Using Growth Models to Measure School Performance
Implications for Gifted Learners
Gail R. Ryser, PhD1 and Karen E. Rambo-Hernandez, PhD2

Abstract: Using assessment data to determine student growth has become an integral part of the accountability movement, and researchers and educators are currently examining how new rules impact the academic assessment of gifted learners. In 2008, the Association for the Gifted’s Annual Symposium at the Council for Exceptional Children Convention focused on policy and legislative issues and their effects on gifted and talented students. One presentation in the symposium centered on the No Child Left Behind Growth Model Pilot Program (GMPP). This article provides an historical overview of the GMPP, describes the current use of growth models used by states for accountability purposes, defines measuring growth from a statistical modeling perspective, discusses implications for gifted learners who are above proficiency, and makes recommendations for policy makers and administrators to begin measuring growth.

Keywords: growth models, accountability, gifted students

Using assessment data to determine student growth has become an integral part of the accountability movement (American Recovery and Reinvestment Act, 2009), and researchers and educators are currently examining how new rules impact the academic assessment of gifted learners. In 2008, the Association for the Gifted’s Annual Symposium at the Council for Exceptional Children Talented and Gifted (CEC-TAG) Convention focused on policy and legislative issues and their effects on gifted and talented students. One presentation in the symposium centered on the No Child Left Behind Growth Model Pilot Program (GMPP). At that time, using growth models to provide estimates of whether student achievement would meet Adequate Yearly Progress (AYP), part of the requirements of the NCLB Act, was in its infancy. Unfortunately, the GMPP did not hold promise for measuring the growth of gifted students because its sole purpose was to identify students scoring below proficient in reading or language arts and mathematics as being on-track to proficiency. Part of the symposium’s purpose was to produce position papers about the various topics presented. As a result, a subcommittee developed a position paper on measuring growth in achievement of gifted students (National Association of Gifted Children, 2008). The position paper made several recommendations including that growth models need to reflect growth that goes beyond proficiency. This article provides an historical overview of the GMPP, describes the current use of growth models used by states for accountability purposes, defines measuring growth from a statistical modeling perspective, discusses implications for gifted learners who score above proficiency, and makes recommendations for policy makers and administrators to begin measuring growth.

NCLB GMPP

Under the 2001 amended Elementary and Secondary Education Act (NCLB), states began conducting annual assessments of student achievement for AYP determination using status models (U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Office of Policy and Program Studies Service, 2011). Status models require schools to test 95% of their students in both reading or language arts and mathematics. In addition, the percentage of tested students scoring proficient or higher must meet or exceed the standard or the annual measurable objective (AMO) that schools and districts set. The main characteristic of status models is that they only focus on a school’s level of student proficiency at one point in time. Thus, status models do not recognize real improvements in individual student’s achievement scores. Furthermore, there is considerable agreement among educators that monitoring schools based on the percentage of their students meeting proficiency at one static point in time places school with diverse populations at a disadvantage (Novak & Fuller, 2003). Schools must assess the performance of several subgroups such as English Language Learners, children from economically disadvantaged families, and each ethnic group. In demographically diverse schools, students may be counted in two or three of these subgroups and if a student is not...
proficient, then they are counted as not proficient multiple times. In 2005, Margaret Spellings, the Secretary of Education, initiated the GMPP (written into regulation in 2008) as another method to determine compliance with NCLB achievement mandates. The pilot program had two important goals: to predict whether non-proficient students would attain proficiency within a given time period and to identify the performance levels students must achieve at each grade level to be proficient within that given time period (U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). Nine states were initially approved for participation and began incorporating growth into their accountability models.

Pilot growth models could be classified into three categories (although different sources call these by different names): transition matrix models, trajectory models (also called growth to proficiency models), and projection models (U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). Transition matrix models measure growth in relationship to performance categories such as basic, proficient, and advanced (Auty et al., 2008). Student growth is indexed by movement from lower performance categories to higher performance categories across 2 years. Trajectory models calculate the difference from a performance target (i.e., test score) in some future grade to a baseline test score. The required gain is divided into annual goals that a student must attain to be considered on-target for proficiency (O’Malley et al., 2009). Projection or regression models estimate students’ future performance based on previous test scores from a reference cohort of students (Bonk et al., 2012). The regression coefficients from the reference cohort are used to calculate the future test scores for younger cohorts of students. A student is considered on-target to proficiency if the student’s predicted future performance is at or above the AMO.

All GMPPs augmented rather than replaced status models. In other words, the pilot states added students who were on-target toward proficiency via the growth model to students who were already proficient via the status model or “safe harbor” (or vice versa). The average increase in the percent of schools that made AYP as a result of using a growth model was 2% in eight of the pilot states; in the ninth pilot state, Ohio, there was 34% increase in the number of schools that made AYP using the growth model. Ohio uses a projection model and adds two standard error units to each predicted score (U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). This reduces the probability that a student who is on-track for proficiency is misclassified as not on-track; conversely, it increases the probability that a student who is not on-track for proficiency is misclassified as being on-track.

Current Growth Models

While the use of growth models to determine proficiency toward AYP was an important step, it was only a beginning. The problem with the pilot growth models is that they focused exclusively on proficiency. In other words, their sole purpose was to count more students scoring below proficiency on state assessments as being on-track to proficient. Thus, growth models in the GMPP did little to track growth in gifted students who scored at or above proficient on state assessments. Recently, state education departments have prioritized the development of growth models to focus on not only determining AYP in reading or language arts and mathematics but also to include other subjects and all students, including students identified as gifted. By 2010, 12 states used a growth model matching NCLB requirements and 13 states used a growth model based on state policy (Blank, 2010). The types and uses of growth models used for accountability purposes has expanded and can be defined as follows: “A growth model is a collection of definitions, calculations, or rules that summarizes student performance over two or more time points and supports interpretations about students, their classrooms, their educators, or their schools” (Castellano & Ho, 2013, p. 16). Castellano and Ho identified three primary growth model interpretations—growth description, growth prediction, and value-added. Growth description provides a growth metric about the magnitude of growth for an individual or group. Growth prediction provides information about the future scores of student given current and past achievements. Value-added provides information about what causes growth, for example particular educators and schools.

In 2011, the Council of Chief State School Officers (CCSSO, 2011) categorized five types of growth models currently used by states: categorical, gain score, regression, value-added, and normative. The first three models correspond to transition matrix (categorical), trajectory (gain scores), and projection (regression) models described in an earlier section. Value-added models are also regression models but take into account student or school characteristics. Value-added models predict expected growth and value-added occurs when the actual growth is more than the expected growth. These types of models require the use of a vertical scale, or performance scores that are comparable across time. Normative models compare student growth to that of a normative group to determine if the growth
or change in student performance is typical, higher, or lower than that of the norm group. Normative models do not require a vertical scale nor do they address whether the observed growth is adequate to reach the AMO.

Castellano and Ho (2013) provide guidance to districts and schools about the interpretation they can make when using a particular type of growth model. Categorical growth models can provide information about growth description and growth prediction because they index growth by examining if students make transitions among performance categories over time. For example, a student who moves from a lower performance category such as basic to a higher performance category would be counted as showing growth. Most gain score growth models are best suited to provide growth description but not prediction or value-added. Simple gain score growth models use only two time points and thus are not reliable for individual comparisons of gains. Regression-based growth models provide information about growth prediction because they predict a student’s future score based on previous scores or information from a cohort group. In regression models, the cohort used to estimate the regression coefficients must be as similar as possible to the cohort whose scores are predicted. Value-added growth models answer how students’ actual performance compares with their expected performance. In addition, teacher and student characteristics are entered into the model to provide information about the causes of growth. Normative models support interpretation of growth description and prediction because they compare students’ percentile ranks to a reference group of academic peers.

Recently, Goldschmidt, Choi, and Beaudoin (2012) conducted a growth model comparison study. Their findings indicated that the various types of models estimated school effects differently based on level (elementary or middle school) and state. In addition, models from which similar inferences about growth can be made were more likely to place schools in the same performance levels than models that differed in purpose. While this is true, none of the currently used growth models demonstrated extreme variations in school ratings—a school would not be rated as a high performer by one model and a low performer by another model. Other results, however, indicated that there is considerable variability in model results based on student characteristics that are beyond school control, such as disabilities, stability, and school size. Status models are most influenced by student characteristics whereas models that use prior information, such as value-added, are least influenced. The Goldschmidt et al. (2012) study concluded that different growth models answer different questions, and thus no single model can be touted as the model that produces the best results. Therefore, schools should consider what inferences they want to make from growth model results and choose the model that helps them address those inferences.

**Statistically Sound Growth Models**

The growth models discussed to this point are based on the definition provided earlier in the article by Castellano and Ho (2013), and were developed for NCLB accountability purposes. Most of these models however do not fit the statistical definition of growth modeling. Since gifted students often score well above the mean, many of the NCLB proficiency-based growth models are unlikely to capture accurately how gifted students are growing academically (McCoach et al., 2013). From a statistical standpoint, growth modeling implies that a very specific set of criteria be met: (a) there must be at least three observations or test scores, (b) the test scores should be comparable across time, and (c) a measure of time must be collected for every testing occasion (McCoach et al., 2013; Singer & Willett, 2003).

The first criterion, at least three test scores, allows us to estimate a growth path for each student. Every test score contains some error (Crocker & Algina, 1986). Sometimes the scores that we get from students are too low because of that error (e.g., the student was distracted during the test), and other times the scores are too high because of error (e.g., the student guessed correctly). With only two time points, we can only connect the dots. There is no way to know what part of the score was based on the student’s actual knowledge and what part of the score was because of error. With three or more time points, we can estimate a line that best captures all of the points, which allows us to approximate the student’s true score at each time point (Cronbach & Furby, 1970; Raudenbush, 2001; Rogosa, Brandt, & Zimowski, 1982).

The second criterion, test scores should be comparable across time (Singer & Willett, 2003), is much harder to obtain. One simple way to meet this criterion is to administer exactly the same assessment at every time point. However, a test that might be appropriate for a second-grade mathematics student to assess mastery of mathematics material is unlikely to be an appropriate measure of mathematics mastery for a sixth-grade student (Martineau, 2006; McCoach et al., 2013). Therefore, those interested in growth over longer periods of time typically rely on assessments that have been placed onto the same metric through a process called vertical scaling. Vertical scaling assumes that the construct being assessed does not change across time and links student scores so they can be compared on a common scale (Singer & Willett, 2003). This process requires that a large number of students take the various assessments and that the test items are subjected to a process that essentially vets the items for specific characteristics such as difficulty and the ability to separate those students who know the material well from those who do not. This process would be impractical for classroom teachers or even administrators.

The final criterion is an accurate measure of time (Singer & Willett, 2003). This can be collected in lots of different ways, for example, the student’s age at each test administration, the months between test administrations, or even years between administrations. The time that the students took the tests does not have to be the same for all students (e.g., some students could take the assessment in September and others in November; the differences in time can be incorporated into the statistical model), but when each student took each assessment must be known (McCoach et al., 2013).
Implications for Gifted Learners Who Are Above Proficient

After all of the above criteria are met, we can use statistical growth modeling to illustrate how students are growing. We often run into some problems when we try to illustrate how gifted students are growing. Here are the three most common ones.

First, most assessments are designed to measure typical students, and, as a result, gifted (or high scoring) students' test scores are more likely to contain error (McCoach et al., 2013). Test items are typically written to see how much average students know. There are usually very few items on an assessment that can be used to determine how much gifted students know. So, we might get a pretty good idea of what an average student is capable of doing because there are a lot of items to examine. However, gifted students might only find a handful of items that are challenging to them. Therefore, missing or guessing correctly just one of those challenging items introduces more error into their scores. Essentially, we are trying to tease out what these students know with fewer items than we use to determine what an average student knows. There are a couple of ways to prevent this (McCoach et al., 2013). One is to give students above-level tests because the performance of the gifted student is more likely to align with the older group for whom the test was designed. Another approach is to use computer-adaptive testing, so that the gifted student takes an assessment that tracks the student's progress—getting increasingly more difficult until their “sweet spot” is determined. This allows the student's score to be essentially as accurate or contain about as much error as an average student.

Ceiling effects in assessments are also common for gifted students (McCoach et al., 2013; Wang, Zhang, McArdle, & Salthouse, 2008). One example of this is when a gifted student takes a test designed for average students and knows so much that the gifted student gets a perfect or near perfect score. The gifted student simply cannot show growth on a later assessment because the student already maxed out the score on the first assessment. We can address this issue in the same way that we can address the previous problem—above-level testing or computer-adaptive tests (McCoach et al., 2013).

Another problem that is common with gifted students’ scores is regression to the mean (Lohman & Korb, 2006; McCoach et al., 2013). Statistically speaking, this is the tendency for people with extreme scores (e.g., two standard deviations above the mean or the 98th percentile) on an initial assessment to have predicted scores on future assessments that look more like average students (Campbell & Kenny, 1999). Because gifted students are more likely to have extreme scores, they are more susceptible to regression to the mean (Lohman & Korb, 2006).

Implications for Policy Makers and Administrators

With many Race-to-the-Top initiatives, policy makers and administrators are trying to capture teacher effectiveness by calculating how much students taught by the same teacher grew academically (American Recovery and Reinvestment Act, 2009). Based on some of the problems listed above, we can see that teachers of gifted students are likely to have students who show less growth than teachers of average students.

This is not intended to be a get-out-of-jail free card for teachers of gifted students. Simply because it is not easy to assess the growth of gifted students does not excuse educators from doing so. Furthermore, educators should be fully aware of what academic assessments can and cannot tell us about growth and thus choose assessments wisely. Following the recommendations of McCoach et al. (2013), we also endorse the use of vertically scaled above-level tests or computer-adaptive tests for tracking the academic growth of gifted students. In the absence of either type of test, gifted students’ tests scores are more susceptible to problems with measurement error (Lohman & Korb, 2006). Educators should examine the reliability and validity properties of any assessment prior to use and know the purposes and limitations of the test being recommended for use (see Carlson, Geisinger, & Spies, in press; Murphy, Geisinger, Carlson, & Spies, 2011; Robins & Jolly, 2011).

Even if administrators have access to data from assessments that are more likely to accurately assess the growth of gifted students, we recommend supplementing growth scores obtained from these assessments with other types of information, such as scores on both traditional and alternative classroom assessments. Adding to the problem, many academic subjects such as social studies or art have few or no assessments available for tracking student growth. Therefore, teachers of gifted students will likely need to take some extra steps to illustrate their students’ academic growth. We recommend the following: (a) assess students at least twice—pretest and posttest—with the intention to move to assessing students at least three times, (b) use more nuanced assessments that are less likely to have issues with ceiling effects, and (c) illustrate student growth.

To assess growth, students must take the same exam or vertically scaled exams three times at a minimum. For administrators who want to move teachers toward assessing student growth, asking them to assess students three times over the same content may be too big an initial step. Therefore, we recommend that teachers start by giving pretests and posttests, if they are not already doing so. This will enable teachers to determine what students knew prior to instruction and how much the students changed as a result of instruction. However, we should add the caveat that if students are only assessed twice, this will only capture change and not growth as statistically defined previously in this manuscript.

The classroom assessments will need to be more nuanced so they can be more sensitive to student change. The assessments should be the same for the pretest and posttest. We encourage teachers to find ways to integrate these assessments into the classroom and not take up substantial amounts of class time. A teacher could base the assessment on a critical question(s) that is essential to the student demonstrating mastery of a concept.
Here are a couple of examples to help illustrate this idea. First, a mathematics teacher who wishes to know how much her gifted students know about the concept of linear change could ask students at the beginning of the unit to write down all of the possible ways that they could use to determine if the relationship between two variables were linear. At the conclusion of the unit, the teacher could imbed that question into the final assessment and compare the responses both in the number of ways and the quality of the descriptions at the two time points. For the second example, an English teacher who wishes to know how much her gifted students know about how to evaluate a source could ask his students to write as many questions as possible to ask to verify that a source is credible. Similar to the mathematics example, the teacher could then imbed that question on the final assessment and compare the responses. Both of these examples would take away very little classroom instructional time but allow for demonstration of what students knew prior to and after instruction.

Finally, educators should illustrate the students’ growth. To do so, the teacher could collect all of the assessments and summarize student performance both before and after instruction. By using assessments frequently and as an integrated component of instruction, multiple examples would be available to demonstrate student change or growth. These examples might include summary data (e.g., in our example, the number of ways students listed to tell if a relationship were linear before and after instruction or if a source were credible) and samples of students’ answers (e.g., several students answers prior to and after instruction).

These recommendations are mostly a reiteration of best practices for classroom assessments for all students, namely, frequent assessments that are imbedded within instruction (Black & Wiliam, 2010) that take into account how to use these assessments to illustrate student growth. These assessment data should be viewed as complementary to the growth data from more formal standardized assessments.

In summary, this article sought to provide an overview of the various types of growth models that are used and how they evolved, the requirements for statistical growth modeling, some potential problems with assessing the growth of gifted students, and recommendations for assessing the growth of gifted learners. It is our hope that the appropriate use of assessment data will help teachers, administrators, and policy makers better understand how best to meet the academic needs of our gifted students.

Glossary

**Categorical growth models**: Express growth as the change in a student’s performance category placement from one year to the next. For example, a student who places in the below-basic category in year one but in the basic category in year two would be showing growth, even though the student did not place in the proficient category. In the categorical growth model schools are rewarded for students who are closing the gap toward or maintaining proficiency (if already proficient). Unlike other growth models, the categorical growth model does not rely on standard scores; rather, it focuses on the standards-based performance levels established by states.

**Gain score growth models**: Calculate a gain score for each student and use it to determine how much growth in achievement a student is making. A gain score is calculated by finding the difference between two scores at different points in time. For this difference to be meaningful the two test scores must be on a common scale. For example, if the gain score is calculated at two different grade levels, a vertically scaled assessment should be used. The gain score growth model compares the student only to himself or herself.

**Normative growth models**: Compare a student’s growth with that of other students with historically similar scores. The model uses quantile regression to predict a student’s future score. Predictors in this model are the student’s prior scores with the student’s current score expressed as a growth percentile. The growth percentile is computed based on the scores of students with similar prior scores, not to the scores of all students. This type of growth model provides easy to interpret information. For example, a growth percentile of 42 is interpreted as 42% of scores falling below this particular score. The model can also provide the amount of growth the student must attain to reach a target-scaled score in a future grade.

**Regression growth models**: Use linear regression to predict a student’s future test score. The model uses test scores from a past cohort of students to estimate a prediction equation. The prediction equation is then used to predict future test scores for the current cohort. The predicted scores are compared with a standard and if the predicted score is above this standard then the student is making adequate growth.

**Value-added growth models**: Control for the influence of selected variables such as demographic information or prior performance on student achievement. Controlling for these variables allows one to say the growth beyond the controlled variables is the value added by the school. In other words, the value-added portion of the model is the growth that goes beyond what one expects.

**Vertical scales**: Allow scores on an achievement or other type of test to be comparable across grade levels. Test scores are linked across several grade levels so that scores from higher grade levels can be meaningfully compared with test scores in previous grade levels. Another feature of vertical scales is that they are designed to be equal-interval scales. In other words, an increase of 25 points from 300 to 325 is the same as an increase from 550 to 575 and so on. The National Assessment of Education Progress assessments are an example of vertically scaled assessments and are available in many subject areas including mathematics, writing, and reading.

Conflict of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

References
Ryser, G. (2008, April). No Child Left Behind (NCLB) growth model pilot program. In S. K. Johnsen (Chair), Policy and legislative issues: Their effects on gifted and talented students. Symposium conducted at the meeting of Council for Exceptional Children-Talented and Gifted (CEC-TAG), Boston, MA.
Bios
Gail R. Ryser is director of the Testing, Research-Support, and Evaluation Center at Texas State University. She is an Associate Editor of Gifted Child Quarterly. Her research focuses on mathematics education, gifted education, and measurement.
Karen E. Rambo-Hernandez is an assistant professor at Colorado State University in the School of Education and the School of Teacher Education and Principal Preparation, where she teaches courses in teacher preparation and educational statistics. In her research, she is interested the assessment of student learning, particularly the assessment of academic growth and modeling of student growth data.