments with the growth model have already occurred using the 2007 to 2008 data for all grades in reading and mathematics, and the results of those analyses are provided for the peer review. That document referenced as Appendix A is a standalone manuscript that brings full transparency to the growth model and its application in the District. Second, we provide a manuscript, referenced as Appendix B, which comprehensively describes the Levenshtein algorithm and how it is applied to form reliable longitudinal data files. We have already worked with this algorithm to assess the degree to which this method is useful for the District and have found that its use exceeds our expectations in terms of joining different yearly data files to form a longitudinal database.

These issues make our proposal extremely valuable for the ED pilot program. To date, no growth model for No Child Left Behind (NCLB) purposes forms projections in the manner we do. Second, this will be the first wide-scale application of the Levenshtein algorithm for merging student records to form a longitudinal data file. This is a significant opportunity to demonstrate how many of the challenges often encountered in tracking students over time can be resolved. Last, no other state has implemented full-color, variable information score reports specifically for their growth model.

The culmination of these issues is a growth model proposal that we believe is well-conceived and practical and can serve as a model for other states given the many unique practices we implement.

Background

State Assessment Background

The District of Columbia implemented new standards-based assessments in reading and mathematics in 2005-2006. Standard-setting for the new assessments was completed in July 2006. In fall 2007, the State assessment system's classification was raised to "approval expected" pending final approval of the technical report for the State alternative assessment. The DC CAS-Alternative Technical Report was submitted in January 2008. Technical reports for the general assessment were completed for both 2006 and 2007 as well as the other technical studies (e.g., validity and reliability) needed for Critical Elements 4 *et seq* of the Standards and Assessment Peer Review.

The District of Columbia data tracking systems have a unique history resulting, in part, from the District's ability to closely monitor enrollment and achievement data. Because of the District's size, on-site internal and external enrollment audits of all public schools have been conducted each October for the last six years. The audits physically track the enrollment of every student enrolled in the DC public schools. In addition, external observers monitor the State assessment reading and mathematics administrations in every public school in the District each spring.

As a result, the District is uniquely able to ensure the accuracy of student achievement data across years. Longitudinal achievement data from the District were used in some of the earliest large-scale studies of growth and value-added studies including those by the New American Schools in 2002-2003.

In the fiscal year 2008 budget, the District of Columbia invested $3 million to create an integrated statewide longitudinal data warehouse. Overall, this investment is $19 million over five years in addition to a three-year $5.7 million grant from ED.

Our Current Accountability Plan

The District currently uses only two methods for making AYP decisions: status and safe harbor. We currently do not use uniform averaging, confidence intervals, or an index system. As in most states, if the District is permitted to use the growth model, it will be applied as the third step in our AYP process after safe harbor.

The results of our proposed growth model will be applied separately for reading and mathematics, they will align with the same Annual Measurable Objectives (AMOs) established for status, and all students will be expected to be proficient in 2014.

The Seven Core Principles and the Washington DC Growth Model Program

Core Principle 1: 100% Proficiency by 2014 and Incorporating Decisions about Student Growth into School Accountability

"The accountability model must ensure that all students are proficient by 2013-14 and set annual goals to ensure that the achievement gap is closing for all groups of students."

1.1 How does the State accountability model hold schools accountable for universal proficiency by 2013-14?

The State will maintain the same AMO and intermediate steps that were approved in the August 2006 revision of the State Accountability Plan resulting in the goal of universal proficiency in 2013-14. These objectives apply to the State, Local Educational Agencies (LEAs), and schools. The current status model and safe harbor determinations will be applied first in all cases. The growth model determinations will apply after these determinations are applied.

The exact same AMOs used for status will be used for the growth model; hence, this model will also require 100% of DC students to be proficient in 2013-2014. In many previously submitted plans, some percentage of students might not be proficient in 2014 but instead might be on track to proficiency in 2017. Our model does not operate in that manner. We view 2014 as the end of the timeline, and schools are held accountable for these goals until legislative changes permit otherwise. The methods by which the AMOs are applied to the DC growth model are described under Core Principle 1.3.

1.2 Has the State proposed technically and educationally sound criteria for "growth targets" for schools and subgroups?

The model proposed is a GTS model. As such, the growth target for every student is the same: proficiency. Our model differs from most submitted GTS models in that our model asks, "What percentage of students is on track to proficiency **next year**?" Many submitted models allow for students to be on a three- or even four-year trajectory. In many respects, allowing such a long timeline precludes educators from having a sense of urgency regarding student achievement.

As comprehensively described in Appendix A, our model generates the probability that each student will be proficient in the next school year. In theory, low probabilities are designed as a call to action leading parents and teachers to act to improve a student's performance. For example, if a parent or teacher learned that his or her student has only a 28% chance of being proficient next year, and if this probability were accompanied with some diagnostic information regarding that student's test performance, then parents or educators can act on that information for the benefit of the student.

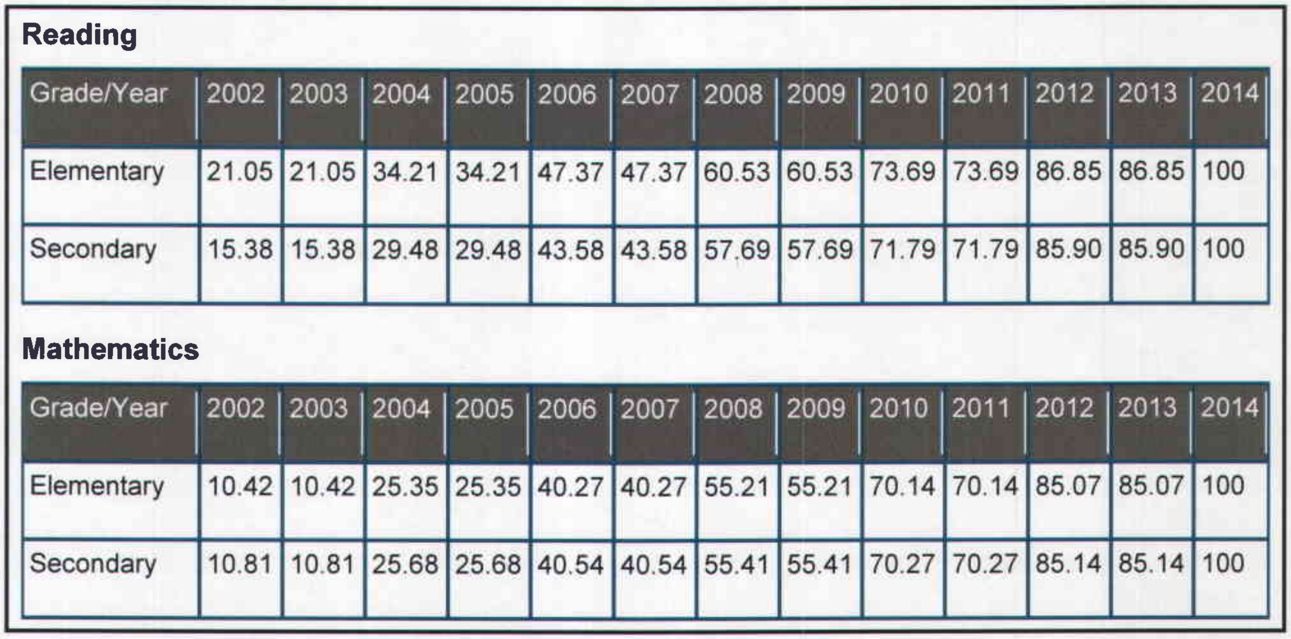
In fact, this is exactly the theory of action our model follows. We derive each student's probability of becoming proficient in the next school year. Subsequently, this statistical information is conveyed to parents and educators using the variable information score reports described under Core Principle 5.2 to provide diagnostic information and improve student performance.

1.3 Has the State proposed a technically and educationally sound method of making annual judgments about school performance using growth?

Under the proposed method, State, district, and school AYP determinations would first follow all rules currently employed under the District's accountability plan. The current status model and safe harbor determinations will be applied first in all cases. The growth model determinations will apply after these determinations are applied. The growth model determinations will be used as a final control to decrease false negatives (e.g., schools on a trajectory to proficiency that do not achieve safe harbor).

The table below shows annual measurable objectives for reading and mathematics from 2001-2002 to 2013-2014. Intermediate goals have been set to measure whether schools make AYP toward meeting the goal of 100% proficiency by 2014 as is called for in the NCLB legislation. As shown in the table, the goals for the percentage of students who must achieve proficiency rise every other year.

Table 1:
AYP: Annual Measurable Objectives for Reading and Mathematics
(Percentage Scoring at Proficient or Above Level)



Reading

Grade/Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Elementary	21.05	21.05	34.21	34.21	47.37	47.37	60.53	60.53	73.69	73.69	86.85	86.85	100
Secondary	15.38	15.38	29.48	29.48	43.58	43.58	57.69	57.69	71.79	71.79	85.90	85.90	100

Mathematics

Grade/Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Elementary	10.42	10.42	25.35	25.35	40.27	40.27	55.21	55.21	70.14	70.14	85.07	85.07	100
Secondary	10.81	10.81	25.68	25.68	40.54	40.54	55.41	55.41	70.27	70.27	85.14	85.14	100

In the District, AMOs were determined using the method prescribed in the NCLB legislation and subsequent regulations. Using stakeholder input, the decision was made to have two-year increases in proficiency goals on the way to universal (100%) proficiency. The growth proposal takes a similar approach by aligning the growth targets to the established proficiency goals (i.e., AMOs). Each year, schools will be required to grow at a rate that will at least follow the trajectory to reach 100% proficiency by 2013-2014.

Since the growth model yields the percentage of the students likely to be proficient in year $t+1$, given the observed scores for students in school j in year t, this percentage can be compared to the State AMO for the following year to make AYP decisions.

For example, under the current status model, if 50% of School A's (an elementary school) students are proficient or above in reading in 2009, School A will not meet AYP for 2009 because the AMO is 63.5%.

Safe harbor determinations will next be used to establish how School A is performing. If School A does not achieve safe harbor, the growth model will be applied.

Suppose the growth model predicts that 80% of students in School A are on track to be proficient next year (i.e, we are projecting that 80% will be proficient in 2010). This projected number is compared to the 2010 AMO, which is 73.69%. In this case, School A would make AYP under growth.

This method is proposed because it is internally consistent with what the projections are stating about a school. That is, the projections form an estimate of next year's performance. Logic dictates that this projected percentage should be compared to next year's AMO and not the current AMO.

The same methods and calculations will be used to make annual judgments about the performance of subgroups and the District. Appendix A shows how the projected percentage of students on track to proficiency for all students and each student group is determined.

Core Principle 2: Establishing Appropriate Growth Targets at the Student Level

"The accountability model must establish high expectations for low-achieving students, while not setting expectations for annual achievement based upon student demographic characteristics or school characteristics."

2.1 Has the State proposed a technically and educationally sound method of depicting annual student growth in relation to growth targets?

The proposed District growth model, as detailed in Appendix A, depicts a student's likelihood of becoming proficient in the following school year. Technically, a projection can only be made probabilistically. Many of the growth models submitted extrapolate a student's future score and make the claim that the student is "on track." From a technical perspective, this is indefensible. There are at least two sources of error that would confound that estimate: projection error and measurement error.

Our model only forms a projection for students and their likelihood of future proficiency because these are the only claims the data can actually support. While this is statistically appropriate, we do this for a second and more important reason. If our model claimed that a student was "on track" to be proficient next year but he or she does not achieve proficiency, educators and parents will tend not to rely on the model since it will appear on its face to make false claims regarding the potential future status of student achievement.

The District model is a regression model that does not bring in any demographic information regarding students. The model uses the most current year's level of proficiency as the outcome variables and the prior year scaled score as the covariate.

When forming projections, every tested student in grades 3 to 7 is included in the model, both students scoring above and those scoring below proficiency. In addition, each student is held to the same high expectation of proficiency. If a student has a single test score in grades 3 to 7, then a projection is made for that student so that the model is applied to all grades 3-8.

The growth model is proposed only for students in grades 3-8 and will not include students in high school. Because of the way our testing system is designed, we do not measure students in grade 9. Forming projections from grade 8 to grade 10 is very difficult because it spans a large time period. It can be easily done statistically, but from a substantive (educational) perspective, we question whether such estimates are meaningful. This does not preclude the high schools from being included in the State Accountability Plan. Indeed, they are included in status and safe harbor. But, it does preclude the high schools from having the additional benefit of the growth model results.

Core Principle 3: Accountability for Reading/Language Arts and Mathematics Separately

"The accountability model must produce separate accountability decisions about student achievement in reading/language arts and in mathematics."

3.1 Has the State proposed a technically and educationally sound method of holding schools accountable for student growth separately in reading/language arts and mathematics?

As with status calculations, separate growth determinations would be completed for reading and mathematics. Examples of how our AYP decisions are made are provided under Core Principle 1.3.

Core Principle 4: Inclusion of All Students

"The accountability model must ensure that all students in the tested grades are included in the assessment and accountability system. Schools and districts must be held accountable for the performance of student subgroups. The accountability model, applied statewide, must include all schools and districts."

4.1 Does the State's growth model proposal address the inclusion of all students, subgroups and schools appropriately?

The description of our model in Appendix A shows how every student with a test score is included in forming the projections used to make the AYP decisions. To be clear, every student included in the DC-CAS status model is also included in the growth model projections. Our model as proposed is the only growth model that can make this claim, as all other State proposals would exclude some students if they had patterns of missing scores.

This is possible because the conditional probability of future success is obtained by examining growth of a cohort of students and then applying those probabilities to the current student group.

All NCLB subgroups are also reported. The method by which the subgroups are included is described in Equation 6 of Section 4 in Appendix A.

The growth model, participation, status, and safer harbor determinations would be calculated independently. As in the status model, all groups would be required to meet the 95% participation rate. If a group does not achieve the 95% participation rate criterion, this group will not achieve AYP regardless of the status or growth model determinations.

Core Principle 5: State Assessment System and Methodology

"The State's NCLB assessment system, the basis for the accountability model, must include annual assessments in each of grades three through eight and high school in both reading/language arts and mathematics, must have been operational for more than one year, and must receive approval through the NCLB peer review process for the 2005-06 school year. The assessment system must also produce comparable results from grade to grade and year to year."

5.1 Has the State designed and implemented a statewide assessment system that measures all students annually in grades 3-8 and one high school grade in reading and mathematics in accordance with NCLB requirements for 2005-06, and have the annual assessments been in place since the 2004-05 school year?

In 2006-2007, the District of Columbia implemented a new system of standards-based assessments in reading and mathematics called the District of Columbia Comprehensive Assessment System (DC-CAS). This is a criterion-referenced test made up of constructed-response and multiple-choice questions and is based on the

District of Columbia standards. Accommodations for the DC-CAS such as accommodations for English language learners and an alternate assessment (DC-CAS Alternate Assessment) for students with IEPs have also been developed.

All students in grades 3-8 and in grade 10 participate in the assessment and are tested in reading and mathematics as well as composition in grades 4, 7, and 10; science in grades 5 and 8; and biology in grades 9 through 12 (if students took a biology class).

The status and safe harbor systems required by NCLB and presented in the State Accountability Plan will continue to apply to all students in grades 3-8 and 10. The proposed growth model as described in Appendix A will be applied to grades 3-8.

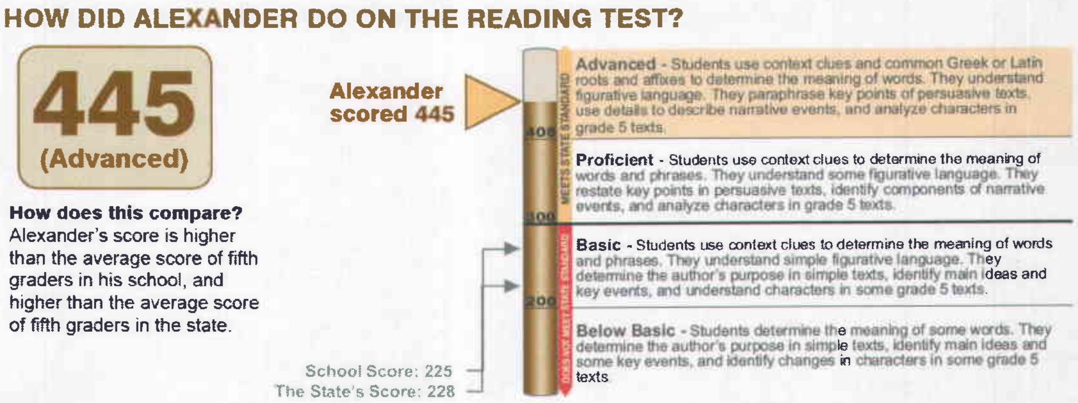
5.2 How will the State report individual student growth to parents?

The State will develop transparent, clear, and easily understood student score reports for the families of students taking the DC-CAS. These reports will contain both status and growth information. To develop the reports, we will begin by asking what actions we would want parents to take using the information presented in the reports. We will then develop graphical displays and text that will not only present data but also offer instructional recommendations that parents may be able to use to improve their students' learning during the following year.

These reports will be fully customized, enhanced color booklets. They will present data on how students performed on the DC-CAS in the current and previous years and provide families with the likelihood that their student will be proficient in the following year. Reports will be developed using technology that allows for complete variability and will provide each family with fully customized information on how their student is performing on specific content areas and the next steps they can take to help their student improve in the strands they are struggling with.

Figure 1 provides an example of how we may provide families with status information. The student's scale score and performance level are clearly highlighted using either a yellow (proficient) or red (below proficient) arrow, and the chart indicates how the child scored in the range of all possible scores. Descriptions of each performance level are provided in family-friendly language that allows families to see what is expected at each level of achievement. The school and State average scores are also shown so families are able to see how their student's score compared to the average of all students taking the assessment. The language under the chart further explains how the student performed and compares the student's score with the school and State averages.

Figure 1



HOW DID ALEXANDER DO ON THE READING TEST?

445 (Advanced)

How does this compare? Alexander's score is higher than the average score of fifth graders in his school, and higher than the average score of fifth graders in the state.

Alexander scored 445

Advanced - Students use context clues and common Greek or Latin roots and affixes to determine the meaning of words. They understand figurative language. They paraphrase key points of persuasive texts, use details to describe narrative events, and analyze characters in grade 5 texts.

Proficient - Students use context clues to determine the meaning of words and phrases. They understand some figurative language. They restate key points in persuasive texts, identify components of narrative events, and analyze characters in grade 5 texts.

Basic - Students use context clues to determine the meaning of words and phrases. They understand simple figurative language. They determine the author's purpose in simple texts, identify main ideas and key events, and understand characters in some grade 5 texts.

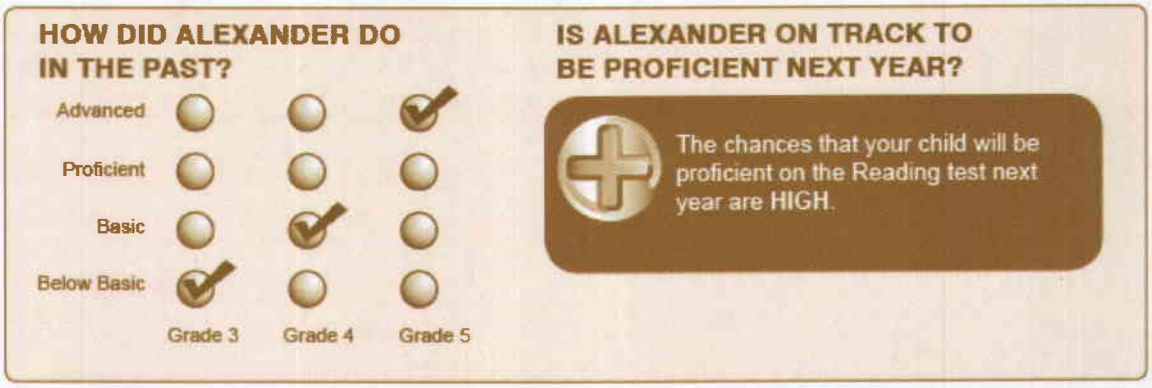
Below Basic - Students determine the meaning of some words. They determine the author's purpose in simple texts, identify main ideas and some key events, and identify changes in characters in some grade 5 texts

School Score: 225

The State's Score: 228

Figure 2 is an example of how we may show how students performed in previous years and their likelihood of reaching proficiency in the following year. This graphic displays the level at which the student performed in reading in all the years that he or she participated in the DC-CAS. The text on the right explains the chances the student will be proficient in the following year.

Figure 2



HOW DID ALEXANDER DO IN THE PAST?

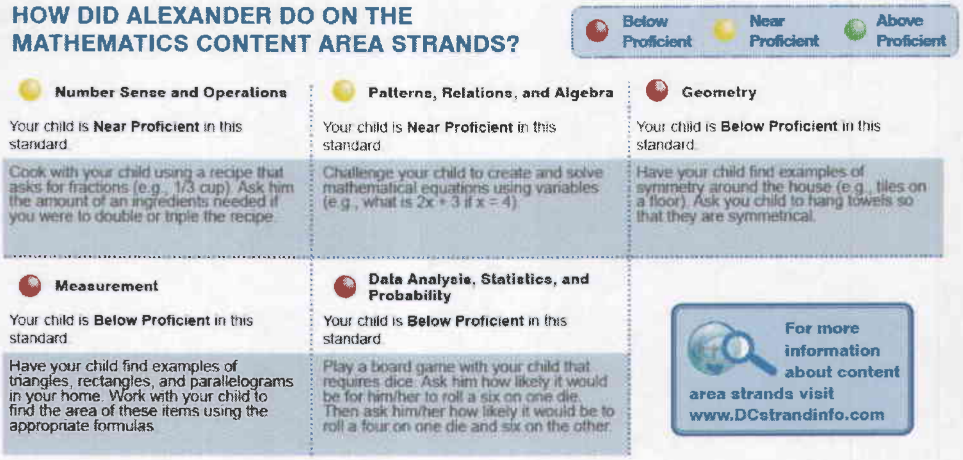
IS ALEXANDER ON TRACK TO BE PROFICIENT NEXT YEAR?

The chances that your child will be proficient on the Reading test next year are HIGH.

In the statistical model, we form specific probabilities. However, when communicating this information to parents, we generalize those probabilities in a manner that makes them more user-friendly. Instead of stating, "your child has a 31% probability of being proficient next year," we trichotomize those probabilities using the Wald confidence intervals. The three categories are low, average and high. Hence, if a student clearly has less than a 50% chance of being proficient, the language on the report will state that the student has a low chance. If the student clearly has better than a 50% chance, then the report will state the student has a high chance. If the probability cannot be statistically distinguished from 50%, then the report will state that the student has an average chance.

Parents could use this information to see how their student has progressed and how their student is projected to perform in the following year. This is designed to motivate parents to take actions to improve their student's chances of being proficient the following year. Figure 3 is provided so that parents can get some diagnostic information regarding their student from the DC-CAS. This is an example of how we may include data on how students performed on each content area strand so parents are able to see whether their student performs well and in which areas they may need some instructional remediation so that they can improve their chances of being proficient next school year. The colored circle shows how the student performed (below, near, or above proficient) on each reporting strand and the text below suggests activities parents could easily do with their students to improve their learning. The text is specifically written to help the student based on his or her performance. That is, the text is variable and is driven by the scores for each student. Different patterns of strengths and weaknesses would result in different instructional recommendations appearing on the reports.

Figure 3



HOW DID ALEXANDER DO ON THE MATHEMATICS CONTENT AREA STRANDS?

● Below Proficient ● Near Proficient ● Above Proficient

● **Number Sense and Operations**

Your child is **Near Proficient** in this standard.

Cook with your child using a recipe that asks for fractions (e.g., 1/3 cup). Ask him the amount of an ingredients needed if you were to double or triple the recipe.

● **Patterns, Relations, and Algebra**

Your child is **Near Proficient** in this standard.

Challenge your child to create and solve mathematical equations using variables (e.g., what is 2x + 3 if x = 4).

● **Geometry**

Your child is **Below Proficient** in this standard.

Have your child find examples of symmetry around the house (e.g., tiles on a floor). Ask you child to hang towels so that they are symmetrical.

● **Measurement**

Your child is **Below Proficient** in this standard.

Have your child find examples of triangles, rectangles, and parallelograms in your home. Work with your child to find the area of these items using the appropriate formulas.

● **Data Analysis, Statistics, and Probability**

Your child is **Below Proficient** in this standard.

Play a board game with your child that requires dice. Ask him how likely it would be for him/her to roll a six on one die. Then ask him/her how likely it would be to roll a four on one die and six on the other.

For more information about content area strands visit www.DCstrandinfo.com

The State may also choose to provide families with information on how their school is performing in comparison with other schools in the State. Using the calculations described in Appendix A, we could determine whether a school is a high or low performing school and whether it is expected to have high or low growth in the following year. This information would guide parents to hold discussions with school administrators on their school's curricular and instructional plans.

Images and descriptions provided in this proposal are examples of what we may choose to include in the reports as we are still in the design phase. The final design of the reports will be vetted with parents and key stakeholders in the District. Through focus groups and meetings, we will determine the status and growth information that is most useful for parents and ensure all language used in the reports is easily understandable.

5.3 Does the Statewide assessment system produce comparable information on each student as he/she moves from one grade level to the next?

Yes, all performance categories (e.g., basic, proficient) were vertically articulated during the standard setting process for both reading and mathematics. There is currently no vertical scale, but this is not needed for the model as described in Appendix A.

5.4 Is the Statewide assessment system stable in its design?

Yes. The District of Columbia implemented new standards-based assessments in reading and mathematics in 2005-2006. Standard-setting for the new assessments was completed in July 2006. In fall 2007, the State assessment system's classification was raised to "approval expected" pending final approval of the technical report for the State alternative assessment and full approval is expected in spring 2008.

Core Principle 6: Tracking Student Progress

"The accountability model and related State data system must track student progress."

6.1 Has the State designed and implemented a technically and educationally sound system for accurately matching student data from one year to the next?

The District will rely on the Levenshtein algorithm to merge and validate that students were properly merged over time. The matching algorithm is completely described in Appendix B of this proposal. That paper also shows how remarkably reliable the merges are.

The District first merges the year 1 and the year 2 data files using the unique student identifier. The Levenshtein algorithm is then used to validate that the merge using ID was performed correctly. Those students retained in the data must meet the following criteria:

- The unique student ID is the same in year 1 and year 2; and
- The Levenshtein normalized distance is greater than or equal to 0.7.

In our review of the data, all students with an LND < 0.4 are incorrect student merges, even though they share the same ID. By manually looking at the names, we can see that some incorrect merges occur. In the range of 0.4 to 0.7, there is some ambiguity in the merge; most of the names reveal incorrect merges and there are some questionable merges.

However, every student with an LND >= 0.7 is clearly the same student. We therefore chose LND = 0.7 as the cutoff point. In our view, the risk of joining incorrect records is greater than the risk of gaining a few correct merges alongside many incorrect merges, which would occur if the cutpoint for the LND were set any lower than 0.7.

Tables 2 and 3 show the number and percentage of students for reading and mathematics who were correctly merged from this process using data from 2007 to 2008. In all grades, the merge rates are high, with the lowest merge of 88.8% occurring for grade 7.

Table 2

Mathematics Table of mergeflag by ESTGRADE2						
mergeflag	ESTGRADE2 (Grade Enrolled)					Total
Frequency Col Pct	4	5	6	7	8	
Not Merged	481 10.50	468 10.18	498 10.71	512 11.17	492 9.96	2451
Merged	4101 89.50	4130 89.82	4154 89.29	4070 88.83	4449 90.04	20,904
Total	4582	4598	4652	4582	4941	23,355

Table 3

Reading Table of mergeflag by ESTGRADE2						
mergeflag	ESTGRADE2 (Grade Enrolled)					Total
Frequency Col Pct	4	5	6	7	8	
Not Merged	478 10.43	467 10.16	494 10.63	510 11.14	490 9.92	2439
Merged	4103 89.57	4131 89.84	4155 89.37	4069 88.86	4448 90.08	20,906
Total	4581	4598	4649	4579	4938	23,345

Tables 4 and 5 show the merge rates by ethnicity. Again, the merge rates are remarkably high. One merge rate for category "I" appears low, but this is an artifact of the small *N* size for that category.

Table 4

Mathematics Table of mergeflag by ETHNICITY2

mergeflag	ETHNICITY2 (Ethnicity)					Total
Frequency Col Pct	A	B	H	I	W	
Not Merged	48 15.38	2041 10.30	256 12.04	2 50.00	102 9.28	2449
Merged	264 84.62	17,770 89.70	1871 87.96	2 50.00	997 90.72	20,904
Total	312	19,811	2127	4	1099	23,353

Frequency Missing = 2

Table 5

Reading Table of mergeflag by ETHNICITY2

mergeflag	ETHNICITY2 (Ethnicity)					Total
Frequency Col Pct	A	B	H	I	W	
Not Merged	46 14.84	2038 10.29	250 11.79	2 50.00	101 9.20	2437
Merged	264 85.16	17,773 89.71	1870 88.21	2 50.00	997 90.80	20,906
Total	310	19,811	2120	4	1098	23,343

Frequency Missing = 2

These are the merge rates that occur using the unique student ID and validating using the LND statistic. However, we can improve these merge rates and increase the number of students that are merged from one year to the next as described in Section 6 of the paper in Appendix B. Our method for doing so is as follows:

- Concatenate the first three letters of the first name and the first three letters of the last name to create string 1 and string 2 variables. Merge the year 1 and year 2 files based on the string 1 and string 2 variables.
- Validate the merge using the Levenshtein normalized distance. For the validation, we use a very stringent set of criteria since this is a "salvage" effort and no unique student IDs are available. Therefore, the string 1 and string 2 variables are a concatenation of first name, last name, grade level, and school attended. Because the Levenshtein algorithm compares similar strings, we subtract 1 from the year 2 grade level so it would match the year 1 grade level.
- We retain only those students if the Levenshtein normalized distance is greater than or equal to 0.9.

After applying this procedure, we are able to recover 281 additional students to our data set. With the addition of these students, the frequencies are provided in Tables 6 through 9. In all cases, this brings the aggregate merge rate for all grades to 90% or better.

Table 6

Mathematics Table of mergeflag by ESTGRADE2						
mergeflag	ESTGRADE2 (Grade Enrolled)					Total
Frequency Col Pct	4	5	6	7	8	
Not Merged	425 9.28	420 9.13	446 9.59	472 10.30	420 8.50	2183
Merged	4157 90.72	4178 90.87	4206 90.41	4110 89.70	4521 91.50	21,172
Total	4582	4598	4652	4582	4941	23,355

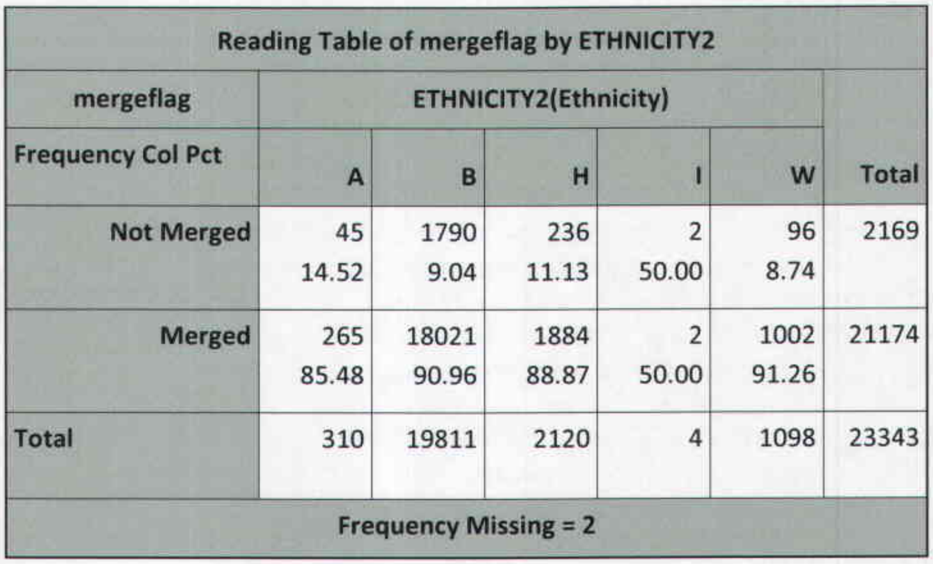
Table 7

Reading Table of mergeflag by ESTGRADE2						
mergeflag	ESTGRADE2 (Grade Enrolled)					Total
Frequency Col Pct	4	5	6	7	8	
Not Merged	422 9.21	419 9.11	442 9.51	470 10.26	418 8.46	2171
Merged	4159 90.79	4179 90.89	4207 90.49	4109 89.74	4520 91.54	21,174
Total	4581	4598	4649	4579	4938	23,345

Table 8

Mathematics Table of mergeflag by ETHNICITY2						
mergeflag	ETHNICITY2 (Ethnicity)					Total
Frequency Col Pct	A	B	H	I	W	
Not Merged	47 15.06	1793 9.05	242 11.38	2 50.00	97 8.83	2181
Merged	265 84.94	18,018 90.95	1885 88.62	2 50.00	1002 91.17	21,172
Total	312	19,811	2127	4	1099	23,353
Frequency Missing = 2						

Table 9



Reading Table of mergeflag by ETHNICITY2

mergeflag	ETHNICITY2(Ethnicity)					Total
Frequency Col Pct	A	B	H	I	W	
Not Merged	45 14.52	1790 9.04	236 11.13	2 50.00	96 8.74	2169
Merged	265 85.48	18021 90.96	1884 88.87	2 50.00	1002 91.26	21174
Total	310	19811	2120	4	1098	23343

Frequency Missing = 2

Our method of merging is state of the art and demonstrates for other states how merges, validation of those merges, and salvage efforts for students lacking IDs can be included in the data. This is the first application of the Levenshtein algorithm to large-scale educational measurement and its demonstration as a part of this growth model pilot program will be very beneficial.

6.2 Does the State data infrastructure have the capacity to implement the proposed growth model?

Yes. First, the merge rates are very high, allowing us to measure growth in order to generate the conditional probabilities very reliably. However, as described in the manuscript in Appendix A, the projections are made for every student in the data irrespective of whether they have a prior test score or not. Because of this, the District is able to implement the proposed model very reliably.

Second, the District has partnered with AIR to implement all technical procedures outlined in this paper. We have a multi-year comprehensive contract with the District to implement this growth model and the score reports, merge the data sets, and perform many other tasks necessary to maintain an operational testing program.

Core Principle 7: Participation Rates and Additional Academic Indicator

The accountability model must include student participation rates in the State's assessment system and student achievement on an additional academic indicator.

7.1 Has the State designed and implemented a statewide accountability system that incorporates the rate of participation as one of the criteria?

Yes, the District will continue to hold the District and schools independently accountable for the NCLB-mandated 95% participation rate.

7.2 Does the proposed State growth accountability model incorporate the additional academic indicator?

Yes, the State will continue to hold the State, districts, and schools independently accountable for the NCLB-mandated other academic indicator – graduation rate for high schools and attendance for all other schools.

Summary

The District is excited about the opportunity to implement this growth model. We view this as a unique opportunity to strengthen our accountability plans and, at the same time, provide information to parents and educators that can be used to spur improvements in student achievement.

Appendix A

Technical Details on the Implementation of the Growth Model in Washington, DC

Harold C. Doran
American Institutes for Research
Washington, DC
hdoran@air.org

Working Draft
Not for Distribution or Citation

October 7, 2008

Abstract

Longitudinal analyses that measure growth towards a standard have become common applications for the U.S. Department of Education's growth model pilot program. However, many of the models implemented make claims beyond what the data can support by assuming a student's path to proficiency is fixed and known with certainty. In this paper, we present a method that forms the conditional probability that a student will become proficient and demonstrate how these probabilities can be used to make adequate yearly progress decisions.

Keywords: longitudinal analysis; growth to standard

1 Background

Originally, the District of Columbia (the District) proposed a model for the U.S. Department of Education (ED) growth model pilot program that depended on a vertical scale. Subsequent studies have illustrated limitations in that scale for measuring student progress. This document provides details on the implementation of a different growth model that does not require a vertical scale. The primary purpose of this model is to estimate the percentage of students on track to proficiency for a given school. Additionally, it may be useful for school evaluation (i.e., value-added).

The class of growth models commonly implemented for NCLB-AYP decisions, sometimes referred to as growth towards the standards models (GTS), as well as value-added models (VAM), tend to be concerned with the following questions regarding students and schools:

1. Given the observed scores for students in school j in year t, what percentage of these students are likely to be proficient in year $t+1$?

2. Given the observed performance of student i in school j at time t, what is the probability that he will be proficient in year $t+1$?

3. How does the performance of students in school j compare to the performance of school j' where $j \neq j'$

For all intents and purposes, NCLB is concerned with the first question regarding school-level performance, classroom educators and parents tend to be concerned with the second question, and school administrators and policymakers are often concerned with the third question. A useful model would attempt to answer all questions at the same time in a reasonably simple, yet reliable way.

This paper presents a statistical model that is designed to provide different audiences/users of data with answers to these questions. The assumptions of the model, a technical description, as well as examples are provided to bring full transparency to the topic.

2 Technical Details of the Model

2.1 Assumptions

In principle, every student in grade $g = (3, \ldots, 7)$ has a chance of being proficient in grade $g+1$. Those chances tend to be improved when they attend a school with a record for improving student performance and when the student's grade g score is high. Formally, this "change" probability can be defined as $p_{g \rightarrow g+1,i}$. This is the probability that a student in grade g will be proficient in grade $g+1$. The probability enjoys the same properties as all probabilities. The two properties of probabilities made use of in the model proposed include:

- $0 \leq p_{g \rightarrow g+1,i} \leq 1 \ \ \forall \ \ i$
- $\mathcal{E}(\gamma) = \sum_i p_{g \rightarrow g+1,i}$

The first property denotes that the probability is is bounded between 0 and 1. The second property implies that the sum of the individual probabilities forms the expected number of students on track to proficiency.

There are three assumptions that motivate the use of the model that is subsequently proposed.

Assumption 1: Many of the growth towards the standards models proposed for the ED growth model pilot program project future student performance, conditional on prior performance, as a fixed score and known with certainty. This is improper as future success cannot be accurately projected, but only estimated with some probability.

Assumption 2: The probability of success in grade $g+1$ depends on the student's grade g score. Student's with high scores in grade g are likely to have a high chance of proficiency in grade $g+1$ whereas students with low scores in grade g are likely to have a low chance.

Assumption 3: Schools have a differential impact on student performance. That is, some schools are more effective with their students than others in terms of growth. As a consequence, two students with the same scores in grade g, but attending different schools are likely to have different probabilities of future success as a function of differences in the instructional program.

Given these assumptions, the aim of the model is to generate an estimate of $p_{g \to g+1,i}$ for students in school j with an observed scaled score in grade g of $\hat{\theta}$. Given that we are now interested in conditioning on these two factors (i.e., scaled scores and schools), let the conditional probability of future proficiency be defined as:

- $p(\hat{\theta}_{(j)i})_{g \to g+1}$ = The probability that student i in school j with a scaled score of $\hat{\theta}$ in grade g will be proficient in grade $g+1$.

We compute this probability for all student scores along the ability scale whether they were proficient in grade g or not. This is done because it is unreasonable to assume a student scoring at or above proficient in grade g will always be proficient in grade $g+1$. Thus, the aim is to include all students, not only those scoring below proficiency.

2.2 Implementation

The estimate of $p(\hat{\theta}_{(j)i})_{g \to g+1}$ is derived from a cohort of students with longitudinally linked test scores moving from grade g to $g+1$. For instance, using data from the spring of 2008 we obtain:

$$y_i = \begin{cases} 1 \text{ if student } i \text{ was proficient or advanced in 2008} \\ 0 \text{ otherwise} \end{cases}$$

We can now regress the observed proficiency status in grade $g+1$ (captured as y_i) on the continuous covariate, $\hat{\theta}_{gi}$, which is their observed scaled score in grade g (spring 2007). The following linear predictor with a school random effect would first be implemented using a generalized linear mixed model (McCulloch & Searle, 2002; Pinheiro & Bates, 2000):

$$\eta = \mu + \beta(\hat{\theta}_{gi} - \theta_c) + \nu_j, \nu_j \sim N(0, \sigma^2) \tag{1}$$

where μ is the average log-odds of success for a student with a scaled score equal to the proficiency cutpoint, $\hat{\theta}_{gi}$ is the observed scaled score for student i in grade g, θ_c is the lower bound cutscore for proficiency, β is the effect of score $\hat{\theta}_{gi}$, and ν_j is the random effect for school j. We include the school level random effect to account for the different instructional experiences likely to occur in different schools. For instance, two students with the same $\hat{\theta}_{gi}$, but attending different schools are likely to have a different probability of success in grade $g+1$. Hence, the random effect will account for this difference. In sum, this form of the linear predictor operationalizes all of the assumptions made explicit in Section 2.1.

With the estimates from Equation (1) in hand, we can now estimate the probability of interest, $p(\hat{\theta}_{(j)i})_{g \to g+1}$ via the following non-linear transformation of the log-odds to the probability scale:

$$p(\hat{\theta}_{(j)i})_{g \to g+1} = [1 + \exp(-\eta)]^{-1} \tag{2}$$

2.3 Wald Confidence Intervals

The model as proposed for NCLB reasons does not make use of confidence intervals. However, in the score reports generated for parents, we attempt to communicate the uncertainty in the estimated probabilities. Because confidence intervals are not allowed for the NCLB growth model pilot program, these are excluded from being used in all calculations related to AYP.

The confidence intervals for each of the probabilities are obtained using the Wald confidence intervals on the original logit scale and then transformed to the probability scale. The lower and upper bound estimates are obtained as from the variance of the linear combination as:

$$var(\eta) = var(\mu) + b^2 \times var(\beta) + 2 \times b \times cov(\mu, \beta) + var(\nu_j) \tag{3}$$

where $b = \hat{\theta}_{gi} - \theta_c$. Upper and lower bound confidence intervals on η on the logit scale are

$$\begin{aligned} \eta_{lower} &= \eta - 1.96\sqrt{var(\eta)} \\ \eta_{upper} &= \eta + 1.96\sqrt{var(\eta)} \end{aligned} \tag{4}$$

The estimates of η_{upper} and η_{lower} are then transformed to the probability scale via Equation 2.

2.4 Generation and Application of Conditional Probabilities

The generation and application of the probabilities is relatively straightforward and proceeds in two steps.

Step 1 : Estimate $p(\hat{\theta}_{(j)i})_{g \rightarrow g+1}$ using data from a cohort with a full compliment of data. A full compliment is defined as two years of longitudinally linked test scores.

Step 2 : Apply the estimates of $p(\hat{\theta}_{(j)i})_{g \rightarrow g+1}$ to the group of interest to make future projections.

Table 1 illustrates how this process operates. The green cells highlight a cohort of students that moved from grade 3 in 2007 to grade 4 in 2008. The parameters in Equation (1) are estimated on the basis of all students in this cohort with longitudinally linked scores. Additionally, the probability estimates from Equation (2) are generated from this same cohort.

For NCLB purposes, the interest is in making a prediction for the grade 3 students in 2008 (light blue cell). That is, the desired inference is the number of grade 3 students in 2008 that are expected to score at the proficiency level in 2009. Therefore, the estimates of $p(\hat{\theta}_{(j)i})_{g \rightarrow g+1}$ generated from the 2007/2008 cohort moving from grade 3 to 4 are applied to the grade 3 2008 student group to make projections regarding their 2009 performance. The method by which the estimates of $p(\hat{\theta}_{(j)i})_{g \rightarrow g+1}$ are applied to make AYP decisions is detailed in Section 4.2.

In sum, the estimates of the conditional probabilities are garnered from the observed growth of students progressing from grade 3 to 4 between 2007 and 2008. Those estimates are then applied to grade 3 students in 2008 to form their 2009 projections.

Table 1: Generation and Application of $p(\hat{\theta}_{(j)i})_{g \to g+1}$

2007	2008	2009
3	3	
4	4	4

3 Instructional Differences Across Schools

One of the assumptions made explicit in Section 2.1 is that schools have a differential impact on student performance as a function of instructional quality. This school effect is captured via the conditional mode of the random effect, ν_j, from Equation (1).

There are two ways to visually display these data to illuminate the fact that schools have a clear differential impact on student achievement. Figure 1 shows three logistic curves. The curve furthest to the left on the x-axis is the highest performing school in the District in terms of math growth from grades 3 to 4. The curve in the middle is the average statewide performance and the curve furthest to the right is the lowest performing school in the District in terms of math growth from grades 3 to 4.

These curves show that a student with a scaled score exactly at the proficient cutscore in grade 3 math (0 on the x-axis due to centering) has less than a 20% chance of being proficient in grade 4 if they attend Brightwood Elementary School. However, a student with the same scaled score, but attends Langdon Elementary School, has better than a 90% chance of being proficient in grade 4. Clearly, the impact on students, after controlling for initial status, varies between schools.

Figure 1 shows only two schools and how they compare to the statewide average performance. A second way to display these data is via a caterpillar plot. Figure 2 shows the variation in school effectiveness in terms of growth for all schools in the District. In this plot, the dark blue dot is the conditional mode of the random effect, ν_j, and the symmetric vertical bars around the dot are the 95% confidence intervals. The horizontal line extending from 0 across the plot is the mean performance of the state. When a school's error bars intersect with the horizontal line, it is not possible to distinguish this school as being either high or low performing vis-a-vis the performance of students at a typical school in the District. Schools that are lower than the horizontal line, and the error bars do not intersect with horizontal line, are clearly lower performing school with respect to the typical school. Schools that are higher than the horizontal line, and the error bars do not intersect with horizontal line, are clearly higher performing schools with respect to the typical school.

4 Use in Educational Accountability Systems: Making AYP Decisions

Once the estimates of $p(\hat{\theta}_{(j)i})_{g \to g+1}$ are obtained, we can immediately apply them to make a probabilistic statement regarding each student's likelihood of being proficient in the following school year. For example, in lay terms (using only a hypothetical example) we might say:

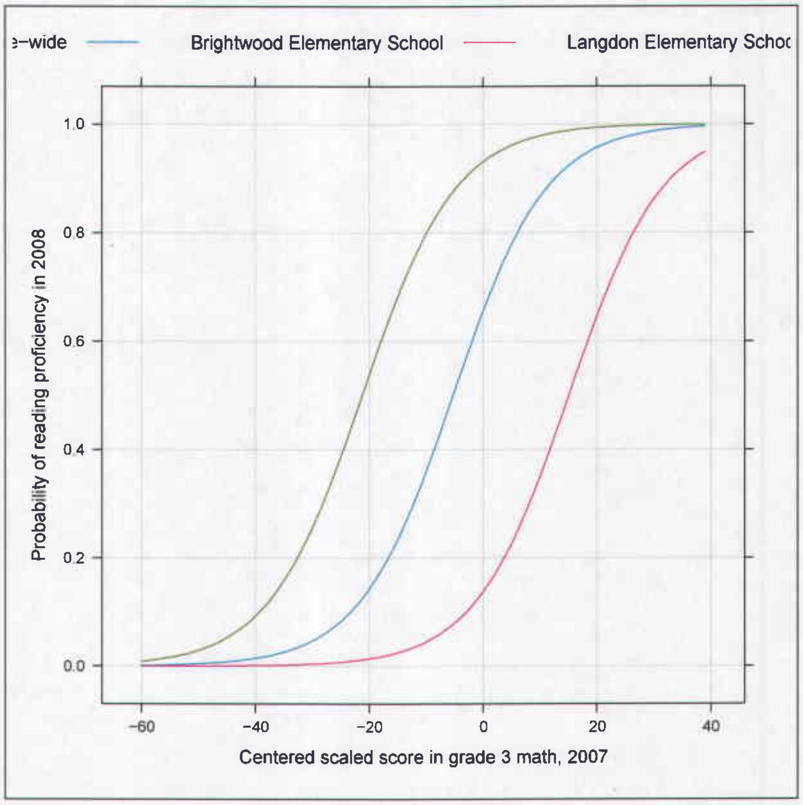


Figure 1: Logistic Curves

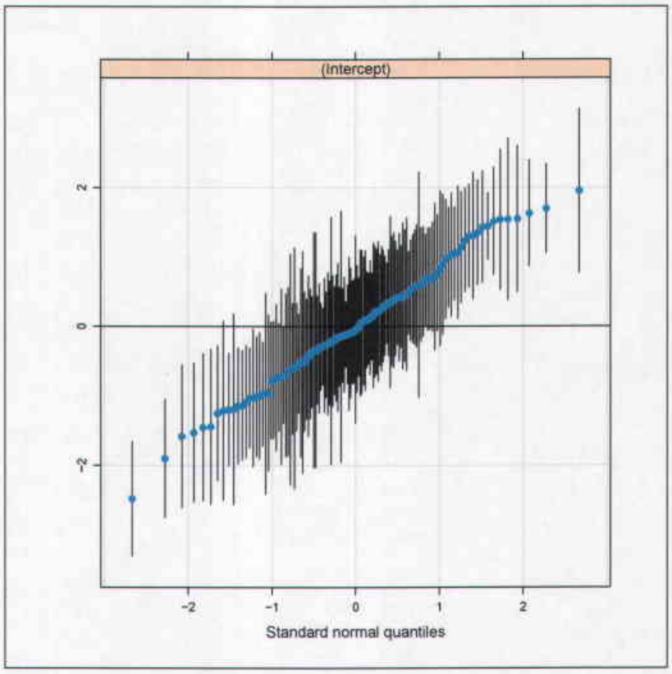


Figure 2: Caterpillar Plot: Grade 3 to 4 Math

> "Any student in grade 3 at Fake DC Elementary School with a scaled score of 350 in reading has a 63% chance of being proficient in grade 4"

At the student level, statements such as this are instructionally useful because teachers and/or parents can see if their student has a low probability of success. If so, then that might be a call to action for some form of instructional remediation in an attempt to improve the student's likelihood of success in the following school year. In this section, we describe how the probabilities are used to form AYP decisions. Before describing the application of the probabilities for NCLB-AYP decisions, some clarification regarding the reliability of the prediction is needed.

4.1 Reliability of the Prediction

It is well known that regression models that condition on variables measured with error yield some bias in the estimates of the fixed effects (Greene, 2000). In this section, we show how the influence of the variable measured with error is minimized given the way test is designed and how the probabilities are generated.

With fixed form assessments, there is typically more measurement error at the extremes of the ability scale than in the middle. This is because the test information function (TIF) is designed to peak nearest the proficiency cutscore; hence the test provides more information near this point than at any other point along the scale. Taking the inverse of the TIF gives the "lack of information", or the measurement error at each point along the ability scale (Lord, 1980).

Figure 3 illustrates this using data from an operational testing program in another state where this model has been implemented. The leftmost plot shows the conditional standard errors for all points along the ability scale for grade 3 math. The rightmost plot is the statewide logistic curve showing the probability for grade 4 proficiency conditional on grade 3 performance.

Because the slope of the logistic curve is not very steep at the high and low ends of the scale, the probability of proficiency in grade 4 does not change much with large changes in the x-axis. In other words, the probability of grade 4 proficiency is approximately the same for students with grade 3 scaled scores of -200 or -50 (again, these scaled scores are centered). This is also true at the high end of the curve as students with scores of 50 or 150 have virtually equivalent probabilities of proficiency in grade 4. Yet, these are the points on the scale where the asymptotic standard errors of $\hat{\theta}$ are the largest. In other words, in places along the scale where we would expect large swings in grade 3 performance as a function of measurement error, there is very little impact on the probability of the prediction because the low slope of the logistic curve dampens the effect of this measurement error.

Now, the slope of the logistic curve is very steep in the middle. As a result, large changes in the variable on the x-axis would result in substantial changes in the probabilities of proficiency in grade 4. However, this is the point on the test scale where the asymptotic standard errors of $\hat{\theta}$ are the smallest; thus it is improbable that we would observe large fluctuations in the observed scores in this region. Consequently, at the point on the logistic curve where large changes would have substantial impact, this effect is minimized given the way the test is designed to be most precise in this region.

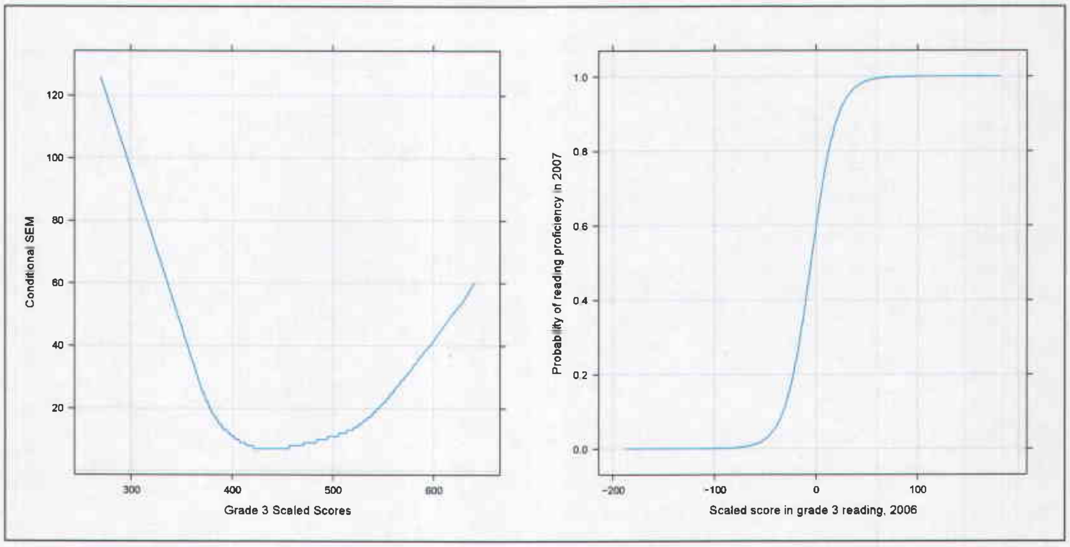


Figure 3: Conditional Standard Errors of Measurement and Logistic Curve

A second way to examine the reliability of the model is to examine the plot of the model fit as provided in Figure 4. The red line in this plot is the model-based prediction of the probability of proficiency in grade 4 conditional on grade 3 performance for the average school in the state. The blue circle is the observed proportion of students that actually obtained proficiency in grade 4 conditional on their grade 3 performance. For example, the model predicts that a student with a scaled score of 340 in grade 3 has about a 20% chance of being proficient in grade 4. The blue dot at this point shows that of those students with grade 3 math scaled scores of 340 in grade 3, about 20% of them actually became proficient the next school year. In sum, this fit plot shows a high degree of internal consistency between the model-based predictions and the observed proportions. Fit plots for all grades in reading and math are provided in Sections 7 and 8.

4.2 Accountability Implementation

For accountability purposes (the school level NCLB question), the quantity of interest is in $\mathcal{E}(\gamma_j)$– the number of students in school j expected to score at or above proficient in year $y+1$. Given the second property of probability stated in Section 2.1 we can aggregate these individual probabilities (i.e., $p(\hat{\theta}_{(j)i})_{g\rightarrow g+1}$) to the school level and for any subgroups of interest to form a statistic to determine the percentage of students likely to be proficient in year $y+1$. This is sometimes phrased as the percentage of kids "on track" to proficiency. This percentage can be compared to the state AMO for making AYP decisions.

This would be done as follows:

$$\mathcal{E}(\gamma_j) = \sum_{g=1}^{K}\sum_{i=1}^{n} p(\hat{\theta}_{(j)i})_{g\rightarrow g+1} \tag{5}$$

where n is the total number of tested students in school j in grade g and K is the number of grades included. The double summation is here because most states have school-based

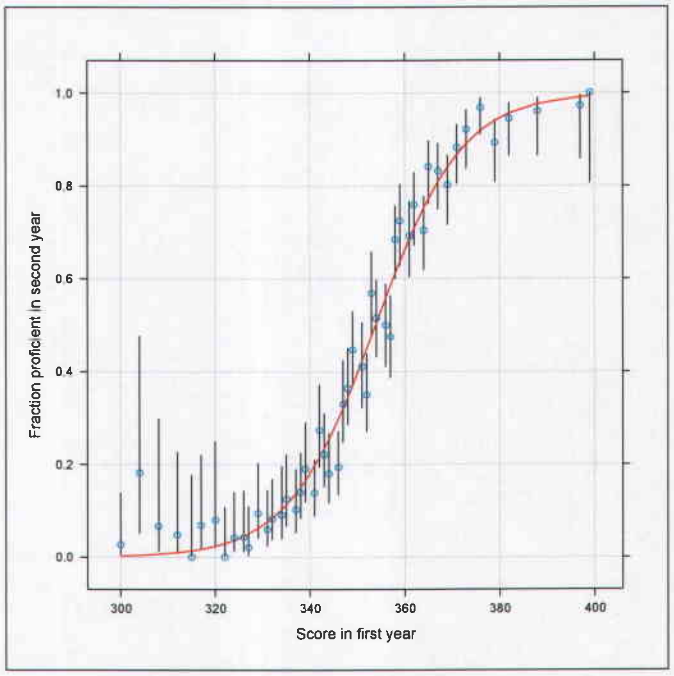


Figure 4: Fit Plot: Grade 3 to 4 Math

AMOs and not grade-specific AMOs. Hence, we first take the sum over all students within a grade and then take the summation over all grades within a school. If the state has grade-specific AMOs, then only the inner summation would be used.

We can also do this for specific NCLB subgroups such as:

$$\mathcal{E}(\gamma_H) = \sum_{g=1}^{K} \sum_{i \in H} p(\hat{\theta}_{(j)i})_{g \rightarrow g+1} \tag{6}$$

where H represents the various NCLB subgroups. The summation over all individuals within a school or within a subgroup yields the number of students projected to score at or above proficient in the following school year. The percentage is derived by dividing this expected number by the total number of students included in the analysis. Because these probabilities are student-specific, we can aggregate to any level of interest.

5 Computational Example

Continuing with the Grade 3 to 4 math example, the model in Equation 1 was implemented using the `lmer` function in the **lme4** package (Bates, Maechler, & Dai, 2008) in the `R` software program (R Development Core Team, 2008), yielding the results in Table 2.

Given the results in Table 2, the probability that a grade 3 student with a scaled score at the proficienct cutpoint (i.e., 360) will be proficient at a typical school in the state can be estimated using Equation 2:

$$[1 + \exp(-(.6455 + .122 \times 0))]^{-1} = .66 \tag{7}$$

Table 2: Grade 3 to 4 Math Results

Fixed Effects	Estimate	SE
μ	0.6455	0.1036
β	0.122	0.0044
$var(\nu)$	0.95835	

Similarly, the probability for a student with a scaled score 10 points above the proficient cutscore (i.e., 370) is:

$$[1 + \exp(-(.6455 + .122 \times 10))]^{-1} = .87 \tag{8}$$

As expected by Assumption 2, a student's likelihood of proficiency in grade $g+1$ depends on their grade g score such that students with higher scores are likely to have larger probabilities of future success. To account for Assumption 3, the calculation of η would include the conditional mode of the random effects for each school, ν_j.

For sake of illustration, assume a school had the following frequency of observed scores in 2008 as displayed in Table 3. Assume the goal is to determine the number of students on track to proficiency for the 2009 school year.

Table 3: Sample Grade 3 to 4 Projections

N	$\hat{\theta}$	$p_{3\to4}$	Projection
10	300	.0013	.013
20	320	.014	.28
15	350	.36	5.4
6	380	.95	5.7
8	390	.98	7.84

The first column `N` is the total number of students in the school at score point $\hat{\theta}$. For example, we can see that this school has 10 students with a grade 3 scaled score of 300. The column $p_{3\to4}$ is the probability of becoming proficient in the following school year for a given scaled score (we assume a typical school for the current example using the log-odds as provided in Table 2). For example, students with a grade 3 scaled score of 350 have about a 36% chance of being proficient in grade 4.

The column `Projection` shows the number of students at each score point who are likely to be proficient in grade 4. This is obtained by simply multiplying the value in the cell in column `N` by the value in the corresponding cell for $p_{3\to4}$.

Using Equation (5) we can see that 19 (actually 19.23) students are expected to be proficient next year. This is the marginal sum of the `Projection` column. So, we can compute the expected percentage of students likely to be proficient next year as $26.1/59 = 33\%$. This expected percentage is compared to the AMO to determine if the school made AYP or not.

One particular benefit of this model is that all students with observed test scores in 2008 would be included in the 2009 projections no exceptions. If a student is included in the status calculations, then they are also included in this analysis as well. It is also important to note that even students scoring above the grade 3 cutscore for proficiency are included in the grade 4 projection. Many of the current ED growth models only implement growth for those students scoring below proficiency. This model accounts for the expected growth of all students.

6 Summary

The description of the model was illustrated using data from grades 3 and 4 for math. However, full implementation of this model occurs for all grades and subjects in exactly the same way as described. For example, a separate regression is performed for grade 4 to 5 math to generate the conditional probabilities and so forth for each grade.

The benefits of this model are many. First, every student with an observed test score can be included when forming the projections. Second, the data only offer the chances that a student will be proficient in the subsequent school year rather than assuming the projected score is known with certainty. As such, this model is more conservative since it does not make claims beyond what the data can support. Last, the model can be used to determine which schools are producing gains in student achievement larger or smaller than other schools in the District.

References

Bates, D., Maechler, M., & Dai, B. (2008). *lme4: Linear mixed-effects models using s4 classes.* (R package version 0.999375-26)

Greene, W. H. (2000). *Econometric analysis* (Fourth ed.). Saddle River, New Jersey.

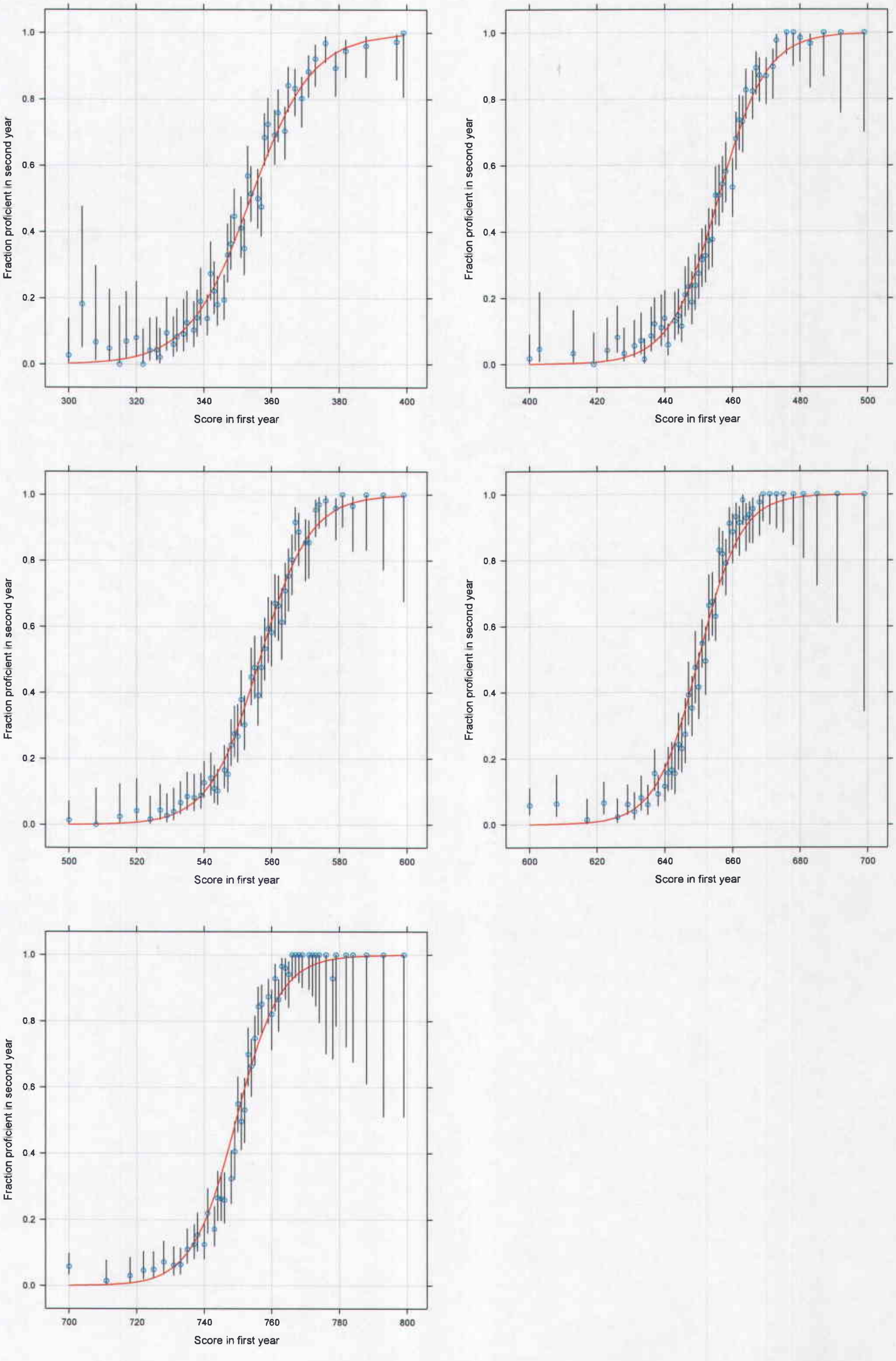
Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Erlbaum.

McCulloch, C. E., & Searle, S. (2002). *Generalized, linear, and mixed models.* New York: Wiley Interscience.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-Plus.* New York, NY: Springer.

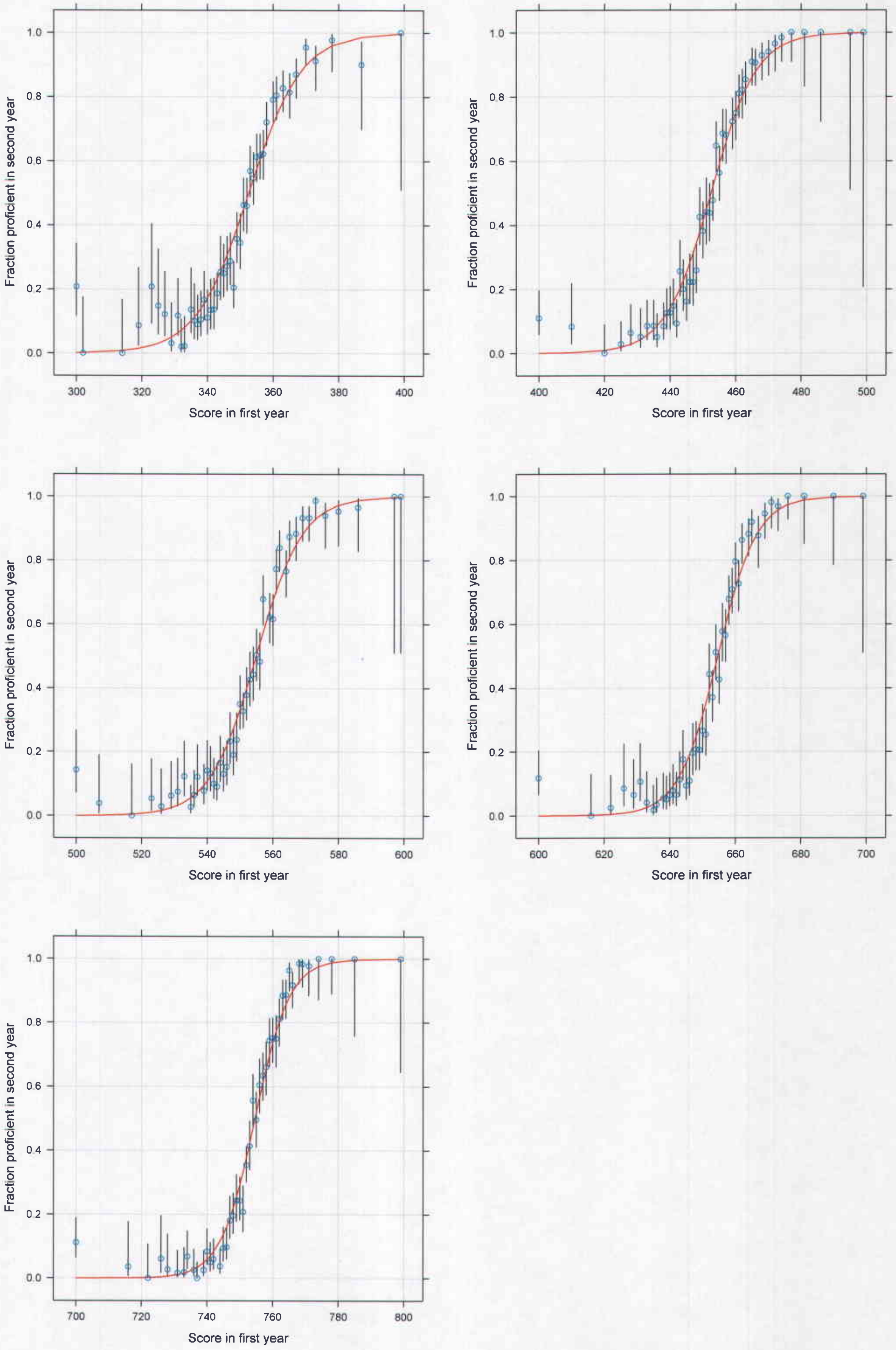
R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria. (ISBN 3-900051-07-0)

7 Math Fit Plots for All Grades



Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
300
320
340
360
380
400
Score in first year
Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
400
420
440
460
480
500
Score in first year
Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
500
520
540
560
580
600
Score in first year
Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
600
620
640
660
680
700
Score in first year
Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
700
720
740
760
780
800
Score in first year

8 Reading Fit Plots for All Grades


Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
300
320
340
360
380
400
Score in first year
Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
400
420
440
460
480
500
Score in first year
Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
500
520
540
560
580
600
Score in first year
Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
600
620
640
660
680
700
Score in first year
Fraction proficient in second year
0.0
0.2
0.4
0.6
0.8
1.0
700
720
740
760
780
800
Score in first year

Appendix B

Application of the Levenshtein Distance Metric For the Validation of Merged Data Sets

Harold C. Doran
hdoran@air.org

Paul Van Wamelen
pvanwamelen@air.org

American Institutes for Research
Washington, DC

Working Draft

October 6, 2008

Abstract

The analysis of longitudinal data in education is becoming more prevalent given the nature of testing systems constructed for No Child Left Behind. While these longitudinal analyses are more common, constructing the longitudinal data files remains a significant challenge. Students tend to move into new schools and districts, and in many cases, the unique identifiers (ID) that should remain constant for students change. It often occurs that different students share the same ID, in which case merging records for different students with the same ID is clearly problematic. In the absence of methods for confirmation, it is not possible to determine the integrity of a merge. In very small data sets, quality control can proceed through human reviews of the data to ensure all merges were properly performed. However, in data sets with hundreds of thousands of cases or more, quality control via human review is impossible. While some informal protocols may be in place for quality control, the educational measurement literature lacks formal protocols to monitor the integrity of merged databases. This paper presents an empirical quality control procedure that may be used to verify the integrity of the merges performed for longitudinal analysis. We also discuss possible extensions that would permit for merges to occur even when unique identifiers are not available.

Keywords: longitudinal analysis; Levenshtein algorithm; quality control; R program

1 Introduction

Under the No Child Left Behind Act (NCLB) students in grades 3 to 8 and high school are assessed at least once per year in reading and math. The scores from these yearly assessments are now becoming valuable pieces of information as they can be used in various longitudinal statistical analyses, such as value-added models and other growth models.

A prerequisite for these longitudinal analyses based on student level data is a data file that links scores for the same student over multiple time periods. However, the construction of a longitudinal database remains a significant challenge given the instability or availability of unique student identifiers (ID) – the variable commonly used to merge the files. If student IDs were perfectly reliable in the sense that they never changed or are never missing, the construction of the longitudinal database could be easily developed and there would be no need to validate whether the merge using ID actually joined the records for the same students.

However, this is rarely the case. Student IDs change, different students share the same ID, or the ID is often missing. The effects of these ID issues will reduce the validity of the merged records by either merging incorrect records together or dropping students from the data files.

While the literature regarding different forms of longitudinal analysis has substantially increased, there is no literature on the topic of protocols for the quality control of merged student data sets in education. This is especially surprising given that high-stakes decisions may be attached to the results of longitudinal analyses, such as adequate yearly progress decisions from the U.S. Department of Education (ED) growth model pilot program or value-added models with an explicit purpose on teacher and school evaluation (Harris, 2008).

In other disciplines, methods for ensuring the accuracy of merged records from different databases has a bit of an empirical history (Fellegi & Sunter, 1969). For instance, the medical field implements various probabilistic matching methods to ensure that the same records for the same patient are properly merged (Gill, Goldacre, Simmons, Bettley, & Griffith, 1993). Clearly, merging incorrect patient records could create false patient histories with far reaching consequences.

If longitudinal analyses are to be informative or to have any consequences, then methods for confirmation that student records are properly merged should be implemented. Consequently, the purpose of this paper is to present a general purpose algorithm that may be useful for guiding the quality control procedures needed to ensure that the correct student records are merged. Additionally, this algorithm can be used to improve merge rates when students are lacking unique IDs.

2 The Levenshtein Distance Metric

The Levenshtein Distance (LD) is a metric useful for determining the similarity of two character strings. The LD, or the edit distance, is defined as the minimum number of operations needed to transform `string 1` into `string 2` where an operation is either an insertion, deletion, or substitution of a character (Levenshtein, 1966). When the edit distance is 0, the two character strings are exactly the same. That is, no changes are needed to transform string 1 into string 2. When the edit distance is non-zero, this is an indication of differences between strings with larger edit distances indicative of larger differences.

Because the LD provides an empirical basis for comparing the similarity of character strings, it is extremely valuable as a confirmatory tool in judging whether the merges from data set 1 and data set 2 properly join the records for the same student. The following sections provide substantive details and examples on the edit distance. We also demonstrate how the LD, and its variants, can be used to verify the integrity of merged data sets.

2.1 Example of the Levenshtein Distance

Assume we have two character strings we wish to compare. The first string is Bill Clinton and the second string is William Clinton. The purpose of the LD procedure is to empirically determine how similar these two character strings are. In this example, the last name is exactly the same and no edits, insertions, or substitutions are necessary. The first name differs, however. Transforming Bill to William would require the following operations:

Operation 1: Substitute W for the B

Operation 2: Insert i after the Will

Operation 3: Insert a after the Willi

Operation 4: Insert m after the Willia

There are a total of four operations needed to transform Bill to William; hence the edit distance is 4. In order to facilitate the use of this procedure, the stringMatch function was developed in the R statistical computing environment (R Development Core Team, 2008). The stringMatch function is available in the **MiscPsycho** package (Doran, 2008).

The following are additional examples of the LD using this R function:

```
> stringMatch('William Clinton', 'Bill Clinton', normalize='no')
[1] 4
> stringMatch('Barack Obama', 'Barry Obama', normalize='no')
[1] 3
> stringMatch('John McCain', 'Jon McCain', normalize='no')
[1] 1
> stringMatch('John McCain', 'Barack Obama', normalize='no')
[1] 11
```

The comparison with the smallest edit distance is the John McCain to Jon McCain comparison, an indication that there is a high degree of similarity between these strings. The largest edit distance is the John McCain to Barack Obama comparison, an indication that these two strings not similar.

2.2 The Normalized Edit Distance

The edit distance is useful, but normalizing the distance to fall within the interval [0,1] is preferred. This normalization is preferred as it is somewhat difficult to judge whether an LD of 4 suggests a high or low degree of similarity. In our implementation, the Levenshtein distance is transformed to fall in this interval as follows:

$$LND = 1 - \frac{LD}{max(s1, s2)} \tag{1}$$

where LD is the edit distance and $max(s1, s2)$ denotes that we divide by the length of the larger of the two character strings. This normalization, referred to as the Levenshtein normalized distance (LND), yields a statistic where 1 indicates perfect agreement between the two strings and a 0 denotes imperfect agreement. The closer a value is to 1, the more certain we can be that the character strings are the same. The closer to 0, the less certain. For example:

```
> stringMatch('Bill Clinton', 'William Clinton', normalize='yes')
[1] 0.7333333
> stringMatch('Barack Obama', 'Barry Obama', normalize='yes')
[1] 0.75
> stringMatch('John McCain', 'Jon McCain', normalize='yes')
[1] 0.9090909
> stringMatch('John McCain', 'Barack Obama', normalize='yes')
[1] 0.08333333
```

Recall that the edit distance for transforming `Bill Clinton` into `William Clinton` is 4. The normalization in this case occurs as:

$$1 - \frac{4}{15} = .73 \tag{2}$$

We divide by 15 because the length of the character string `William Clinton` is 15, which is the larger of the two strings (the space between the first and last name is included in this example). As in Section 2.2, the `John McCain` to `Jon McCain` yields the best comparison with an LND value close to 1, a strong indication of similarity. Additionally, the `John McCain` to `Barack Obama` comparison yields the worst comparison with a value of .08, a strong indication of dissimilarity.

2.3 Probability of the Normalized Edit Distance

In addition to the LND, it is useful to determine the chances of observing an LND value of x or larger in a population of names. In other words, the desired inference is the probability that a comparison of two random character strings would yield an LND statistic of x. This probability can be used to help determine if the LND statistic obtained between the two character strings is indicative of a high match or a low match.

For example, if a comparison of two strings returned an LND of .7 and the $\mathbb{P}(\text{LND} \geq .7) = .001$, we could be relatively confident that any two comparisons yielding an LND of

.7 are similar names given that it is so unlikely that an LND = .7 would occur from a chance comparison of two random strings. On the other hand, if a comparison of two strings returned an LND of .3 and the $\mathbb{P}(\text{LND} \geq .3) = .6$, then we have some assurance that an LND of .3 yields an incorrect comparison given that an LND of .3 would occur about 60% of the time when two random strings are compared.

Given a large data set with many names (such as a student test score database), the following procedure can be used to empirically obtain these probabilities:

1. Take a random sample without replacement of n names from the data.
2. Compare each of these n names to all N student names in the data to obtain the LND.
3. Count the number of times the LND of x_i is observed.
4. Divide x_i by the total number of comparisons made to obtain $p(x_i)$

Because the intended inference is $\mathbb{P}(x \geq x_i)$, we compute the cumulative probabilities as:

$$\mathbb{P}(x \geq x_i) = 1 - \sum_{x_i \leq x} p(x_i) \tag{3}$$

To illustrate this process, assume we have the following data:

```
          fname1          fname2
1  Joseph McCall      Joe McCall
2     Paul Jones      Paul Jones
3  Larry Everett   Barry Everett
4   Sam Thompson Samuel Thompson
5   Sally Fields    Sally Fields
6    Doug Carter  Douglas Carter
7  Bill Friendly  William Friend
8    Tom Davison   Tommy Davison
9     Frank Mann   Franklin Mann
10    Mary Jones      Cary Jones
```

where `fname1` is the name in year 1 and `fname2` is the name in year 2. Assume we take a random sample without replacement of five names from `fname1`. This might result in `Joseph McCall, Sally Fields, Frank Mann, Paul Jones, Larry Everett`. Now, we compute the normalized edit distance for these sampled names against every name in `fname2`.

Table 1 shows how the names from the random sample are compared to every other name in the data yielding a normalized edit distance for each comparison. Retrieving the probability now only requires that we count the number of times the same normalized distance is observed divided by the total number of cells in the table. For instance, the value of 1 occurs twice, once in the `Sally Fields` comparison and again for the `Paul Jones` comparison. There are 50 total cells in the table. Hence, the probability of observing a 1 in these data is 2/50 = .04.

To facilitate use of this procedure, a second `R` program is also available in the **MiscPsycho** package called `stringProbs` to automate this process. Use of the function would proceed as:

	Joseph McCall	Sally Fields	Frank Mann	Paul Jones	Larry Everett
Joe McCall	0.77	0.08	0.10	0.00	0.08
Paul Jones	0.08	0.33	0.20	1.00	0.23
Barry Everett	0.00	0.31	0.08	0.23	0.92
Samuel Thompson	0.13	0.27	0.20	0.40	0.13
Sally Fields	0.08	1.00	0.08	0.33	0.31
Douglas Carter	0.14	0.14	0.14	0.29	0.07
William Friend	0.07	0.43	0.14	0.21	0.14
Tommy Davison	0.08	0.15	0.23	0.15	0.15
Franklin Mann	0.08	0.08	0.77	0.23	0.00
Cary Jones	0.08	0.33	0.20	0.70	0.38

Table 1: LND Statistics Under Multiple Comparisons

```
> stringProbs(dat, N=5)
```

In this function, the user simply provides a dataframe (`dat`) with the names to be compared. The argument `N=5` is used to determine how many names are randomly sampled for the comparison. The value of N can be equal to the total number of rows in the dataframe `dat`, but it cannot be larger.

Figure 1 plots the cumulative probability of occurrence against the LND statistic for these data. This figure shows that the $\mathbb{P}(\text{LND} \geq 1) \approx 0$, $\mathbb{P}(\text{LND} \geq 0) \approx 1$, and $\mathbb{P}(\text{LND} \geq .2)$ or larger is approximately .4.

3 Application of the Levenshtein Normalized Distance for Merging Student Records

Many states maintain unique student identifier codes that should remain constant over time. If these IDs were perfectly reliable, then it would be possible to merge using these IDs alone. However, these student IDs can and do change, thus making it feasible that two different students can share the same ID over time. It would of course be improper (and would yield unreliable statistical estimates) if the records for these different students were merged as it would create a false student history.

In small data sets with a few hundred students, it would be possible to manually review the merge to assess whether the correct records were joined between the different data files. However, in data sets with hundreds of thousands of cases or more, an automated and empirical procedure is needed to verify if merged data sets properly joined records for the same student.

Therefore, the following strategy can be used to merge and validate that the correct student records were merged. Doing so would be performed in three steps:

1. Merge the year 1 data file with the year 2 data file using the available unique student identifiers.

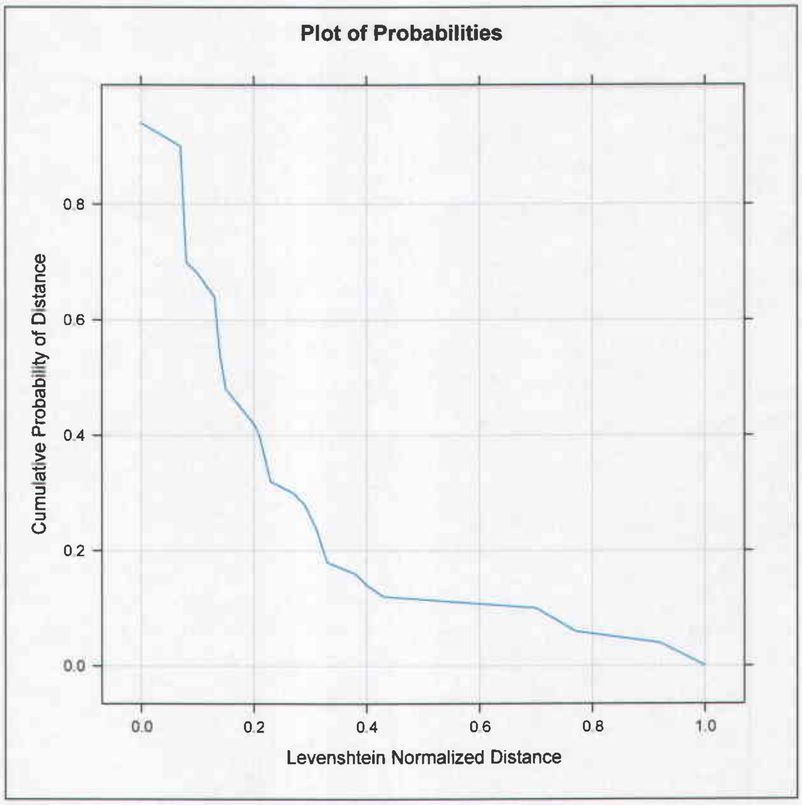


Figure 1: Sample Plot of Cumulative Probabilities

2. Compute the LND for each record in the data using the student first and last names in both years.

3. Subset the data and keep only those records where the normalized edit distance obtained by comparing the first and last names in year 1 against the first and last names in year 2 is $\geq x$ where x is a value determined through examination of the data and the probability of occurrence.

When this process is applied to student test score data, the LND is used to verify that the correct student records are merged. One particular challenge, discussed in the next section, is what value of x should be used as the cutpoint.

It sometimes occurs that in student achievement data sets that the names are transposed from one year to the next. For example, a student's first name may be properly recorded in year 1 as `William Clinton` but in year two, the first and last names are switched as `Clinton Bill`. If the LND were obtained under this circumstance it would yield:

```
> stringMatch('William Clinton', 'Clinton Bill')
[1] 0.1333333
```

which is a very low value and may suggest this student should be dropped from the database even if the IDs were the same. One possible enhancement is to derive the LND under multiple permutations of the first and last name and use the maximum value of the LND under each permutation. The following three permutations may be useful:

Permutation 1: Compare year 1 (first, last) to year 2 (first, last)

Permutation 2: Compare year 1 (first, last) to year 2 (last, first)

Permutation 3: Compare year 1 (last, first) to year 2 (first, last)

In this example, each of the three permutations would result in the following:

```
> stringMatch('William Clinton', 'Clinton Bill') # Permutation 1
[1] 0.1333333
> stringMatch('William Clinton', 'Bill Clinton') # Permutation 2
[1] 0.7333333
> stringMatch('Clinton William', 'Bill Clinton') # Permutation 3
[1] 0.1333333
```

The maximum LND of .73 would be used as the value for this student in the verification process.

4 Demonstration Using Data from Washington, DC

Two years of test score data were merged using the unique student identifier codes in year 1 and in year 2. Only students with first and last names in years 1 and 2 are retained in the data, resulting in a total of 22,890 students.

As a first step, the first and last names in year 1 are concatenated to form `string 1` and the first and last names in year 2 are concatenated to form `string 2`. The `stringMatch` function is applied to all students comparing `string 1` and `string 2` to compute the LND. For the current example, only **Permutation 1** is provided. Greater reliability would likely result if all three permutations were applied and the maximum LND from each were used.

Figure 2 is a set of conditional density plots showing the distribution of the LND statistic for all students in the data by grade level. The very large spike near 1 for each grade shows that the LND is very high for almost all cases in the data, suggesting that these data are very clean. That is, the very large LND for almost all students is indicative of the fact the the merge using student ID most likely merged the correct students together as verified using the `string 1` to `string 2` LND comparison.

As a second step, the `stringProbs` function is used to get an empirical estimate of the probability of observing an LND of x. In this application, a random sample of 10 names are drawn from `string 1`. These 10 names are compared to every `string 2` name in the data. The purpose of this step is to help determine what cutpoint might be used for subsetting the data. That is, we may want to keep only those students whose IDs are the same and the $LND \geq x$.

Figure 3 plots the cumulative probability of occurrence against the LND statistic. The results in this figure suggest that the probability of occurrence begins to diverge from 0

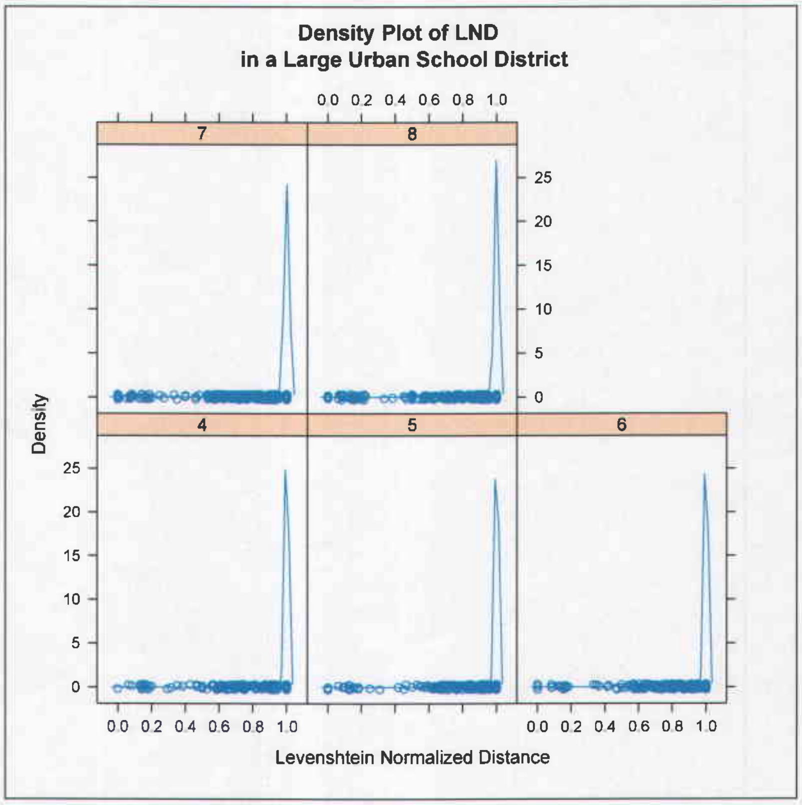


Figure 2: LND Density Plot

around an LND of .4. This suggests that an observed LND of about .4 is the point at where we can become suspicious that the merges may be incorrect because the chance occurrence begins an upwards trend towards 1 near this point.

The purpose in obtaining these probabilities is to offer an empirical method such that the process can be automated. For example, we may first merge on ID and then, based on Figure 3, retain only those records with an LND $\geq .4$. However, for purposes of further exploration, we manually review the data to assess the degree to which the merge was properly performed.

While the data used in this example cannot be publicly shared because of FERPA regulations [1], the matching of correct students is remarkably correct. In our review, students with LND statistics lower than .4 are always different students even though they shared the same student ID. Had these students been retained in the data without the LND verification, the wrong records would have been merged and subsequent statistical analyses would reflect this inaccuracy. This nicely conforms to the probabilities as displayed in Figure 3 because the plot suggests that $LND \leq .4$ is the point where incorrect merges tend to become more likely. Within the range of .4 to .7, there is quite a bit of ambiguity in the student names compared, therefore it is not always possible to discern if the names are the same. Within this range are many misspellings, reversals (i.e., first name and last name are switched), and

[1] FERPA is the Family Education Rights and Privacy Act and prevents any identifiable information from being released

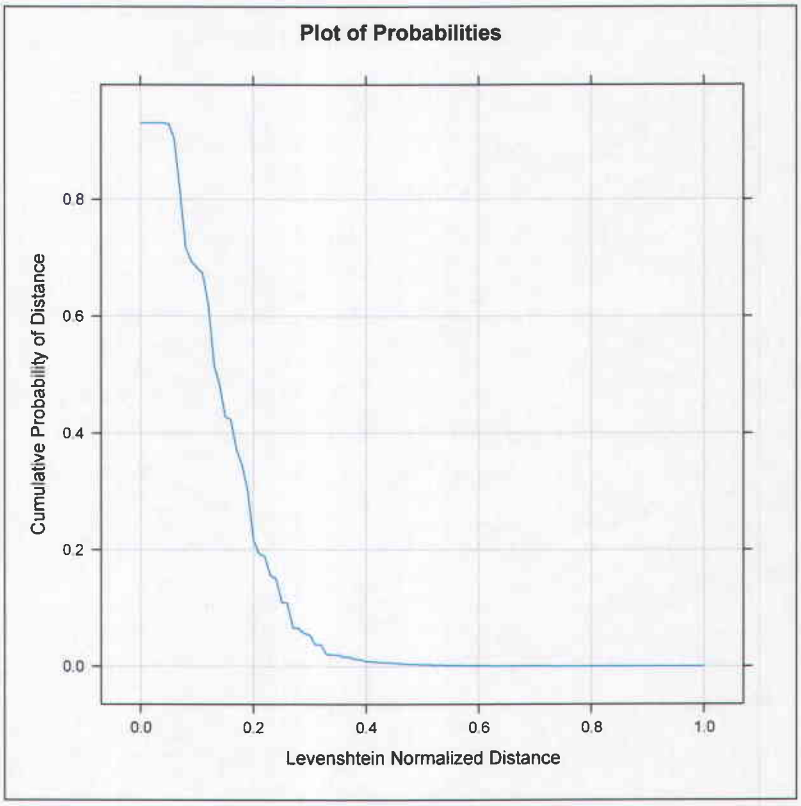


Figure 3: Plot of Cumulative Probabilities

obvious non-matches. For all students with an LND above .7, our manual review indicates that these students are always correctly merged. Hence, LND = .7 would be a good cutpoint chosen for retaining students.

5 Generalizations of the LND

The LND is a general purpose statistic that can be used quite variably and reliably depending on the variables in the data. In this section, we propose two examples of how it may be extended to further verify if records were properly merged or how records might be merged in situations where unique identifiers are completely lacking.

5.1 Use of LND and Other Demographics to Seek Further Validation

In the examples used in this paper, only first and last names were combined to compute the LND and the probability of its occurrence. However, the method is clearly general and other demographic characteristics of students could be brought in for purposes of additional assurance.

For instance, assume that gender and date of birth are available in the year 1 and year 2 data files that are to be merged in addition to the first and last names and the unique student ID. These indicators can be used to further determine the chances of a correct merge. To illustrate, assume there are three students in the data as shown in Table 2.

Student	DOB1	DOB2	Gender1	Gender2	LND	$\mathbb{P}$(LND)
1	10.20.2001	10.20.2001	F	F	.7	.001
2	12.15.2001	1.15.2001	F	M	.2	.2
3	4.17.2001	3.18.2002	M	F	.1	.4

Table 2: Sample Table of Demographic Characteristics

The issue at hand is answering the question, "what is the probability of two random students having an exact match on all of these characteristics?". This probability can be used to determine a point of cutoff. For instance, the chances of an exact match on month in the date of birth is 1/12, the chances of an exact match on day in the date of birth is roughly 1/31, the chances of an exact match on year in the date of birth is $1/Q$ where Q is the number of years listed in the data file, and the chances of an exact match on gender is 1/2.

For the current illustration, assume a frequency count on year in the data shows there are only four years (2000, 2001, 2002, 2003), thus $Q = 4$. In the first case, Student 1 matches exactly on all of these demographic characteristics. So, we compute the probability of two random students having an exact match on all of these characteristics as well as the LND as $1/31 \times 1/12 \times 1/4 \times 1/2 \times .001 = 3.360e - 07$. In the case of the Student 2, there is an exact match only on day and year in the date of birth. So, the probability of two random students matching on these two characteristics and the LND is $1/31 \times 1/4 \times .2 = 0.0016$. The last student matches on none of the demographic characteristics and the probability of the LND is rather high. In this case, the chance of two random students matching on this single characteristic is .4.

In the case of Student 1, the probability that two random students would match on each of these characteristics is very low and the LND is relatively high. Therefore there is a good degree of assurance that the records for this student are properly merged. This is less true for Students 2 and 3. The match probability for student 2 is very low, but the LND is also very low, suggesting that an incorrect merge occurred. With large-scale data sets, this probability would be computed for all students in the data using any available demographic characteristics. Subsequently, the retention of students in the merged data might include only those with an LND $\geq .7$ and with a match probability less than p.

5.2 Merges When Unique ID is Unavailable

Implementation of this algorithm could conceivably make it possible to merge student records even if a unique student ID were not available at all. In this case, the LND could be computed by comparing all N_1 names in the year 1 data file to all N_2 names in the year 2 data file. One could then subset the data and retain any record for which the LND is greater than a predetermined cutoff point. While this idea is possible, it is less than ideal. When the number of names is large, the computational burden is rather tremendous as the number of comparisons made would be $N_1 \times N_2$. The `stringProbs` function implements exactly this comparison to compute probabilities. The example given in Section 4 compares 10 random names to all other 22,890 names in the data. Even this small comparison of 22,890 × 10 is rather time consuming. A comparison of 22,890 × 22,890 would yield more than 5 million comparisons, a less than desirable computational scenario.

On the other hand, matching on names is conceivable for a small number of individuals. Hence, if the $N_1 \times N_2$ array is small, then such a merge is feasible. This kind of merge validation may be valuable for joining teacher records over time. It is often the case that unique teacher IDs do not exist, but there may be a need to join records for teachers together from different data files.

Last, the LND can be used in a "salvage" effort used to recover or improve merge rates for those individuals that may have been lost in the merge using student ID. For instance, say 500 individuals could not be merged because they lacked a unique ID in one of the years or the ID from year 1 to year 2 changed. It is undesirable to lose these students completely. Consequently, it is possible that some percentage of these 500 students could be recovered by joining the data files using some creativity and a potentially reliable variable. For example, a merge could occur using the first n letters of the last name (or do a merge on birthdate) in year 1 and year 2 and then validate that merge by computing the LND on the first and last names.

For example, consider the sample data in Table 3. The column `Name 1` is the student's name in the year 1 file, `Name 2` is the student's name in the year 2 file, `Lname1` is the first four letters of the students year 1 last name, `Lname2` is the first four letters of the students year 2 last name, `LND` is the LND statistic computed by comparing the year 1 and year 2 names, and `ID1` and `ID2` are the year 1 and year 2 student IDs, respectively.

In this toy example, the data files are joined by merging on the first four characters of the last name. Now, because of similarities in the last names across different students, this is likely to yield some incorrect merges. For instance, the data in row 1 are merged based on `jone`, but the LND statistic as well as visual comparisons show this is clearly an incorrect merge. The student in row 2 has no year 2 ID. As such, this student would have been dropped if the merge occurred on ID. However, merging on the first four characters of the last name and computing the LND shows this is indeed the same student. The student in row 3 has no year 1 ID. However, the LND statistic computed for this individual suggests it is most likely the same student.

Name 1	Name 2	Lname1	Lname2	LND	ID1	ID2
Mary Jones	William Jones	jone	jone	.46	123	456
Doug Carter	Doug Carter	cart	cart	1	321	
William Thompson	Bill Thompson	thom	thom	.75		654

Table 3: Sample Fuzzy Merge

A fuzzy merge such as this has potential risks and so a very stringent value of the LND should be chosen to guarantee a proper merge. It would also be very useful in situations such as this to further compute match probabilities as demonstrated in Section 5.1.

6 Summary

Merging student records from different yearly data files is a difficult task. Student IDs change or they may be missing in one year even though the student has valid test score attempts in both years. Additionally, quality control procedures for validating the merged records are sorely lacking from the educational literature. As longitudinal analyses become more common, and as they are used for purposes such as adequate yearly progress decisions, such quality control protocols should be in place.

The methods presented in this paper can be used as a general framework for quality control or for merging individuals where unique student identifiers are not available. As states or school districts implement longitudinal analyses for school or teacher accountability, consideration of quality control procedures such as those presented in this paper should become a part of the data quality plan.

Future research in this area can expand the probabilistic matching methods described by drawing upon the implementation of pattern matching algorithms used in other fields. For instance, the extreme value distribution has been used in genetics to explore the degree to which to two different DNA structures share similar features (Arratia, Gordon, & Waterman, 1986) and the field of music has explored how audio files can be matched based on their similar features (Yang, n.d.).

References

Arratia, R., Gordon, L., & Waterman, M. (1986). An extreme value theory for sequence matching. *The Annals of Statistics*, *14*(3), 971-993.

Doran, H. C. (2008). *MiscPsycho: Miscellaneous psychometrics.* (R package version 1.2)

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Gill, L., Goldacre, M., Simmons, H., Bettley, G., & Griffith, M. (1993). Computerised linking of medical records: methodological guidelines. *Journal of Epidemiology and Community Health*, *47*, 316-319.

Harris, D. N. (2008, June). *Would accountability based on teacher value-added be smart policy? an examination of the statistical properties and policy alternatives* (Tech. Rep.). University of Wisconsin at Madison.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*, 707-710.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria. (ISBN 3-900051-07-0)

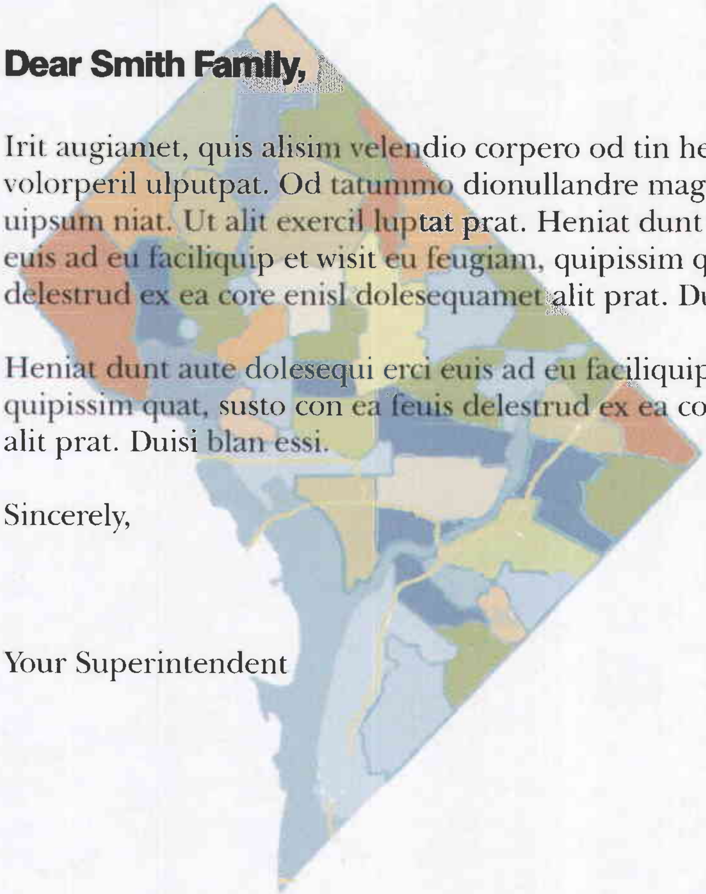
Yang, C. (n.d.). *Music database retrieval based on spectral similarity* (Tech. Rep.). Stanford University. http://kom.aau.dk/ oa/Teaching/IN6-03/ProjectProposals/Yang01.pdf.

Appendix C

Alexander Smith

Test Date: Spring 2009
School: DC School (157896)
Grade: 5
Subjects: Reading and Mathematics

Reading This Report Can Help Your Child Succeed!



Dear Smith Family,

Irit augiamet, quis alisim velendio corpero od tin henim erci te et vent volorperil ulputpat. Od tatummo dionullandre magnim volorer alisseq uipsum niat. Ut alit exercil luptat prat. Heniat dunt aute dolesequi erci euis ad eu faciliquip et wisit eu feugiam, quipissim quat, susto con ea feuis delestrud ex ea core enisl dolesequamet alit prat. Duisi blan essi.

Heniat dunt aute dolesequi erci euis ad eu faciliquip et wisit eu feugiam, quipissim quat, susto con ea feuis delestrud ex ea core enisl dolesequamet alit prat. Duisi blan essi.

Sincerely,

Your Superintendent

Online Resources:

www.feuguezcvzccveroetum.com

www.feuguevcvvvvvvvvveroetum.com

www.feuguxxeveroetum.com

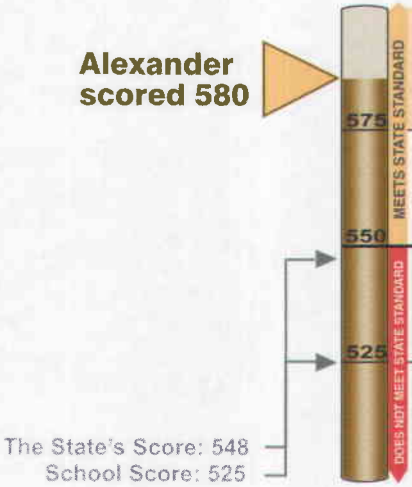
www.feuguxxxeveroetum.com

HOW DID ALEXANDER DO ON THE READING TEST?

580

(Advanced)

How does this compare?
Alexander's score is higher than the average score of fifth graders in his school, and higher than the average score of fifth graders in the state.



Advanced - Students use context clues and common Greek or Latin roots and affixes to determine the meaning of words. They understand figurative language. They paraphrase key points of persuasive texts, use details to describe narrative events, and analyze characters in grade 5 texts.

Proficient - Students use context clues to determine the meaning of words and phrases. They understand some figurative language. They restate key points in persuasive texts, identify components of narrative events, and analyze characters in grade 5 texts.

Basic - Students use context clues to determine the meaning of words and phrases. They understand simple figurative language. They determine the author's purpose in simple texts, identify main ideas and key events, and understand characters in some grade 5 texts.

Below Basic - Students determine the meaning of some words. They determine the author's purpose in simple texts, identify main ideas and some key events, and identify changes in characters in some grade 5 texts.

HOW DID ALEXANDER DO IN THE PAST?

	Grade 3	Grade 4	Grade 5
Advanced	○	○	✓
Proficient	○	○	○
Basic	○	✓	○
Below Basic	✓	○	○

IS ALEXANDER ON TRACK TO BE PROFICIENT NEXT YEAR?

The chances that your child will be proficient on the Reading test next year are **HIGH**.

HOW DID ALEXANDER DO ON THE READING CONTENT AREA STRANDS?

🔴 Below Proficient 🟡 Near Proficient 🟢 Above Proficient

🟡 Language Development

Your child is **Near Proficient** in this standard.

Play a "word of the day" game with your child. Have him/her find and define a different word every day from different types of literature.

🟢 Beginning Reading

Your child is **Above Proficient** in this standard.

Have your child read a favorite story or poem out loud. Then ask him/her to explain the story in his/her own words.

🟢 Informational Text

Your child is **Above Proficient** in this standard.

Read "Letters to the Editor with your child. Have him/her identify ways the author supports his/ her position.

🟢 Literary Text

Your child is **Above Proficient** in this standard.

Pick out several different types of literature to read with your child (e.g., story, article, poem). Have him/her list the differences between each form.

🟢 Research

Your child is **Above Proficient** in this standard.

Have your child use a several sources (e.g., dictionary, encyclopedia, internet) to look up facts on their favorite animal (e.g., diet, habitat).

🟢 Writing

Your child is **Above Proficient** in this standard.

Have your child write an explanation of a process (e.g., how to make cookies). Have him/her include a topic statement, supporting details, and a conclusion.

🟡 Media

Your child is **Near Proficient** in this standard.

Collect ads from newspapers or magazines. Have your child point out words and figures of speech that are meant to change your mind about the product (e.g., New!, Exclusive!).

🟡 English Language Conventions

Your child is **Near Proficient** in this standard.

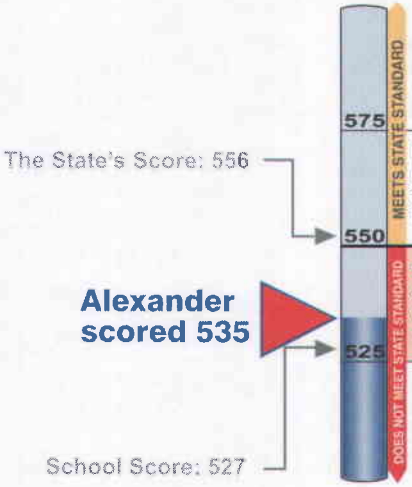
Have your child write a story or poem about a family member or friend. When he/she is finished go back over the story for spelling and grammar errors.

HOW DID ALEXANDER DO ON THE MATHEMATICS TEST?

535

(Basic)

How does this compare?
Alexander's score is higher than the average score of fifth graders in his school, and lower than the average score of fifth graders in the state.



Advanced - Students perform complicated computations. They explain real world problems using expressions and one-step equations. They compare and analyze features of shapes. They solve problems involving proportional relationships and measurement conversions.

Proficient - Students perform computations using decimals. They create and use simple expressions to solve real world problems. They solve simple one-step equations. They use geometric properties to identify and measure angles, and solve simple measurement problems.

Basic - Students perform computations in correct sequence. They extend patterns and simplify basic expressions. They identify and directly measure angles. They provide solutions for straightforward measurement problems. They use mathematical language clearly.

Below Basic - Students perform basic computations on whole numbers and fractions. They identify simple patterns and may be able to extend given patterns. They identify different types of angles. They use scale drawings to show data, and use tools to find measurements.

HOW DID ALEXANDER DO IN THE PAST?

	Grade 3	Grade 4	Grade 5
Advanced	○	○	○
Proficient	○	○	○
Basic	○	✓	✓
Below Basic	✓	○	○

IS ALEXANDER ON TRACK TO BE PROFICIENT NEXT YEAR?

The chances that your child will be proficient on the Mathematics test next year are **AVERAGE**.

HOW DID ALEXANDER DO ON THE MATHEMATICS CONTENT AREA STRANDS?

🔴 Below Proficient 🟡 Near Proficient 🟢 Above Proficient

🟡 Number Sense and Operations

Your child is **Near Proficient** in this standard.

Cook with your child using a recipe that asks for fractions (e.g., 1/3 cup). Ask him the amount of an ingredients needed if you were to double or triple the recipe.

🟡 Patterns, Relations, and Algebra

Your child is **Near Proficient** in this standard.

Challenge your child to create and solve mathematical equations using variables (e.g., what is 2x + 3 if x = 4).

🔴 Geometry

Your child is **Below Proficient** in this standard.

Have your child find examples of symmetry around the house (e.g., tiles on a floor). Ask you child to hang towels so that they are symmetrical.

🔴 Measurement

Your child is **Below Proficient** in this standard.

Have your child find examples of triangles, rectangles, and parallelograms in your home. Work with your child to find the area of these items using the appropriate formulas.

🔴 Data Analysis, Statistics, and Probability

Your child is **Below Proficient** in this standard.

Play a board game with your child that requires dice. Ask him how likely it would be for him/her to roll a six on one die. Then ask him/her how likely it would be to roll a four on one die and six on the other.

For more information about content area strands visit www.DCstrandinfo.com

How Did My School Contribute to My Child's Learning?

READING

Your school is a **LOW PERFORMING** school, and it has **HIGH OVERALL GROWTH**.

Many of the students in your child's school have performed below proficient in reading. However, the school is helping students improve.

Your school contributes a lot to your child's learning. This is more than other schools in your district.

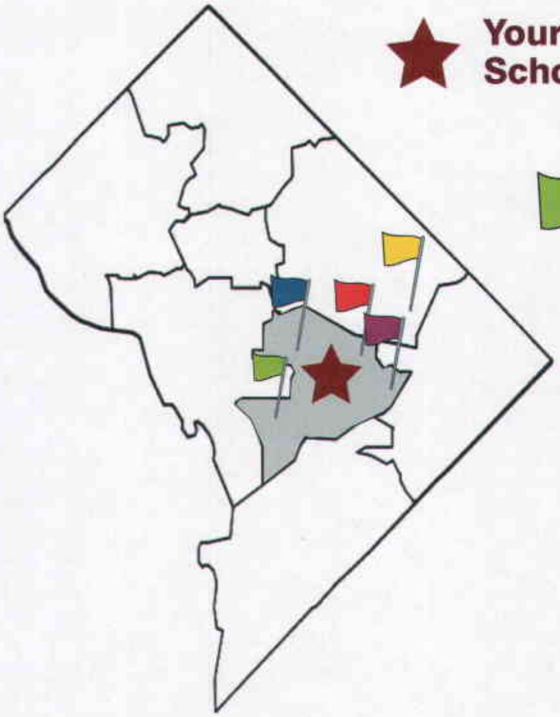
MATHEMATICS

Your school is a **LOW PERFORMING** school, and it has **LOW OVERALL GROWTH**.

Many of the students in your child's school have performed below proficient in mathematics. Your school is somewhat helping students improve.

Your school contributes an average amount to your child's learning. This is similar to other schools in your district.

At which schools will my child have a better chance of being proficient next year?



★ Your School

🟦 Edison Elementary

🟨 Franklin Middle School

🟩 Johnson Elementary

🟥 Eisenhower School

🟪 Lincoln Elementary

This is a list of schools in Washington D.C. at which your child may have a better chance of being proficient next year.

You may consider speaking with your child's school administrator to see what these schools are doing differently or you may consider enrolling your child in one of these schools.

More Information about how to help Alexander reach proficiency

To learn more about what your school is doing to help your child learn, please contact Lynda Caldwell by phone at 234-333-2928 or by email at lcaldwell@dcps.com.

The district is required to provide extra support such as ___, ___, and ___ for students who may not be proficient next year. Please contact Lynda Caldwell by phone at 234-333-2928 or by email at lcaldwell@dcps.com for more information.

If your child does not get help, your child may not reach proficiency next year.