Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 19, Number 15, November 2014 ISSN 1531-7714

A Comparison of Three Conditional Growth Percentile Methods: Student Growth Percentiles, Percentile Rank Residuals, and a Matching Method

Adam E. Wyse, *The American Registry of Radiologic Technologists*

Dong Gi Seo, *The National Registry of Emergency Medical Technicians*

This article provides a brief overview and comparison of three conditional growth percentile methods; student growth percentiles, percentile rank residuals, and a nonparametric matching method. These approaches seek to describe student growth in terms of the relative percentile ranking of a student in relationship to students that had the same profile of prior achievement. It is shown that even though the methods come from a similar conceptual foundation, the methods make different assumptions and use different models to estimate growth percentiles. Reading and Mathematics data from a large-scale assessment program are used to compare the growth percentile estimates in a practical setting. Results suggested that the methods often give somewhat similar results. However, the matching method tended to provide somewhat different estimates compared to the other approaches for students that had extreme scores on the prior year test. The implications of these results for large-scale state accountability programs are discussed.

Since the introduction of the No Child Left Behind (NCLB, 2001) Act, there has been increasing attention and scrutiny on the performance of students, teachers, and schools in the United States. NCLB formalized requirements that every student be proficient by 2014. From the start critics have pointed out challenges in reaching this goal, including that it is statistically impossible for all students to be proficient (Linn, 2003; 2005; Rogosa, 2005), that there is variation in the proficiency rates across states (Braun & Qian, 2007; Linn, 2003; 2005), that proficiency rates are highly dependent on the location of cut scores (Ho, 2008), and that certain types of schools may disproportionately not make accountability targets (Choi, Seltzer, Herman, & Yamashiro, 2007; Sims, 2013). These challenges as well as other factors have resulted in changes to federal policy to allow for the use of growth models in state accountability systems. This has included the process for receiving waivers from NCLB, which introduced the requirement of using student growth as a part of accountability systems (USDOE, 2011b).

One of the most popular approaches that states have employed as a mechanism for measuring student growth has been student growth percentiles (Betebenner, 2008; 2009; 2011) or an alternative known as percentile rank residuals (Castellano & Ho, 2013b). These methods fall under the class of methods known as conditional growth percentile models (Castellano & Ho, 2013a), which attempt to characterize student growth in terms of the relative percentile ranking of a student in relationship to students that had the same profile of prior achievement. In many accountability

systems, the profile of achievement is simply the performance of the student in the previous year and the goal of the modeling is to estimate the percentile rank of the student conditional of how the student performed last year. In some cases, more than two years of prior test scores might be used, but many accountability systems base their estimates on only two years of data (Goldschmidt et al., 2005; USDOE, 2011a).

The purpose of this paper is to provide a brief overview and comparison of student growth percentiles, percentile rank residuals, and an alternative nonparametric approach that simply matches students based on their prior year score and then estimates percentile ranks within these matched distributions. Throughout the discussion and comparisons, key similarities and differences are highlighted between the approaches. Data from a large-scale state testing program are then used to compare the estimates of the conditional growth percentile ranks with each of the three approaches. Results show that in many cases student growth percentiles, percentile rank residuals, and the nonparametric matching approach give fairly similar results. Some notable differences were found in how the methods assigned growth percentiles when students had prior year scores toward the extremes of the prior year score distribution. The article concludes with additional discussion and some recommendations for implementing these methods in the future.

Conditional Growth Percentile Methods

There are several different methods and approaches for measuring student growth in state accountability systems, including methods based on gain scores, trajectories, categorical performance level transitions, residual gains, projections, conditional growth percentiles, and multivariate models (Castellano & Ho, 2013a). Methods based on conditional growth percentiles are among the most popular approaches. These methods attempt to describe the relative location of a student's current score in comparison to students that had the same profile of prior achievement in the metric of percentile ranks. Given that this is the goal of the methods, conditional growth percentile methods are typically explained as providing a measure of growth that indicates how much progress a student has made relative to their academic peers that had the same score on previous assessments. For example, a common way of explaining how these methods work is to talk about a small group of students in a classroom that all had the same score on a grade 3 reading test (e.g., a score of 330) and then show how these students did on the grade 4 reading test relative to other students with the same prior score across the state. The top scoring student on the grade 4 reading test is shown to have the highest growth percentile in the class, the bottom scoring student is shown to have the lowest growth percentile in the class, and middle scoring students are shown to have moderate growth percentiles. Examples of these types of explanations can be found on several state websites, including the websites for the state of Virginia (http://www.doe.virginia.gov/testing/scoring/student_growth_percentiles/), the state of New Jersey (http://www.state.nj.us/education/AchieveNJ/teacher/percentile.shtml), and the state of Washington (https://www.k12.wa.us/assessment/studentgrowth.aspx).

Based on this description, one would think that conditional growth percentile models work by first locating students with the same prior year score and then finding the percentile rank of students within this distribution across the state. This is not exactly how student growth percentiles and percentile ranks residuals work to estimate conditional growth percentiles. In the next few sections, we describe in more detail how student growth percentiles, percentile rank residuals, and a nonparametric matching approach that is aligned with the above description work to estimate conditional growth percentiles.

Student Growth Percentiles

Student growth percentiles were introduced by Betebenner (2008; 2009; 2011) as a normative approach for describing the growth of students. Student growth percentiles can be implemented in the SGP R package and employ sophisticated statistical modeling approaches that involve smoothing the distribution of prior year scores using B-spline functions. One hundred quantile regression lines are then estimated in the intervals from 0.005 to 0.995 in increments of 0.01 to determine the percentile rank of the student conditional on prior achievement. Algebraically, the equations that are estimated can be represented as:

$$Q_{Y_{ti}}\left(\tau \middle| Y_{(t-1)i},\ldots,Y_{1i}\right)=\sum_{k=1}^{t-1}\sum_{j=1}^{3}\phi_{jk}\left(Y_{ki}\right)\beta_{jk}\left(\tau\right) \qquad (1)$$

where Y_{ti} is the observed score on the current assessment at time t for student i, τ is the quantile that one wants to estimate, Y_{ki} is the observed score for prior time k for student i, $\phi_{jk}(Y_{ki})$ is a cubic B-spline basis function of degree j for prior time k, and β_{jk} are the coefficients for the cubic B-splines.

To figure out a student's growth percentile from Equation 1, one inputs the student's prior year test scores into right hand side of each of the 100 quantile regression equations to determine predicted scores. Then, one compares the predicted scores to the student's current year observed score and the midpoint of the two quantiles that the student's score falls between times 100 is the student growth percentile. For example, if the student's current year score fell between the predicted scores of 0.605 and 0.615, then the student growth percentile would be 61.

There are several important observations that one can make based on the above equation and its use to determine student growth percentiles. First, the equation allows for many years of prior year test scores to be included in the estimation of student growth percentiles. Second, the models are non-linear and do not make the same assumptions as are made with linear regression models. This includes allowing heteroscedasticity in the current year test score distributions for different combinations of prior year scores. Third, while it is possible to estimate student growth percentiles for every percentile from 1 to 99 for a given profile of prior year scores, all of these student growth percentiles may not be observed in a particular application. For example, there may only be five students with the same profile of prior year scores and in this case a maximum of five distinct student growth percentiles would be observed if all of the students had different current year scores. These growth percentiles are determined by comparing current year observed scores with predicted scores and involve the use of the estimated equations based on all available data. This implies that the typical explanation of how student growth percentiles work does not completely align with how the method is explained to non-technical audiences because it is the substitution of one's prior year scores into the estimated equations based on all available data and comparing these values to one's current year observed score that determines the student growth percentile. Fourth, it is apparent that Equation 1 is not impacted by covariates. That is, there are not separate predicted scores for males and females that had the same prior year scores. Finally, the process to estimate student growth percentiles can be time consuming since they involve estimating 100 equations and the use of cubic B-splines.

The student growth percentile approach is a well-developed method that has been implemented in several different states. The approach is widely accessible to researchers and practitioners through the SGP R package. This software package offers functions to estimate student growth percentiles as well as tools for producing simple PDF score reports and visualizations of data. The state of Colorado has also done extensive work on communicating SGP results to different stakeholders and has developed several web-based applications that utilize student growth percentile data. This work has been published in many articles and technical reports (see Betebenner 2008; 2009; 2011).

Percentile Rank Residuals

Percentile rank residuals were introduced as an alternative method to student growth percentiles for estimating conditional growth percentiles by Castellano and Ho (2013b). In contrast to student growth percentiles, percentile rank residuals involve estimating a linear regression model. The linear regression model that is estimated as part of computing percentile rank residuals is:

$$F\left(Y_{ti}\middle|Y_{(t-1)i},\ldots,Y_{1i}\right)=\beta_o+\beta_1Y_{1i}+\ldots+\beta_{t-1}Y_{t-1i}+\varepsilon_i \qquad (2)$$

where Y_{ti} is the observed score on the assessment at time t for student i, Y_{1i} is the observed score at prior time 1 for student i, $Y_{(t-1)i}$ is the observed score for prior time t -1 for student i, the β s are the regression coefficients, and ε_i is a residual error. After estimating Equation 2, one puts in the prior year scores to estimate predicted scores and one finds the residuals as:

$$Y_{ti}-\hat{Y}_{ti} \qquad (3)$$

Then, based on the residuals from Equation 3 one finds the number of residuals that are less than or equal to the student's residual divided by the total number of examinees, n, times 100. This is the percentile rank

residual for the student, which can be represented algebraically as:

$$\frac{\#residuals \leq Y_{ti} - \hat{Y}_{ti}}{n} \times 100 \quad (4)$$

It is also possible to use the regression model to simply create predicted scores and a student's observed score could be compared to the distribution of predicted scores or function of predicted scores. This approach is not commonly employed to determine percentile rank residuals and would probably lead to somewhat different results than those obtained from using the residuals computed in Equation 4.

Similar to student growth percentiles, there are a few things that one can infer from looking at how percentile rank residuals are estimated. First, similar to student growth percentiles it is possible to use multiple years of prior test scores when estimating the regression equation. Second, since the model is a linear regression model it involves the typical linear regression assumptions. These assumptions include that the relationship between the dependent and independent variables is linear, that the variables are measured without error, and that the errors are independent, normally distributed, and homoscedastic (Osborne & Waters, 2002). Third, similar to student growth percentiles the typical explanation of how the method works is not how the statistical estimation process works to determine percentile ranks. In this case, it is the comparison of one's residual to the full distribution of residuals across all examinees that determines the conditional growth percentile. Also, similar to student growth percentiles every percentile from 1 to 99 for a given profile of prior year scores may not be observed in a particular application. Fourth, it is apparent that Equation 2 is not impacted by any covariates and it is only the prior year scores that enter into the computation of the predicted values. It would be easy to extend Equation 2 and include covariates if one desired to do so. Finally, it is considerably easier to estimate percentile rank residuals than it is to estimate student growth percentiles since many commercially available software packages can be used to estimate linear regression models and compute corresponding residuals and percentile ranks.

Castellano and Ho (2013b) use simulation methods to compare student growth percentiles and percentile rank residuals using known statistical distributions. They concluded that percentile rank residuals and student growth percentiles both worked pretty well in terms of recovering conditional growth percentiles. They suggested that percentile rank residuals tended to work better with smaller sample sizes and that in practical settings one should investigate model fit to decide between the approaches. They observed in passing that the two methods tended to handle extreme scores somewhat differently. Since the method comes from a common conceptual foundation as student growth percentiles, many of the tools that have been developed for student growth percentiles, such as those in the SGP package, can be utilized with percentile rank residuals to communicate results to parents, teachers, and students.

Nonparametric Matching

There is a third approach for estimating conditional growth percentiles that is usually dismissed by researchers that have written about conditional growth percentile methods because it is hard to implement with many years of prior test scores (Castellano & Ho, 2013a; 2013b). This approach is to simply match students based on their prior year test scores and then find the percentile rank for students within these matched distributions. For example, Castellano and Ho (2013a) state "a strict implementation of this procedure would seem to involve the selection of "academic peers" that have identical previous scores. This is impractical and imprecise with large numbers of prior grade scores (p. 89)." In many accountability systems, however, the growth models that are used only involve looking at the previous year scores. This is often due in part to the fact that as one uses more years of data there are fewer students that have prior year scores for every prior year. The use of more years of data typically decreases the number of students that will have growth scores.

The approach mentioned by Castellano and Ho (2013a; 2013b) that involves matching students based on their prior year test scores is nonparametric and does not involve the estimation of any regression models. The first step of the procedure is to simply find and match students based on their prior year scores. Then, within each of these distributions one calculates percentile ranks. The percentile rank assigned to the student within their matched distribution is simply the number of current year observed scores that are less than or equal the student's current year observed score divided by the total number of

examinees, n, times 100. This can be represented algebraically as:

$$\frac{\#Y_t \leq Y_{ti}}{n} \times 100 \tag{5}$$

There are several things to note about this procedure. First, the basic procedure involves only matching on one prior year test score. Including more than a single year of data makes the procedure harder to implement because as one includes more prior years of data there are fewer students that have the same combination of prior year scores. This can result in the number of distinct observed percentiles for a given set of prior year scores becoming small. In the most extreme case, there would be no other students with the same prior achievement profile and all students would be assigned a percentile rank of 99. This can impact the precision of the percentiles in Equation 5 and is a drawback to the approach, which is noted by Castellano and Ho (2013a; 2013b). Second, since the procedure uses matching and is nonparametric, there are very few required assumptions and different distributions of percentile ranks are possible with different combinations of prior year scores. This can be both a strength and weakness. On the positive side, one is not relying on statistical assumptions or a commonly assumed distribution, which may not hold across all combinations of prior year scores, to estimate growth percentiles. On the negative side, the flexibility of the matching method can lead to situations where the percentile ranks assigned may not necessarily be monotone for different combinations of scores unless one uses a smoothing method. The possibility for results to not be monotone without smoothing may be confusing for some stakeholders to understand when they try to interpret their results. The cubic B-spline functions in the SGP package are used to remove the possibility of results not being monotone with the student growth percentile method. Third, the procedure is fully aligned with the description of how conditional growth percentile models are typically explained to non-technical audiences since it involves matching based on prior year scores. Fourth, similar to student growth percentiles and percentile rank residuals there are no covariates besides test scores that are used in estimating growth percentiles. Finally, similar to percentile rank residuals, this method is easy to implement in commercial software packages and is computationally easier to compute than student growth percentiles

To date this nonparametric approach, although it has been mentioned in passing in the research literature, has not been empirically compared to student growth percentiles or percentile rank residuals because it has been viewed as impractical by some researchers. Since the method comes from a common conceptual foundation as percentile rank residuals and student growth percentiles and attempts to report growth in the metric of percentiles, the SGP R package and tools that have been developed to report and visualize results in these other contexts can also be used with the nonparametric matching approach.

Data and Methods

To investigate the similarity of the three approaches for computing conditional growth percentiles, data from a large-scale state testing program that has been exploring the use of a conditional growth percentile type method as a mechanism for reporting student growth to schools were utilized. The state typically has reported growth by looking at performance level changes across years for students in grades 4 through 8 (see Martineau, 2007; Wyse, Zeng, & Martineau, 2011) and this growth model has been approved by the federal government as part of the United States Department of Education Growth Model Pilot Program (2011a). In this study, Mathematics and Reading assessment data for students in grades 4 through 8 for a single two-year period were used. Only students who had valid scores in their current and their prior year grades were used in the analyses for each subject.

Table 1 presents descriptive statistics of the scale scores for each cohort for both Reading and Mathematics. The Mathematics and Reading tests are not vertically scaled and separate scales are created for each grade and subject combination. The sample sizes of students were higher for Reading compared to Mathematics, but in both cases the sample sizes of students in each cohort were quite large. One can also see that the difference in mean scales between consecutive grades was roughly 100 scale score points. Most of the standard deviations of the scale scores were in the low 20s to the low 30s. The correlations between the prior and current year scores tended to be higher for the Mathematics data compared to the Reading data.

The three conditional growth percentile methods were estimated in R. Student growth percentiles were estimated using the SGP R package with the default settings. Percentile rank residuals and the nonparametric matching method were estimated using custom written programs designed for these purposes.

Table 1: Descriptive Statistics of Scale Scores for Mathematics and Reading Data

Subject	Cohort	N	Previous Mean	Current Mean	Previous SD	Current SD	Correlation
Mathematics	3~4	107,844	329.367	429.415	19.295	22.165	0.788
	4~5	105,019	429.465	527.306	21.765	30.498	0.808
	5~6	101,277	526.922	624.136	29.891	24.232	0.803
	6~7	98,131	623.65	725.771	23.746	26.073	0.807
	7~8	94,044	725.575	819.408	25.724	25.458	0.819
Reading	3~4	108,468	332.721	433.612	26.682	29.100	0.706
	4~5	103,896	431.564	533.339	28.630	27.807	0.711
	5~6	108,557	532.136	630.957	27.985	27.613	0.733
	6~7	105,268	628.888	728.011	27.320	30.983	0.742
	7~8	100,729	726.012	824.898	30.692	23.913	0.722

After estimating each of the three methods, the methods were compared using four separate criteria.

The first criterion was the root mean square difference (RMSD). The RMSD measures the average squared differences between two different methods. The RMSD can be represented algebraically as:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right)^2}{n}} \tag{6}$$

where x_{1i} is the conditional growth percentile for method 1 for observation i, x_{2i} is the conditional growth percentile for method 2 for observation i, and n is the total number of examinees. When the RMSD is closer to zero, it signals that the two methods are more similar to each other.

The second criterion was the mean difference (MD). The MD measures the average differences between two different methods. The MD can be represented algebraically as:

$$MD = \frac{\sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right)}{n} \tag{7}$$

and the terms in the equation have the same meaning as before. When the MD is positive it indicates that the growth percentiles assigned by method 1 are higher on average than method 2, and when the MD is negative it indicates that the growth percentiles assigned by method 1 are lower on average than method 2. When the MD is closer to zero in absolute value, it signals that the two methods are more similar to each other.

The third criterion was the classification consistency of the two methods. This was determined by finding the percent agreement for the two methods in terms of classifying individuals as exhibiting low, typical, or high growth. Low growth was defined as a growth percentile less than the 35th percentile, typical growth was defined as growth percentile between the 35th and 65th percentile, and high growth was defined as growth percentile that was greater than the 65th percentile. The definitions for low, typical, and high growth were based on the growth categories utilized by the state of Colorado in their applications of student growth percentiles (Betebenner, 2011). Methods that have higher classification consistency are more similar to each other in terms of classifying examinees into the different growth categories.

The final criterion was the correlation between methods, which was computed as the Pearson product moment correlation. High correlations signal that there is a higher degree of linear association between the methods and that the methods are more similar to each other. It is expected that most of the correlations between the methods should be fairly high since all of

the methods can be classified as conditional growth percentile methods.

Results

Table 2 shows the pairwise comparisons of the three different conditional growth percentile methods across the four criteria. Several patterns are apparent in the results. First, the three methods exhibited greater similarity across the four criteria for the Mathematics data compared to the Reading data. This corresponded to the situation where the correlations between the observed scores in consecutive grades were higher in Table 1. Second, the methods that were most similar to each other differed depending on the criterion.

For the correlations, the student growth percentile method and matching method had the highest correlation across all cohorts for both grades. The pairwise correlations between the matching method and the percentile rank residual method and between the student growth percentile and the percentile rank residual method were less, but still tended to be high. For these two sets of pairwise comparisons, there were some cases where the correlations were nearly identical between the pairs, while in other cases one pair of methods had a higher correlation than the other pair. Across all pairwise correlations, the correlations ranged from a low of 0.972 to a high of 0.998, indicating a very high degree of association between the three methods.

For the RMSD statistic, the student growth percentile and matching method had the lowest value and were the most similar except for the Reading grade 6 to 7 cohort and the Reading grade 7 to 8 cohort. In these cases, the student growth percentile and percentile rank residual methods were the most similar methods. The matching method and the percentile rank residual method were the least similar pair of methods for both Reading and Mathematics.

The results for the MD statistics were somewhat different than results for the correlations and RMSD. For this criterion, the student growth percentile and percentile rank residual method were the most similar methods. For the Mathematics data, the differences between these two methods were relatively small and positive. For the Reading data, the differences were a bit higher and negative. The matching method tended to produce higher growth percentiles on average than the other two methods. For the Mathematics data, the differences between the matching method and the other two methods were on average two to three percentiles higher. For the Reading data, the differences were about approximately 3.5 to 4.5 percentiles higher. These magnitudes of differences are noteworthy and suggest that there may be some important differences in the average percentiles assigned when using the matching method compared to the other two methods.

In terms of the classification consistency, the student growth percentile and percentile rank residual methods tended to yield the most similar classification into the low, typical, and high growth categories except for the grade 6 to 7 and grade 7 to 8 Mathematics cohort where the student growth percentile method and the matching method had the highest classification

Table 2: Comparison of the Three Conditional Growth Percentile Methods for Mathematics and Reading

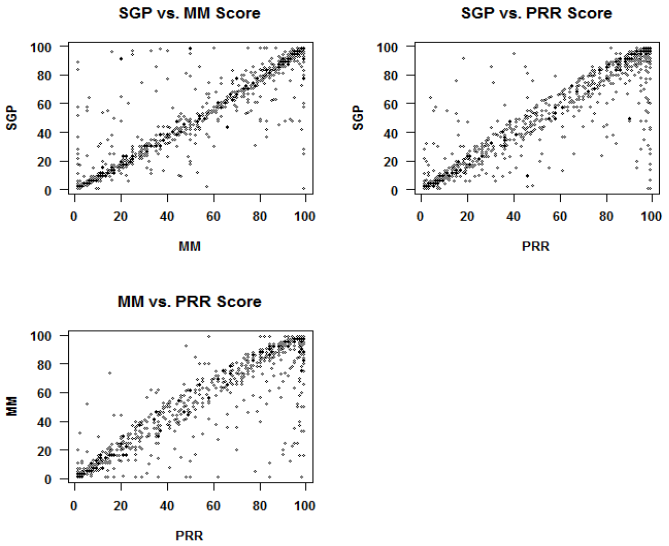
		SGP vs. MM				SGP vs. PRR				MM vs. PRR			
Subject	Cohort	r	RMSD	MD	Consis tency	r	RMSD	MD	Consis tency	r	RMSD	MD	Consis tency
Mathe-matics	3~4	0.997	3.046	-2.157	0.927	0.989	4.354	0.077	0.933	0.988	4.962	2.234	0.906
	4~5	0.997	3.189	-2.321	0.926	0.987	4.611	0.129	0.943	0.988	5.180	2.450	0.909
	5~6	0.998	2.958	-2.241	0.934	0.988	4.537	0.083	0.939	0.988	5.026	2.324	0.908
	6~7	0.998	2.784	-1.934	0.943	0.984	5.107	0.171	0.911	0.984	5.561	2.104	0.908
	7~8	0.997	3.515	-2.856	0.910	0.982	5.411	0.073	0.906	0.982	6.166	2.928	0.877
Reading	3~4	0.992	5.788	-4.515	0.881	0.979	6.014	-0.893	0.894	0.973	7.681	3.622	0.896
	4~5	0.993	5.571	-4.435	0.881	0.979	5.941	-0.721	0.887	0.972	7.837	3.715	0.913
	5~6	0.994	5.238	-4.245	0.879	0.978	6.117	-0.653	0.881	0.974	7.504	3.592	0.873
	6~7	0.993	5.730	-4.618	0.869	0.987	4.754	-0.727	0.915	0.983	6.681	3.891	0.904
	7~8	0.993	5.584	-4.458	0.860	0.984	5.252	-0.750	0.883	0.976	7.323	3.708	0.903

Note: SGP is the student growth percentile method, MM is the nonparametric matching method, and PRR is the percentile rank residual method.

consistency and the Reading grade 3 to 4, grade 4 to 5, and 7 to 8 cohorts where the matching method and the percentile rank residual method had the highest classification consistency. The classification consistency across all grades, cohorts, and pairwise comparisons ranged from 0.860 to 0.943, which indicates that there was a fairly high amount of similarity in classification consistency.

Figure 1 shows scatter plots of the grade 3 to 4 Reading data to help further illustrate some of the similarities and differences that existed among the three different methods. In each of the three panels, one can see that most of the points clustered around the 45 degree line running from the bottom to the top corner of the plots. This explains the high correlations reported in Table 2. Examining the panels more closely shows that the greatest differences between methods appeared to be in the assignment of some of the extreme percentiles. For example, when the percentile rank residual method had a growth percentile close to the 99th percentile it was possible to observe a growth percentile from near the 1st to 99th percentile for both the student growth percentile and the matching method. When the matching method was close the 99th percentile the range of scores for the percentile rank method was smaller and there were no percentiles below 50 assigned and for the student growth percentile method there was only a small number of percentiles less than 50 assigned. One can also see a range of percentiles when the matching method was close to the 1st percentile for the both the student growth percentile and percentile rank residual methods. These patterns held across all of the cohorts in both subjects.

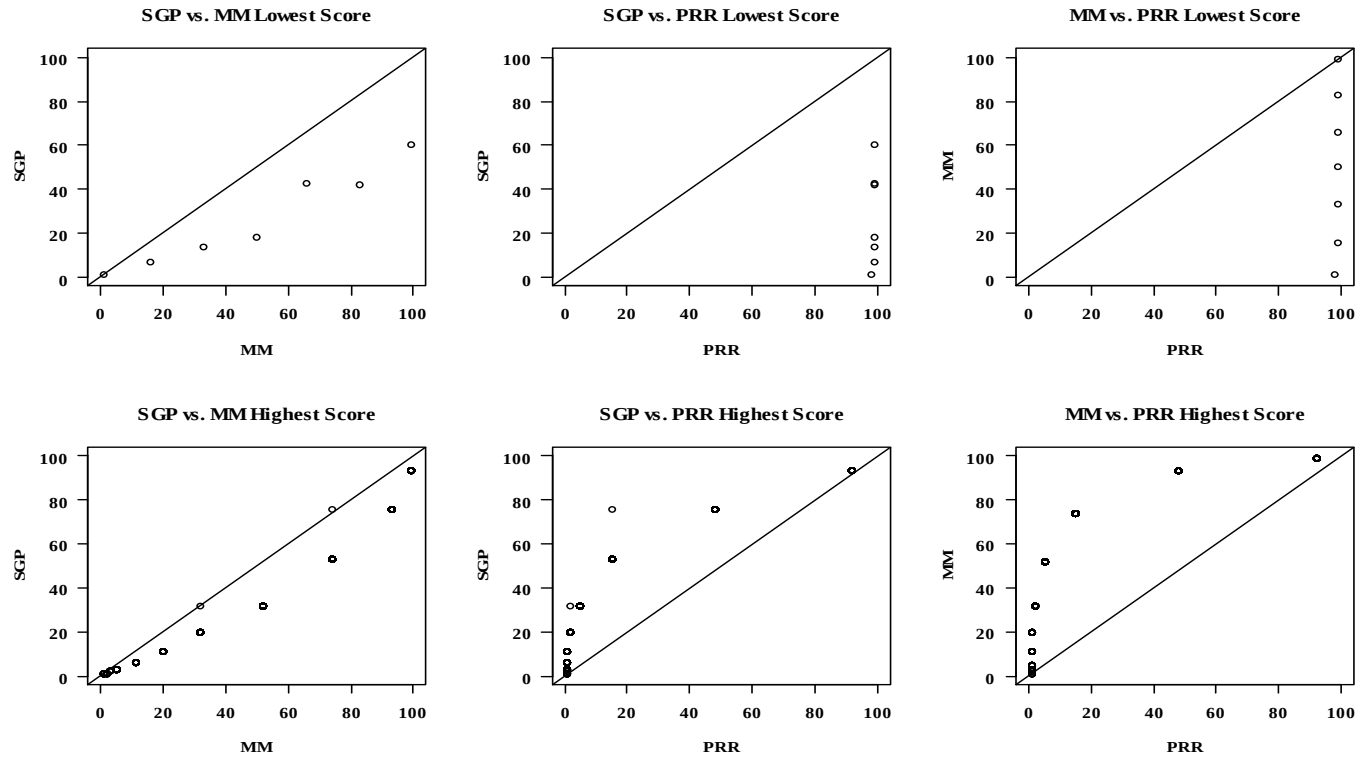
Figure 1: Scatter Plots for Different Methods for Grade 3 to 4 Reading Cohort



Note: SGP is the student growth percentile method, MM is the nonparametric matching method, and PRR is the percentile rank residual method.

To greater understand some of the differences between the methods we examined the assignment of growth percentiles across a range of prior year scores. The most notable differences were found in how the different methods assigned growth percentiles for examinees that achieved the highest and lowest score points on the grade 3 Reading test. Figure 2 shows a scatter plot of these relationships. The top row shows comparisons between methods for the lowest score point on the grade 3 test and the bottom row shows comparison between the methods for the highest score point on the grade 3 test. Some patterns in how the methods assigned growth percentiles to extreme score points are apparent from Figure 2. Namely, the percentile rank residual method assigned people who scored at the lowest scores on the grade 3 test a growth percentile in the upper 90s. This was in contrast to the other two methods, which assigned some lower growth percentiles. One can also see that the matching method had growth percentiles ranging from close to 1 to 99, while the student growth percentile method had growth percentiles less than 60 and lower than the percentiles assigned by the other methods. For students who obtained the highest grade 3 test score, the percentile rank residual method produced much lower growth percentiles than the other two methods with most of the percentiles being less than 60. The matching method yielded growth percentiles that were higher than the other two methods, which covered the full range of possible percentiles. These results were not isolated to the grade 3 to 4 Reading cohort and other grades and content areas had similar results to those found with the grade 3 to 4 Reading cohort.

Figure 2: Method Comparison Plot of Highest and Lowest Two Scores for Grade 3 to Grade 4 Reading Cohort



Note: SGP is the student growth percentile method, MM is the nonparametric matching method, and PRR is the percentile rank residual method.

These patterns may be a function of regression to mean that can occur for students with extreme prior year test scores when they take the test at the next higher grade. The patterns may also be a function of the fact that both the percentile rank residual method and student growth percentile method use all of the available data to estimate equations that are assumed to hold across the range of observed scores. The distributions of current year scores for students that scored at the extremes of the prior year test score distribution are often quite different than distributions of current year scores at other score points. This implies that the fit of the models and estimates of the predicted scores at the extreme score points do not appear to be particularly good with either method for these data. Since many of the tests on which these analyses are based exhibited a ceiling effect with more students achieving some of the higher score points on the test, this provides an explanation as to why the matching method tended to yield average higher growth percentiles when looking at the mean differences in Table 2. These are the situations where the matching method makes better use of data observed at each prior year score and assigned a greater range and higher growth percentiles than the other methods.

Discussion and Conclusion

The purpose of this paper was to provide a brief overview and comparison of student growth percentiles, percentile rank residuals, and a nonparametric matching method. Results suggested that the methods often yielded somewhat similar results, but there were some notable differences in the estimation and assignment of growth percentiles for examinees that had scores at the extremes of the prior year test score distribution. In these cases, the matching method tended to yield a greater range of growth percentiles and assigned higher growth percentiles than the other two methods because the matching method estimates growth percentiles conditional on the prior year score and does not assume that equations based on all available data hold for extreme scores. This is an

implicit assumption of both percentile rank residuals and student growth percentiles.

Many of the extreme score points are where the assumptions for student growth percentiles and percentile rank residuals are least likely to hold. In these cases, it appears that student growth percentiles and percentile rank residuals, in particular, were not very accurate in estimating conditional growth percentiles. These potential challenges with percentile rank residuals and student growth percentiles are worthy of additional investigation in other contexts. It is possible that some of the results observed in the data sets investigated in this study may hold in other contexts as well. Although, it was not the explicit focus on the study by Castellano and Ho (2013b), some of their results also suggested that there may be some notable differences in the way that percentile rank residuals and student growth percentiles assign percentile ranks for students with extreme prior year scores.

There are two conclusions that one may draw based on the results presented in this study. The first is that the methods often give somewhat similar results in practical circumstances and that unless there is a concern about extreme scores the methods will probably yield fairly similar results. The second conclusion is that the matching method appeared to work as well, if not better than the student growth percentiles and percentile rank residuals, for the data used in this study. Of course, one of the challenges when using real data is that there is not a true baseline to compare the different methods against to definitively conclude which method works better. It is important to point out that even though the examples used in this study came from a large-scale state assessment program that utilizes student growth models and were designed to be representative of typical situations in which the three methods would be implemented, data and results may differ in other situations. For example, other data sets may not have the same sample sizes as were used in this study and may have different score distributions and patterns of examinee performance. These and other factors may change results in other practical settings.

Using the matching method does have some additional practical challenges. This includes that it is hard to implement this method when there are multiple prior years of data since the number of examinees in the matched distributions may be small. When there are a small number of matched individuals in the matched distributions, the precision of the percentile ranks decreases. Although it is hard to give a definitive recommendation of what sample size of examinees would lead to enough precision to be able to effectively implement the method for a particular application, having 10 to 20 examinees or more at each prior year score combination seems like a reasonable minimum. The total sample size would then be dependent on the number of prior year score combinations. Of the three methods, the matching method probably is the most restrictive in terms of the required sample size because enough matched cases are needed at each prior year score to obtain precise estimates. Another important factor to consider is the possibility for some of the results to not be monotone unless one uses a smoothing method. The possibility for non-monotone results may make it hard for stakeholders to interpret some of the results. These are important things to consider in the application of this method and they warrant additional investigation in other circumstances. In addition, since the matching method does not estimate a statistical model, it does not produce estimated equations that may be used for other accountability purposes, such as projecting how a student may be expected to perform at some future point in time. The matching method does have the advantage that how the method is implemented aligns with how non-technical audiences think that conditional growth percentile methods work.

One can contrast these considerations with those for the student growth percentile method and percentile rank residual method. Student growth percentiles and percentile rank residuals make stronger assumptions than the matching method. In the case that there is a small amount of data for a prior year score point or prior year score profile, percentile rank residuals and student growth percentiles assume that the relationships estimated across the full distribution of prior year scores hold in this case as well. This assumption may or may not hold in practice and future research could focus on testing the assumptions of these methods with different distributions of scores. The percentile rank residual and student growth percentile methods also have different sample size requirements than the matching method. Castellano and Ho (2013b) found that to get similar precision as the percentile rank residual method with a sample size of 1,000 examinees that the student growth percentile method required a sample size of approximately 5,000

examples. They suggested that the percentile rank residual method was more robust to smaller sample sizes. Based on their research, the percentile rank residuals would appear to require the lowest sample size with the student growth percentile and matching method requiring more data. Additional research is needed to examine how the methods work with different sample sizes for a variety of different types of data. Since the percentile rank residuals and student growth percentiles involve the application of statistical models to estimate equations, estimated equations are typically available for other accountability purposes, such as computing score projections. These score projections are sometimes used to give schools credit for students being on track to reach proficiency at some time point in the future.

Before one uses any of the three conditional growth percentile methods, it is important for researchers and practitioners to consider and examine how the method will work in their context. This can include comparing the results of the different methods against each other using simulated or empirical data to evaluate the performance of the methods across a variety of different criteria. The examples shown in this article illustrate a few example criteria that one might want to consider. Other criteria are possible, including looking at school level criteria or criteria at other levels of aggregation. The use of different criteria and data may highlight other important similarities and differences between the methods that were not captured in this study. Ultimately, the decision of what method to use should be based on a good understanding of the strengths and weaknesses of the approaches as well as evidence of whether the method will function appropriately for the desired purpose in the specific context.

References

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shephard (Eds.), *The future of test-based educational accountability* (pp. 155-170). New York, NY: Taylor-Francis.

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

Betebenner, D. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth trajectories/projections.* The National Center for the Improvement of Educational Assessment. Retrieved March 20, 2014 from http://www.nj.gov.edu/education/njsmart/performance/SGP_Technical_Overview.pdf.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313-338). New York: Springer.

Castellano, K. E., & Ho, A. D. (2013a). *A Practitioner's Guide to Growth Models*. Washington, DC: CCSSO. Retrieved March 20, 2014 from http://scholar.harvard.edu/files/andrewho/files/a_pracitioners_guide_to_growth_models.pdf.

Castellano, K. E., & Ho, A. D. (2013b). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*, *38*, 190-215.

Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2007). Children left behind in AYP and non-AYP schools: Using student progress and the distribution of student gains to validate AYP. *Educational Measurement: Issues and Practice*, *26(3)*, 21-32.

Goldschmidt, P., Roschewski, P., Choi, K. C., Auty, W., Hebbler, S., & Williams, A. (2005). *Policymakers' guide to growth models for school accountability: How do accountability models differ?* Washington, DC: CCSSO. Retrieved March 20, 2014 from http://www.ccssso.org/publications/details/cfm?PublicationID=287.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*, 351-360.

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Educational Policy Analysis Archives, 11(31)*. Retrieved March 20, 2014 from http://epaa.asu.edu/epaa/v11n31.

Linn, R. L. (2005, June). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives*, *13(33)*. Retrieved March 20, 2014 from http://epaa.asu.edu/epaa/v13n33/.

No Child Left Behind Act of 2001, 20 U.S.C. 6311 *et seq*.

Martineau, J. A. (2007, March). *Designing a valid and transparent progress-based value-added accountability model*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), Chicago, IL.

Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, *8(2)*. Retrieved March 20, 2014 from http://PAREonline.net/getvn.asp?v=8&n=2.

Rogosa, D. R. (2005). Statistical misunderstandings of the properties of school scores and school accountability. In J. L. Herman & E. H. Haertel (Eds.), *Yearbook of the National Society for the Study of Education* (pp.147-174). Chicago: National Society for the Study of Education.

Sims, D. P. (2013). Can failure succeed? Using racial subgroup rules to analyze the effect of school accountability failure on student performance. *Economics of Education Review*, *32*, 262-274.

U.S. Department of Education (2011a). *Final report on the evaluation of the growth model pilot project*, Washington, D.C: Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service. Retrieved March 20, 2014 from http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/gmpp-final.pdf.

U.S. Department of Education. (2011b). *Letter from the Education Secretary, September 23, 2011*. Retrieved March 20, 2014 from http://www2.ed.gov/policy/gen/guid/secletter/110923.html.

Wyse, A. E., Zeng, J., & Martineau, J. A. (2011). A graphical transition table for communicating status and growth. *Practical Assessment, Research and Evaluation*, *16(11)*. Retrieved March 20, 2014 from http://pareonline.net/getvn.asp?v=16&n=11

.

Note:

A portion of this work was completed while the authors worked at the Michigan Department of Education. The conclusions, discussion, and views contained in this article are not necessarily the official position of The American Registry of Radiologic Technologists, The National Registry of Emergency Medical Technicians, or the Michigan Department of Education. The authors would like to thank Ji Zeng for comments she provided that helped to improve this manuscript.

Citation:

Wyse, Adam E. & Seo, Dong Gi (2014). A Comparison of Three Conditional Growth Percentile Methods: Student Growth Percentiles, Percentile Rank Residuals, and a Matching Method. *Practical Assessment, Research & Evaluation*, 19(15). Available online: http://pareonline.net/getvn.asp?v=19&n=15

Corresponding Author:

Adam E. Wyse
The American Registry of Radiologic Technologists
1255 Northland Dr.
St. Paul, MN 55120
Email: adam.wyse [at] arrt.org