Alternative Methods for Computing Growth Norms

Joseph J. Stevens

Joseph F. T. Nese

Gerald Tindal

University of Oregon

Key words: Achievement growth, student growth percentiles, growth norms, performance benchmarks

Paper presented at the National Council on Educational Measurement conference, April, 2015, Chicago, Ill.

Address all correspondence to Joseph Stevens, University of Oregon, Department of Educational Methodology, Policy, and Leadership, College of Education, 5267 University of Oregon, Eugene, OR 97403-5267; 541-346-2445, stevensj@uoregon.edu.

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R32C110004 awarded to the University of Oregon. The opinions expressed are those of the author and do not necessarily represent views of the Institute or the U.S. Department of Education.

Abstract

The purpose of this paper is to illustrate and compare several methods for describing and interpreting student achievement growth. There is a long history of the use of normative methods to represent performance on a variety of educational and psychological tests as well as in medical applications like physical measures of child development (i.e., height, weight). The conventional method used to construct norms in these traditional applications is to calculate empirical percentiles from cross-sectional data at one or a series of time points and then fit and smooth the data to produce normative reference centiles or deciles (see Angoff, 1984; Peterson, Kolen, & Hoover, 1989). More recently, Betebenner (2009) has popularized a different kind of percentiles for enriching interpretation of student achievement performance, student growth percentiles (SGP). SGP are actually better described as conditional regression models in that they do not actually examine growth but condition current performance on prior performance. A number of states and school jurisdictions have adopted SGP as a primary mechanism to report and interpret student achievement growth and, in some cases, to aid in evaluating student, teacher, or school performance for accountability purposes. A third class of methods for calculation of growth norms is the application of more elaborate statistical models to represent change such as multilevel growth models (MGM). A potential advantage of MGM is that they actually use longitudinal data rather than cross-sectional or conditional data to represent growth. Using interim assessment data from a large Arizona school district, we demonstrate all three methods (traditional, SGP, multilevel growth models) for estimating student growth norms and we then discuss advantages and disadvantages of each method.

Alternative Methods for Computing Growth Norms

The purpose of this paper is to illustrate and compare several methods for describing and interpreting student achievement growth using data from a large school district's interim assessment. "What does this score mean?" is perhaps the fundamental question asked by test users and consumers. Meaningful answers to this question depend on putting a score into context by supplying information about what content was covered on the test, what testing methods were used, and the scale on which the score is reported. Normative methods have long been used to represent growth in educational, psychological, and medical applications like academic achievement, social and psychological functioning, and physical measures of child development (i.e., height, weight). Norms provide one of the fundamental methods that allow more meaningful interpretation of score information and allow the user to compare an individual's performance to a standard benchmark to enhance interpretation (Bloom, Hill, Black, & Lipsey, 2008). In these applications, norms are usually calculated from cross-sectional data for individuals and then fitted empirically using statistical models. Often model results are also then averaged or smoothed to produce norms that are more consistent with theoretical expectations (e.g., normally distributed).

In educational and psychological testing there is a longstanding tradition of computing norms to provide an interpretive context for individual performance on achievement tests (e.g., CTBS, ITBS, Stanford), intelligence tests (e.g., Stanford-Binet, WISC-III), and other measures of psychological and social functioning (e.g., Wechsler Memory Scale, Kohn Social Competence Scale, MMPI). In all of these applications, norms are developed by engaging in some sampling process to select the norms group, administering the test or instrument under standard conditions, and then analyzing results to produce reference information showing the average performance of individuals at different locations within the score distribution. An extensive variety of physiological and developmental characteristics have also been analyzed to produce normative information for developmental, clinical, and medical applications (see Cole, 1988; Wright & Royston, 1997).

In many of these applications, the fundamental interest is in representing normative patterns of growth or development. Nonetheless, almost all norms development has historically used cross-sectional data. For example, in the development of the Wechlser Memory Scale norms (Pearson Education, 2008), during the instrument standardization process, all individuals were tested in the same year. Age based norms were developed by grouping data according to examinee age. Obviously tracking individuals over many years to create a true longitudinal sample is not feasible for this instrument. However, in many other applications it is possible to track individuals over time and create norms based on the actual longitudinal course of their performance. It is uncertain, however, whether there are any appreciable differences between cross-sectional and longitudinally based norms.

Given the recent upsurge in interest in the growth of academic achievement in education, there is some question regarding the best ways to represent and interpret growth. One recent development that has attracted substantial interest is the Student Growth Percentiles (SGP) method described by Betebenner (2009). A number of states and school jurisdictions have adopted SGP to report and interpret student achievement growth.

Another approach to create growth norms that is less commonly applied, is to use common growth modeling methods. One of the most widely used statistical approaches to modeling growth is multilevel models (Raudenbush & Bryk, 2002; Singer & Willett, 2003). These models estimate a growth trajectory for each individual and can be used to evaluate the rate of growth, individual differences in growth parameters, and predictors of growth. These estimated growth trajectories can also be used to develop normative descriptions of the growth function at each point in time as well as normative descriptions of slope or rate of growth across the period of time studied.

In this paper we describe and illustrate three alternative methods for creating growth norms from a longitudinal sample. Following application and illustration of each method, we draw attention to their relative advantages and disadvantages. It is also useful to distinguish between two broad classes of growth models. Briggs & Betebenner (2009) define *absolute growth models* as models that estimate growth conditional on a time function (e.g., linear, quadratic, nonparametric). They define *relative growth models* as models that estimate growth based on prior achievement. In this paper, two methods presented are examples of absolute growth models: Traditional Growth Norms (TGN) and Multilevel Growth Model (MGM) norms. The third method presented here, Student Growth Percentiles (SGP) is variously described as a relative growth model (Betebenner, 2009) or a conditional status model (Castellano & Ho, 2013) but does not exactly represent growth in the sense that growth is usually defined by developmental theorists or applications of typical longitudinal methods (Willett, Singer, & Martin, 1998). The three alternative models include differences in data and measurement requirements, degree of flexibility in application, dependency on scale properties (i.e., ordinal or interval measurement), and requirements for vertical linking over time, required sample size, and composition of the sample. Thus our purpose was to estimate achievement test norms using the three models (TGN, SGP, and MGM) and compare and contrast the three models in terms of the resulting norms.

Method

Participants

Results presented in this paper are based on three cohorts of fifth grade students taking a district mandated interim assessment in a large school district in Arizona.[1] Data included seasonal (fall, winter, spring) interim mathematics and reading assessments administered in Grade 5. A total of three years (cohorts) of data were included to increase sample size. Although cohort differences and cohort stability are of interest, those issues are part of different studies and we do not separate cohorts in any of the analyses presented here.

A total of 3,985 students across the three cohorts were included in the study. Of these, 3,949 students (99%) had at least one mathematics score and 3,947 students (99%) had at least one reading score. Sample demographics for the analytic sample are displayed in Table 1. The demographics of the sample were characterized by a small percentage of White students and a large percentage of minority students: Hispanic (52%), Black/African American (11%), and American Indian/Alaskan Native students (8%). A large percentage of district students (75%) were eligible for free or reduced price lunch. A substantial proportion of students were also English language learner students (36%) who were classified either as active language learners or exited students being monitored. All ELL students in the district were required to take a test of English language proficiency. Those testing below a specified threshold were enrolled in a

[1] We also conducted analyses on a second sample composed of the annual accountability test, reading/language scores from a cohort of Oregon students who were in the third grade in 2008, 4th grade in 2009, 5th grade in 2010 and 6th grade in 2011. Results for this sample closely parallel those reported here, but in the interests of space and time are not reported. These results are available on request from the first author.

Table 1. *Sample Demographic Characteristics.*

Variable	N	%
Female	1,911	48.0
Special Education	465	11.7
Race/Ethnicity		
American Indian/Alaskan	321	8.1
Asian	102	2.6
Black or African American	438	11.0
Hispanic	2056	51.6
Hawaiian/Pacific Islander	24	0.6
Multiple Categories	120	3.0
White	924	23.2
English Language Learner		
Active	310	7.8
Monitor	1,110	27.8
Free/Reduced Lunch	2,971	74.6

district-wide structured English immersion (SEI) program. After students reached proficiency on the English language assessment, they were placed in a general education classroom, but were monitored for two additional years.

Variables. Students' scores on the mathematics and reading subtests of the Measures of Academic Progress (MAP; Northwest Evaluation Association, 2011) served as the outcomes for this study. The MAP is administered seasonally (fall, winter, spring). Means and standard deviations for each time point are displayed for the analytic sample in Table 2.

Table 2. *RIT Scale Score Means (M) and Standard Deviations (SD) for the MAP Mathematics and Reading Tests.*

	Mathematics		Reading	
Time	*M*	*SD*	*M*	*SD*
Fall	209.06	14.07	203.53	15.22
Winter	215.06	15.03	207.67	14.72
Spring	221.23	16.24	211.42	14.86

The MAP is an untimed computerized adaptive test on which each student is presented different items conditional on his or her estimated ability level. The adaptive algorithm results in consistently higher test information and lower standard errors across a wide range of student abilities (NWEA, 2011). The tests include 50 multiple-choice items with 4 or 5 response options. All items on the MAP were calibrated on a common, vertical scale, using a one parameter, IRT (Rasch) model (NWEA, 2011) that produces scores on a transformed logit or Rasch unit (RIT) scale. Student growth can therefore be tracked both within and across school years.

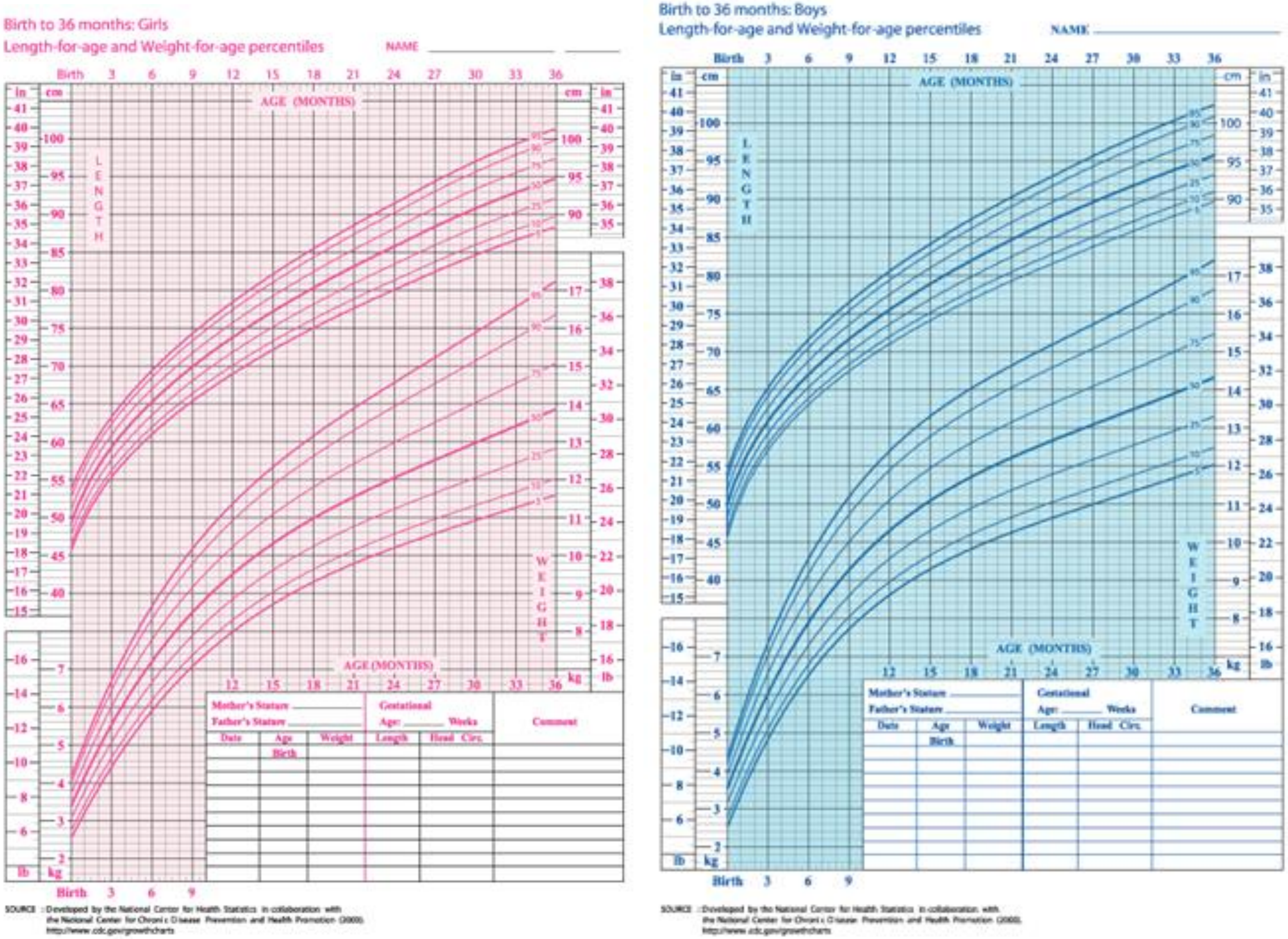
Analytic Models

As described earlier, our purpose was to examine three alternative methods of providing benchmarks for the interpretation of students' academic growth: (a) the traditional approach to norms (TGN), (b) student growth percentiles (SGP), and (c) multilevel growth model (MGM) norms.

Traditional growth norms (TGN). There is a long history of using empirically derived norms as a benchmark for interpretation and comparison of a measurement. These methods have been used for many years on intelligence tests, achievement tests, and measures of social competence or psychological functioning (Anastasi & Urbina, 1997) as well as in many health science applications. One example of the latter approach to normative interpretation, well known to parents, is "pediatrician norms" (see CDC examples in Figure 1; Kuczmarski, Ogden, & Guo et al., 2002), that provide a means to compare the height or weight of a child to a large national sample of other children. Two common uses of this type of normative information are (a) to track developmental progress, and (b) identify individuals at extremes of "reference" intervals or at risk for low performance or slow development. Also note that a key feature of absolute growth model norms is the ability to locate an individual's performance or status at a given point in time and determine how that performance relates to the normative sample.

TGN are almost always based on cross-sectional samples of individuals, not longitudinal examination of individuals over time. That is, to create norms for elementary school students, a cross-sectional approach might analyze data from 3rd, 4th, and 5th graders from a single academic year instead of tracking 3rd graders for two succeeding years into 5th grade. With large sample sizes and stable composition of cohorts, there may be little difference between cross-sectional and longitudinal norms.

In traditional applications, one of the key considerations in building norms is the design of the norming sample. Sometimes almost complete populations are used to develop norms, but more commonly sampling methods are used to select a more manageably sized group. Angoff (1984) and Lord (1959) describe a number of sampling issues in the development of test norms. However, the utility of norms depends on the intended use and it often may be more relevant to make comparisons to a local group or subgroup. Different norms may be useful for different purposes or when there are substantial differences in the composition or characteristics of the local group and the norming sample. It is also important to note that when criterion-referenced or standards-based interpretations are required, it will usually be necessary to relate normative results back to criterion benchmarks of performance.



SOURCE : Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000). http://www.cdc.gov/growthcharts

SOURCE : Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000). http://www.cdc.gov/growthcharts

Figure 1. CDC infant growth norms—length and weight by age and sex.

In the development of many norms, deciles or percentiles are computed empirically. In wholly empirical norms development, especially when sample size is large, norms development primarily involves the description of the score distribution. Additionally when more complex sampling procedures have been used, it may be necessary to apply sampling weights to the observed scores in order to maintain representativeness (Peterson et al., 1989). However, in many applications, especially where there is a belief that the norms represent underlying theoretical entities, additional statistical procedures are used to fit statistical models to the data, and then to transform and/or smooth functions before creating final norms (Kuczmarski et al., 2002).

Student growth percentiles (SGP). Student Growth Percentiles (SGP; Betebenner, 2009) have become increasingly popular as a method of summarizing student and school performance in recent years and are currently used as a primary analysis of performance in dozens of states. The SGP method provides information on a student’s relative rank in a conditional score distribution based on current year performance regressed on the student’s scores in one or more previous years. Another characteristic feature of the SGP approach popularized by Betebenner (2009) is the use of ordinal models (i.e., quantile regression) as well as B-spline, cubic polynomial smoothing of distributions. The analysis can be implemented using the SGP package in *R* (Betebenner & Iwaarden, 2011). The SGP method is founded on the assumption that the score scales commonly used in educational testing do not have interval properties which leads to the need to use ordinal analysis methods. As discussed earlier, it is also important to keep in mind that SGP are a *relative growth model* in which student performance is not conditioned on time but on prior achievement. Thus an individual’s status or performance is compared to the current measurement occasion norm distribution that is conditioned on previous occasions but it is not possible to locate individuals on a temporal scale. Rather, a new SGP conditional distribution is generated for each new measurement occasion.

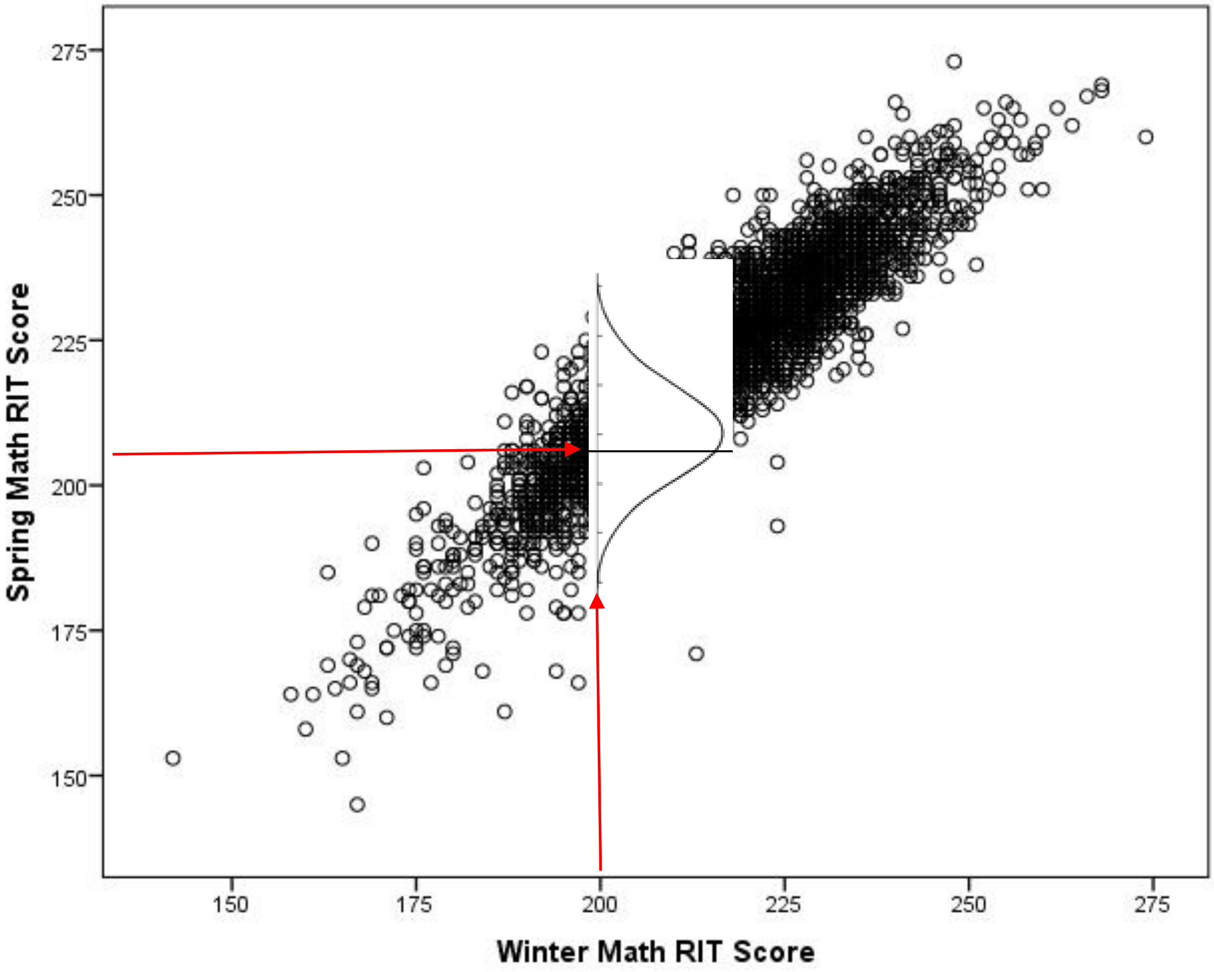


Figure 2. Illustration of the SGP method conditioning spring math RIT scores on winter math RIT scores.

Figure 2 illustrates the SGP process using only one conditioning score for simplicity. The figure depicts an example in which the student had a winter math RIT score of 200. This prior score defines a conditional distribution of all students who had a winter score of 200. The same student had a spring score of 205 which is then located within the “winter 200” conditional distribution and the percentile rank of the spring score of 205 is then determined in the conditional distribution (~45% in this illustration).

Multilevel Growth Model (MGM) Norms. Another method of representing student growth rests on the statistical modeling of change over time. Multilevel growth models (MGM) have become an increasingly common method for estimating change over time (Raudenbush & Bryk; Singer & Willett, 2003). These models are absolute growth models in that they relate change to a time function and maintain the metric of the score scale as an outcome and metric of change. As a result, a vertically linked score scale is necessary to support valid interpretations of results. We demonstrate the application of MGM here on using a two level MGM (time nested within student). Results of this model are then used to create normative deciles describing student academic growth. These MGM analysis produces ordinary least squares (OLS) and empirical Bayes (EB) estimates of each student's growth function.

Results

We applied the three alternative methods to the within-year, seasonal interim assessment scores of Arizona students on the MAP mathematics and reading tests. For the TGN, we created empirical growth norms by calculating deciles at each measurement occasion (fall, winter, spring) for the longitudinally matched sample of students described above. Figure 3 shows normative growth for mathematics achievement in the panel on the left and for reading achievement in the panel on the right. Each line in the figures represents a decile (percentile ranks of 10, 20, 30, 40, 50, 60, 70, 80, and 90) for the within-year MAP test with the median or 50th percentile presented using a black line.

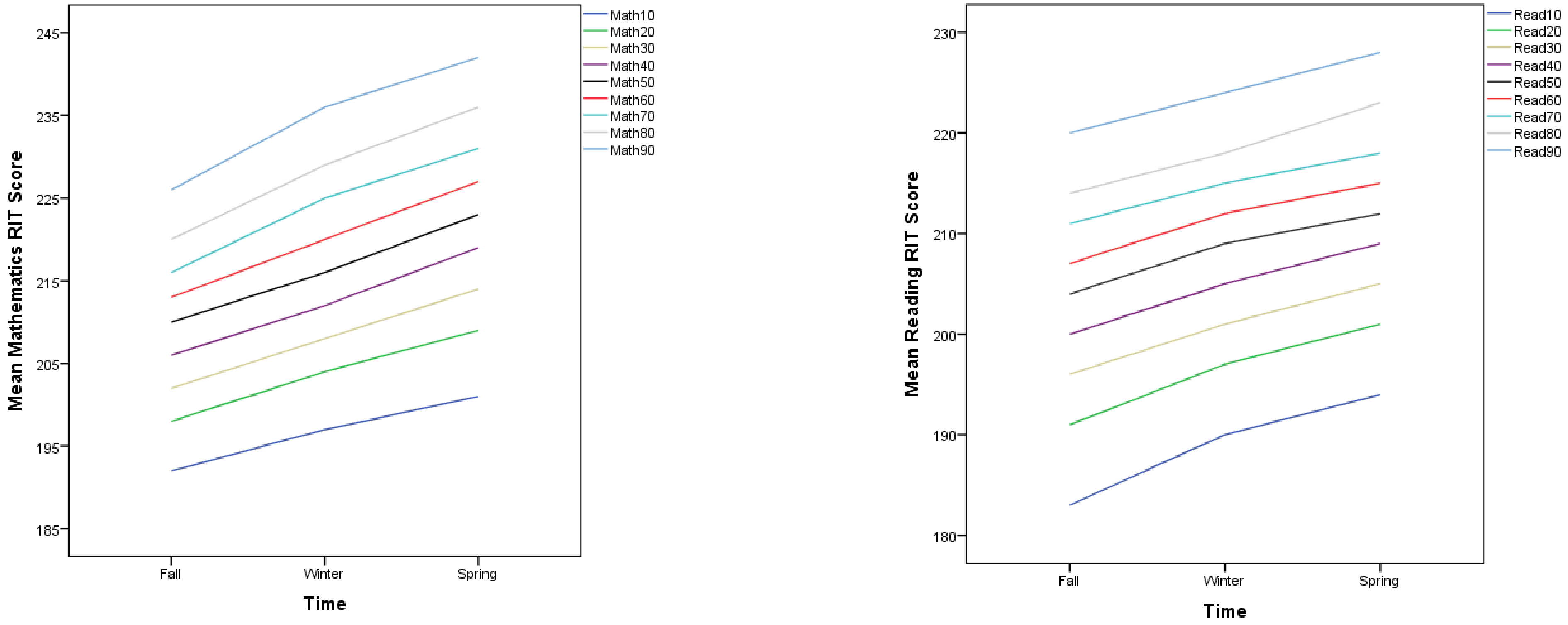


Figure 3. Traditional growth norm deciles for mathematics (panel on left) and reading (panel on right).

Next, we analyzed the same data using the SGP procedure as implemented with the *R* package SGP. SGP produces a conditional percentile rank for each student that, in our case, represented the student's relative standing in the residual score distribution produced by that portion of spring performance not predicted by the student's winter and fall scores. Thus, in a distribution of spring MAP scores residualized with respect to winter and fall scores, each student received a percentile rank corresponding to their relative standing in the distribution of residuals. An SGP was calculated for each student's spring mathematics and reading scores conditioned on their winter and fall scores.

Figure 4 shows the SGP results for mathematics in the upper panel and reading in the lower panel using the fall and winter MAP scores to predict the spring MAP scores. The red lines in each figure define SGP deciles within the spring score distribution. As can be seen in the figures, there are a range of spring RIT scores that receive any particular SGP percentile depending on the previously obtained scores. Also apparent is the relative growth nature of SGP in that the results have no direct reference to the time at which previous scores were obtained.

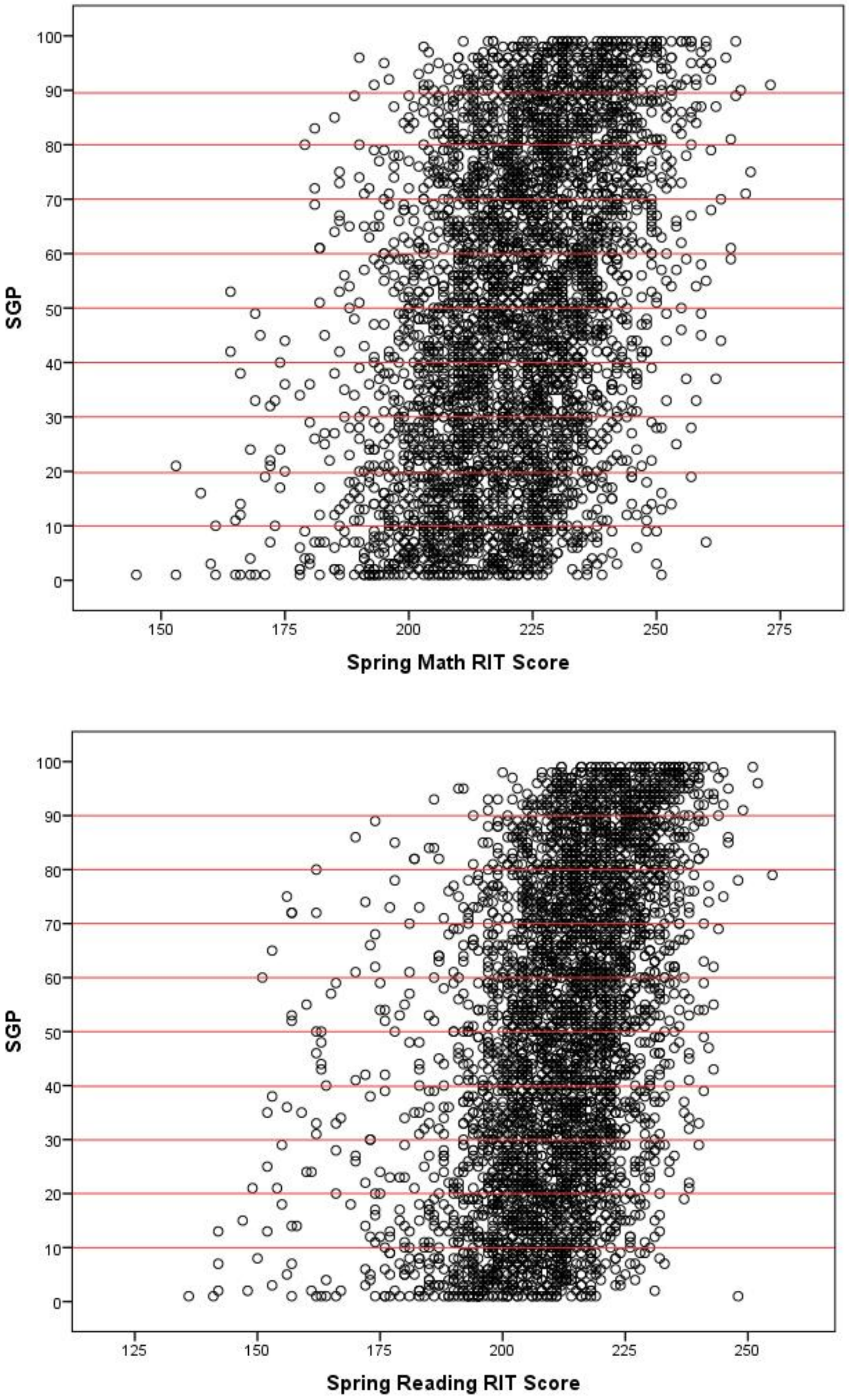


Figure 4. SGP deciles for mathematics (upper panel) and reading (lower panel).

Finally, we applied multilevel growth models to the within-year interim mathematics and reading assessment data. The multilevel models were two-level, linear models estimated using HLM 7.0 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). Because there were only three measurement occasions, quadratic or other functional forms were not examined. Time was coded 0, 1, or 2 to reflect the fall, winter, and spring test administrations. The HLM models were specified as:

Within-person, level-1 (measurement occasions, 1-t):

$$\text{MAP Score}_{ij} = \beta_{0j} + \beta_{1j}\,(\text{Time}_{ij}) + r_{ij}$$

Between-person, level-2 (persons, 1-i):

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

After estimating the MGM we calculated growth norms by taking the deciles of student's growth functions calculated using either ordinary least squares (OLS) or empirical Bayes (EB) estimation. Figure 5 below shows the EB estimated deciles for mathematics and reading.

The TGN and MGM methods allow comparisons of student performance at any time point in the growth trajectory while SGP provide normative information only for the current year. Therefore for purposes of comparison, we compared the percentile ranks (PR) of the spring TGN norms, the SGP norms, and the spring EB estimated MGM norms. But because another advantage of the MGM method is the creation of estimated growth rates (slopes), we also calculated PRs for both OLS and EB estimates of growth. Because understanding SGP can be complex, we also calculated conditional regression models following Castellano and Ho (2013). The conditional regression model was an ordinary least squares multiple regression with

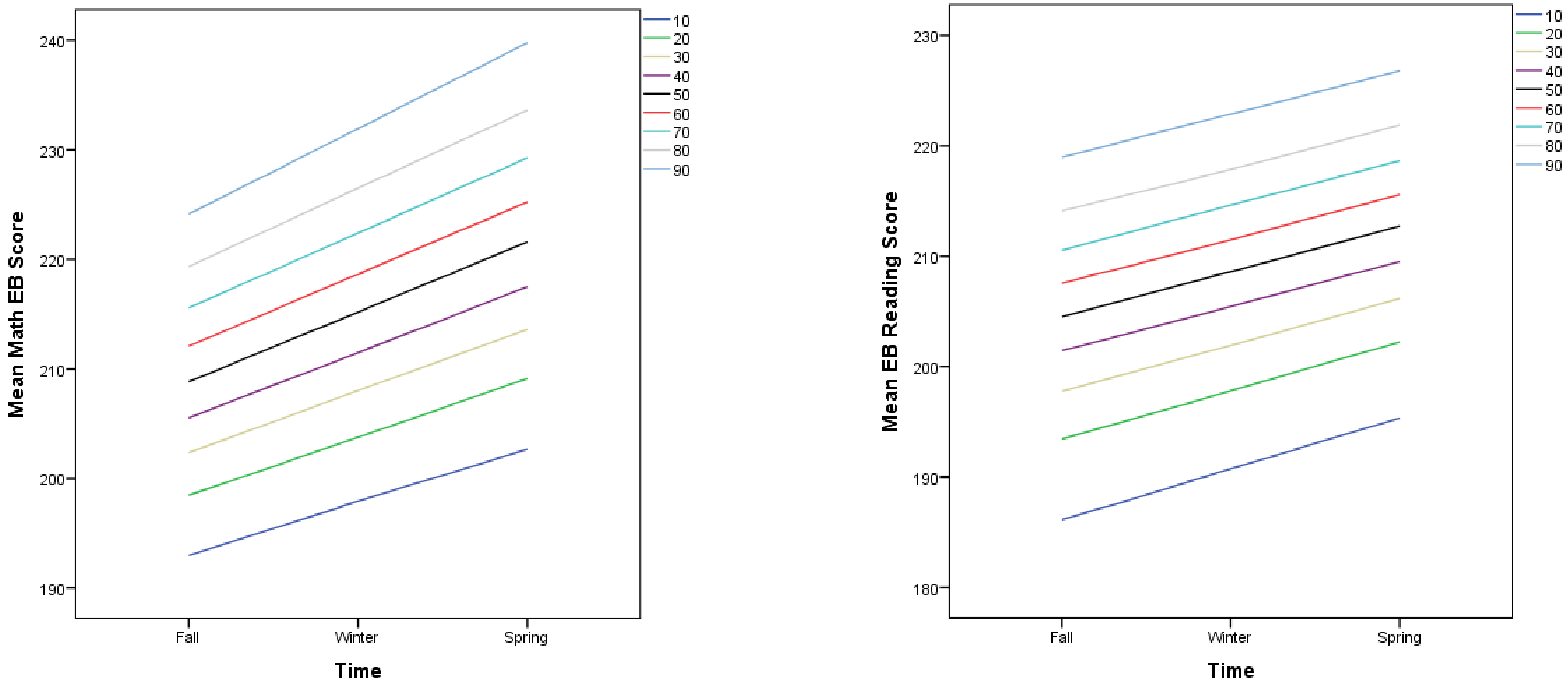


Figure 5. MGM empirical Bayes deciles for mathematics (panel on left) and reading (panel on right).

fall and winter MAP scores predicting spring MAP scores. Following analysis, the model residuals were calculated (CR in the tables below) for comparison to the SGP. The only real difference between CR and SGP results should be estimation methods (OLS vs. smoothed quantile regression). This provided an additional benchmark for understanding the SGP results.

Table 3 shows correlations among the percentile ranks for these alternative methods with the mathematics measures on the lower diagonal and reading measures on the upper diagonal. It

Table 3. *Correlations among Norming Methods, Mathematics in the Lower Diagonal, Reading in the Upper Diagonal*.

	TGN spring	MGM spring	SGP spring	CSR spring	OLS slope	EB slope
TGN spring	1.00	.96	.46	.44	.21	-.09
MGM spring	.98	1.00	.19	.18	-.02	-.31
SGP spring	.39	.20	1.00	.99	.81	.70
CSR spring	.40	.21	.99	1.00	.80	.69
OLS slope	.50	.35	.82	.82	1.00	.93
EB slope	.94	.87	.55	.57	.74	1.00

can be seen that, as expected, the CR and SGP ranks are almost identical (*r* = .99). The correlation between SGP and the slope estimates is also fairly high for OLS slope (*r* = .82 mathematics and *r* = .81 reading) but substantially lower for EB slope (*r* = .55 mathematics and *r* = .70 reading). Correlations between the two slope estimates were higher in reading (*r* = .93) than in mathematics (*r* = .74). In mathematics, correlations between the TGN spring percentile ranks were high for MGM spring estimates (*r* = .98) and EB slope (*r* = .94) but substantially lower for SGP (*r* = .39) and OLS slope (*r* = .50). In reading, the correlation between the TGN spring percentile ranks was also high for MGM spring estimates (*r* = .96) but near zero for EB slopes (*r* = -.09). In reading, the correlation with SGP ranks was similar to mathematics (*r* = .46) and lower for OLS slope (*r* = .21). Examination of these correlations makes clear that relations may differ from mathematics to reading and there may be substantial differences between the alternative norming methods.

As a further comparison of methods, Table 4 shows students' percentile ranks based on the different models for five students randomly sampled from the results for mathematics (upper table) and reading (lower table). The purpose of the tables is simply to further demonstrate how rankings may differ as a function of the norms method used. For example, in mathematics, student A's percentile rank would be 55, just above average, using the MGM spring score estimate, but only 8, substantially below average, using the OLS slope estimate. Or, another example from the reading table, Student F has a percentile rank of 79 based on the MGM spring estimate but only 2 based on the SGP estimate.

Table 4. *Mathematics and Reading Percentile Ranks by Norming Method for Five Students Randomly Selected in each Content Area.*

Mathematics

Student	TGN	MGM	SGP	CR	OLS slope	EB slope
A	43	55	10	13	8	25
B	34	37	34	36	29	30
C	73	60	63	65	43	60
D	66	39	79	78	33	47
E	73	84	21	22	38	64

Reading

Student	TGN	MGM	SGP	CR	OLS slope	EB slope
F	57	79	2	5	9	6
G	29	40	11	12	18	19
H	45	50	32	33	30	29
I	94	96	49	46	60	40
J	51	54	42	40	74	74

Discussion

We computed and compared three different methods of representing normative growth on school district interim assessments in mathematics and reading. An important initial consideration in reviewing these methods is the distinction between absolute and relative growth models. Of the three methods considered here, TGN and MGM represent absolute growth models while SGP represent a “relative growth” model (Briggs & Betebenner, 2009). Thus an important early consideration in choosing among these methods is what is meant by “growth” and what are the purposes of using growth norms. The three methods provide different information about student progress and answer different questions about student performance and academic growth. These three methods depend on different assumptions and have different data requirements as well.

Traditional Growth Norms

TGN are well known to many users and are likely to be more transparent and easier to understand than the other methods considered here. TGN provide information on absolute growth and are based on empirical description of either cross-sectional or longitudinal data. TGN are often refined based on assumptions about the underlying theoretical distribution of performance that lead to use of various smoothing and/or estimation methods. The TGN method allows the user to readily locate an individual performance at a particular point in time and make comparisons to the larger norms sample. Another important strength of the TGN method as usually implemented in educational and psychological testing is an emphasis on careful sampling (e.g., stratified random sampling) to ensure the representativeness of the norms group (Peterson et al., 1989). To our knowledge, within the area of standards-based achievement tests used for accountability purposes, these sampling procedures are not often used.

As usually implemented, TGN allow evaluation of empirical growth and the opportunity to apply model fitting and smoothing methods, apply sample weighting to ensure representativeness when appropriate, and represent a growth function across a defined temporal interval. TGN methods may require large samples (depending on purpose and kind of norms), a constant vertically linked scale over time, and representative sampling methods. TGN methods usually do not report fitted model results and therefore do not allow examination of growth rate at the individual or aggregate level (although there is no reason this could not be done).

Student Growth Percentiles

SGP provide information on the relative ranking of students on this year's test conditioned on the scores received in a previous year's test(s). The name SGP is somewhat misleading and likely creates some difficulties and misunderstanding in interpretation. SGP are described as a relative growth model by Betebenner (2009) but Castellano & Ho (2013) argue that the term "conditional status percentile ranks" is a more accurate description. This characterization of SGP is replicated here by the near perfect correlation of SGP estimates with conditional regression residuals.

The SGP method is the least demanding in its attendant assumptions and data requirements as it only involves ordinal properties of measurement, does not require a score scale that is comparable, equated, or vertically linked across measurement occasions, and only requires one prior data point in addition to the current measurement occasion. Some disadvantages of SGP are their basis in complex modeling methods that are unlikely to be understood by stakeholders, and complex interpretation. Because of the language regularly associated with SGP, it is also likely that results will often be misinterpreted. Because SGP represent a student's relative position within a statistical conditional distribution it is hard to say how much growth has actually occurred (see Figure 4 for example). Students can decline in performance and still receive a high SGP or increase in performance and receive a low SGP depending on prior scores.

Multilevel Growth Model Norms

Unlike SGP methods, MGM provide information on absolute growth and explicitly include times of measurement into estimation of the growth function. This means that MGM norms are meaningfully tied to the temporal course of development, maturation, or matriculation whereas SGP models have no explicit connection to the times at which previous measurements were administered (i.e., the model is moot on whether prior achievement was one month or two years prior). Another advantage of MGM methods is the ability to condition on other factors thought to be relevant to student performance at either the student (e.g., special education status) or school levels (e.g., region). If there is interest in estimating an underlying theoretical distribution of student performance, the MGM method has some additional advantages. The estimation methods used (i.e., maximum likelihood estimation, empirical Bayes estimation) are highly efficient. The MGM method also allows students to have missing data on some measurement occasions and through the method used to represent time in the model, student assessments can occur at different times. Another advantage of the MGM approach is that unit "weakness" is adjusted through EB shrinkage. That is, estimation of a student's growth trajectory is shrunk towards the grand mean growth trajectory as an individual's data is sparse or unreliable (Raudenbush & Bryk, 2002).

MGM have a number of potential disadvantages as well. They are based on fairly complex statistical methods that may be difficult for some users or stakeholders to understand. They also require moderately sized samples ($N > 200$) for estimation, and a vertically linked scale over time. Because use of a temporal function is integral to the MGM approach, longitudinally matched samples are necessary and multiple assessment occasions are required.

Summary and Conclusions

Despite a long history of the use of normative methods to aid score interpretation in a variety of educational, psychological, and health sciences applications, norms are rarely used in the interpretation of standards-based achievement testing common in accountability contexts. SGP have enjoyed rapid growth in popularity and application, provide results expressed as percentile ranks, and use interpretive language (i.e., "comparison to the student's peer group") that suggest norm-referenced interpretations of score performance and academic growth. However, Castellano & Ho (2013) argue that SGP do not directly reflect "achievement score gains" but change in a student's relative position in the current score distribution compared to previous score distributions. We examined two other methods for developing achievement growth norms, traditional norms and multilevel growth model norms. We found that each representation resulted in somewhat different representations of test performance. It is also evident that if the user wishes to explicitly locate a student's performance along a temporal trajectory, TGN or MGM methods are better suited. An additional advantage of MGM methods is the model based fitting and smoothing that occurs when empirical Bayes methods are applied. Use of normative information can provide useful information for score interpretation but further research is needed to guide use and application in the context of modern accountability testing.

References

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th Ed.). Upper Saddle River, NJ: Prentice Hall.

Angoff, (1984).

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

Betebenner, D. W., & Iwaarden, A. V. (2011). *SGP: An R package for the calculation and visualization of student growth percentiles* [Computer software manual]. (R package version 0.4-0.0 available at http://cran.r-project.org/web/packages/SGP/)

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*, 289–328.

Briggs, D., & Betebenner, D. (2009, April). *Is Growth in Student Achievement Scale Dependent?* Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and Quantile Regression Approaches to Student ‘‘Growth’’ Percentiles. *Journal of Educational and Behavioral Statistics,* 38, 190-215

Cole, T. J. (1988). Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society (Series A), 151*, 385-418.

Kuczmarski, R. J., Ogden, C. L., Guo, S. S., et al. (2002). *2000 CDC growth charts for the United States: Methods and development*. Hyattsville, Maryland: Department of Health and Human Services.

Lord, F. M. (1959). Test norms and sampling theory. *Journal of Experimental Education, 27*, 247-263.

Northwest Evaluation Association. (2011). *Technical manual for Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG).* Portland, OR: Author.

Pearson Education. (2008). *Wechsler Memory Scale* (4th Ed.): *Clinical features of the new edition*. San Antonio, TX: Author.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York: Oxford University Press.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations, *Development and Psychopathology, 10*, 395-426.

Wright, E. M. and Royston, P. (1997). A comparison of statistical methods for age-related reference intervals. *J. Roy. Statist. Soc. Ser. A* 160 47-69