

Different Growth Measures on Different Vertical Scales

Dongmei Li

Abstract Vertical scales have been used by testing programs for decades to facilitate the tracking of student performance over time. With the recent emphasis on the measuring of student growth for accountability purposes, scores from vertically scaled tests have been used to evaluate school or teacher performance. Because there are different types of growth measures and there are also different ways to construct a vertical scale, it is important to understand the impact of the vertical scales on various growth measures for important educational decisions. Based on some mathematical relationships that have been shown to exist among certain growth measures and with the use of empirical data, this study investigated the impact of different vertical scales on the relationships among simple gain scores, residual gain scores, and three growth measures based on conditional status percentile ranks (CSPR). Results showed that the correlations between simple gain scores and the rest of the growth measures were affected by the extent of scale expansion or scale shrinkage across grades.

Keywords Growth measures • Vertical scales • Gain scores • Student growth percentiles

Vertical scales are constructed by linking scores of tests from different grade levels through various data collection designs and statistical methods to a common scale. It has been well-acknowledged in the literature that different choices in the vertical scaling process may result in different scales (Li, 2015a; Paek, Young & Yi 2008; Patz 2007; Tong & Kolen 2007). With the increasing high-stakes use of growth measures based on vertically scaled scores in state accountability systems, the effect of different potential vertical scales on growth measures has been a topic of continuous investigation. Research has shown that the rank ordering of students or schools may change not only with the change of growth measures (Dunn & Allen 2009; Goldschmidt, Choi & Beaudon 2012; Li & Kolen 2008), but also with the change of the underlying vertical scales (Briggs & Domingue 2013;

D. Li (✉)
ACT, Inc., 500 ACT Drive, Iowa City, IA 52243, USA
e-mail: dongmei.li@act.org

Briggs & Weeks 2009; Lei & Zhao 2012; Li & Kolen 2011). Based on some mathematical relationships that have been shown to exist among certain growth measures, this study sought to investigate, in some greater depth, the effect of different vertical scales on the relationships among a few growth measures. This study also investigated whether certain types of vertical scales tended to yield more consistent results across growth measures than other types of vertical scales. Specifically, this study investigated how certain features of the vertical scales might impact the relationships among three types of growth measures: simple gain scores, residual gain scores, and CSPR.

In the sections below, an introduction is first given on the potential differences among vertical scales. Then, a few growth measures and their known relationships are described. After that, theoretical and empirical investigation results on the impact of vertical scales on the relationships among these measures are described, followed by conclusions and discussion.

1 Properties of Vertical Scales

Strictly speaking, there has been no formal classification of vertical scales into different types in the literature. Vertical scales are often described or compared in terms of grade to grade means, grade to grade score variabilities, or the effect sizes of between-grade mean changes.

Among these features, the differences in grade to grade score variability have probably drawn the most attention in the literature (e.g., Burket 1984; Hoover 1984; Phillips & Clarizio 1988). A quantity depicting the change of score variability across grades, that is, the ratio of the standard deviations of scores for the upper and lower grades, has been shown to be predictive of correlations among a few growth measures (Li 2015b; Li & Kolen 2011; Roberts & Burrill 1995). Relevant to the purpose of this paper is the mathematical relationship between simple gain scores and residual gain scores given by Li and Kolen (2011). This relationship is discussed in the next section.

2 A Few Growth Measures and Their Relationships

This study investigated three types of growth measures that are commonly used in the context of measuring student growth between years: simple gain scores, residual gain scores, and CSPR. For the simplicity of discussion, this study focused on two years' scores, referred to as the current year and the the previous year scores.

The simple gain score is the difference between scores earned between two years. The residual gain score is the difference between a student's observed current year score and the student's predicted current year score, usually based

on a linear regression of the current year scores on the previous year scores. CSPR is a term suggested by Castellano and Ho (2013) to describe the use of percentile ranks of students' current status scores conditioning on previous scores as a growth measure. One example of CSPR is the well-known student growth percentiles (SGPs) (Betebenner 2008, 2009) which have been used in many states' accountability systems. SGPs and two other less commonly used CSPR alternatives were considered in this paper: the ordinary least squares percentile ranks of residuals (PRRs) (Castellano & Ho 2013), and the empirical conditional percentile ranks (ECPRs).

Before investigating how certain features of vertical scales may affect the relationships among these growth measures, some of their known relationships from previous research are described below.

2.1 Simple Gain and Residual Gain Scores

Whether simple gain scores or residual gain scores should be used in measuring change has been debated in the literature (e.g. Harris 1963; Maris 1998). Lord's paradox (Lord 1967) described the inconsistency of results when the evaluation was based on simple gain scores or residual gain scores. Li and Kolen (2011) showed that correlations between simple gain scores and residual gain scores are determined by two quantities: the correlation between time 1 and time 2 scores, and the ratio of standard deviations (SDs) between time 2 and time 1 scores. Let $k = \frac{\sigma_{X_2}}{\sigma_{X_1}}$, where X_1 and X_2 are the time 1 and time 2 scores, and let σ_{X_1} and σ_{X_2} represent the SDs of the time 1 and time 2 scores, respectively. Let $\rho_{X_1X_2}$ represent the correlation between time 1 and time 2 scores. The correlation between simple gain and residual gain scores (ρ_{DR} , where R stands for the residual gain score, and D stands for the simple gain score) is determined by the values of k and $\rho_{X_1X_2}$. Specifically,

$$\rho_{DR} = \frac{k\sqrt{1 - \rho_{X_1X_2}^2}}{\sqrt{1 + k^2 - 2k\rho_{X_1X_2}}}, \tag{1}$$

where $R = X_2 - \widehat{X}_2$. The predicted time 2 score \widehat{X}_2 is based on linear regression of X_2 on X_1 . Li and Kolen (2011) pointed out that Eq. (1) could be used to predict the consistency of results when these two measures were used for evaluating school effectiveness, and that it could also be helpful for understanding in what situations Lord's paradox (Lord 1967) tended to be more severe. This paper applied this equation in the context of vertical scaling and showed its merit in predicting the impact of vertical scales on the relationships between simple gain scores and residual gain scores.

2.2 *Three CSPR Measures*

As mentioned earlier, this study considered three approaches for estimating growth using CSPR. One was the SGP statistic that has been implemented in many state accountability systems. The other two (PRR and ECPR) were alternatives that have been mainly investigated in research studies.

SGPs are the percentile ranks of current scores conditioning on prior scores based on quantile regression. PRRs, percentile ranks of residuals based on ordinary least square regression, are the percentile ranks of residual gain scores. In addition to the regression methods, one major difference between SGPs and PRRs is that the SGP is the percentile rank in the conditional distributions, but the PRR is the non conditional percentile rank of all residual scores. In other words, SGPs compare a student's current score with the current scores of those who had the same previous scores, whereas PRRs compare a student with all other students in terms of the difference between their current scores and their own expected scores, with the expected scores defined by the linear regression of current year scores on previous scores. Despite the above conceptual differences, it is expected that these two approaches would give similar results as long as the bivariate distribution of the two year scores satisfies the assumptions of a linear relationship and homoscedasticity. This expectation was confirmed by Castellano and Ho (2013). They compared SGPs and PRRs in great depth and concluded that these two metrics were very similar in practice. Using simulations, they also found that PRRs out performed SGPs under multivariate normal distributions (MVN), but SGPs outperformed PRRs with greater deviation from MVN and with greater non linear transformations of the score scales.

ECPRs are conceptually very close to SGPs, except that the ECPR estimates are not based on conditional percentile ranks from quantile regression, but based on the empirical cumulative frequency distributions of current status scores conditioning on prior scores. Some research used these empirical percentile ranks from large data sets as the criteria for evaluating SGPs (e.g., Grady, Lewis & Gao 2010), but the author of this paper considered ECPRs as an alternative approach to SGPs when sufficiently large data are available. Due to the conceptual closeness between ECPRs and SGPs, it was expected that these two approaches would give similar estimates, especially when sample sizes are large.

2.3 *Scale Dependency of the Growth Measures*

Note that among these growth measures, only simple gain scores directly measure the magnitude of growth between grades and thus require the availability of a vertical scale for their proper use. The other two are based on the conditional distributions and can be used regardless of whether the scores are vertically scaled or not. As confirmed by some studies (Briggs & Domingue 2013; Li 2015b), growth

measures that do not require a vertical scale should be unaffected or at least affected less by changes in vertical scales than growth measures that do require a vertical scale. However, little can be found in the literature regarding how certain features of the vertical scale may affect the relationships among different growth measures. The next section shows how Eq. (1) can be used for this purpose.

3 Vertical Scales and the Relationships Among the Growth Measures

Equation (1) shows that the correlation between simple gain scores and residual gain scores is related to two quantities, the ratio of standard deviations between the upper grade and lower grade (k) and the correlations of scores between the two grades ($\rho_{X_1X_2}$). The quantity k depicts the change of score variability across grades—one important feature of vertical scales. With $k > 1$, the variability of scores increases from the lower grade to the higher grade; with $k = 1$, the variability stays constant; with $k < 1$, the variability decreases, which is often referred to as scale shrinkage. Though the values of the correlations may not be known before longitudinal data are gathered, a few likely values can be used for calculation.

Kolen and Brennan (2014) described several statistical methods that could be used to construct vertical scales, among which the IRT method and the Thurstone (1938) method are the most commonly used. The Thurstone method establishes linkages across grades through linear transformations of z-scores, and the IRT method establishes linkages across grades through linear transformation of the theta scores. Unless there is some other transformation, all variations of scales using the same method (Thurstone or IRT) are linearly related. For those vertical scales that are linearly related, the correlation of scores between grades remains the same no matter which vertical scale is used. In this case, the change of correlations between simple gain and residual gain scores from one vertical scale to another only depends on k .

Figure 1 shows the calculated correlations between simple gain and residual gain scores for a selected range of k and a few between-grade correlation values that are realistic in educational tests based on a review of the literature. This plot shows that the correlations between simple gain and residual gain scores may vary from below .5 (when the scale has a severe shrinkage from lower to upper grade) to well above .9 (when the scale expands in variability to a certain extent). However, the change of correlations between simple gain and residual gain scores also depends on the correlations between scores from the two grades—a greater extent of scale expansion may not necessarily result in a higher correlation between the two growth measures. When score variability stays constant across grades ($k = 1$), the correlations between simple gain and residual gain scores are around .9 or higher for the selected between-grade correlations (.6, .7, .8, and .9).

Equation (1) can be used to predict the correlations of simple gain scores and residual gain scores based on two year data, but what about their relationships with

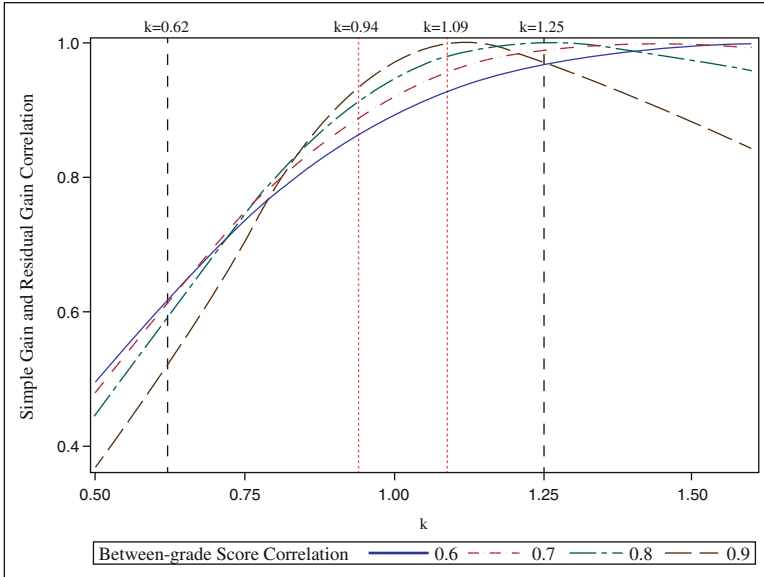


Fig. 1 Correlation of simple gain scores and residual gain scores for different between-grade score correlations and different values of k

the CSPP measures? Based on the known relationships between these measures, the following hypotheses are made.

1. Residual gain scores and PRRs are non linearly perfectly related because PRRs are just the percentile ranks of residual gain scores. The observed correlations between them could be less than perfect because of non linearity of their relationship but may still be very high.
2. The high correlations among the three CSPP measures and the perfect relationship between residual gain scores and PRRs may indicate that residual gain scores and other CSPP measures would also be highly correlated. Since none of these four measures require a vertical scale, the high correlations among them are not likely to be affected by the change of the vertical scale.
3. The high correlations among the three CSPP measures and the expected high correlations between residual gain scores and CSPP measures suggest that the correlations between simple gain scores and CSPPs may be similar to the correlations between simple gain scores and residual gain scores. For example, if the correlations between simple gain scores and residual gain scores are low, then it is expected that the correlations between simple gain scores and the CSPP measures would also be low.

Therefore, though residual gain scores and CSPPs do not require vertical scores, when these measures are based on scores from vertical scales, the extent of scale expansion or scale shrinkage of the vertical scale is expected to affect the corre-

lations between these growth measures and simple gain scores. The relationship revealed in Eq. (1) could be used not only to predict the correlations between simple gain and residual gain scores, but also to roughly predict the correlations between simple gain scores and the CSPR measures.

Below are two examples from the literature showing the variability of vertical scales that could result from different choices in the scaling process. These examples are used to demonstrate how Eq. (1) could be used to predict the relationships between simple gain scores and residual gain scores as well as their relationships with the CSPR measures.

3.1 Two Vertical Scaling Examples

Li (2015a) described the results of various scaling options using the Thurstone (1938) vertical scaling methodology, including different linking approaches, different data collection designs, and different variations within the Thurstone method. The study was based on simulated data mimicking the real vertical scaling data collection designs of a large scale testing program for grades 3 through 10. The k values of between-grade SD ratio calculated from the SDs provided in that study ranged from 0.94 to 1.09 across grade pairs and across different scaling options. This range of k values is represented by the two dotted vertical lines in Fig. 1. Correlations between simple gain scores and residual gain scores are expected to be high (above .8) for this set of vertical scales.

Another example came from Briggs and Domingue (2013). They discussed a few vertical scales constructed from empirical data using the IRT methods. Calculated from the SDs provided in their paper, the k values between adjacent grades across the different vertical scales ranged from 0.62 to 1.25, as indicated by the two dashed vertical lines in Fig. 1. Unlike the set of vertical scales from Example 1, this set of vertical scales were more variable and the consistency of rank orders based on simple gain and residual gain scores could vary a lot among these scales.

These two examples show how Eq. (1) can be used for predicting the consistency of growth measures based on simple gain and residual gain scores by the features of the vertical scales even at the stage of scale construction before longitudinal data are available. Hypotheses (1) through (3) stated previously can be applied to predict the relationships among other growth measures, but these hypotheses cannot be tested with the statistics reported in the two example research studies. The next section describes the use of empirical data in examining the relationships among all the growth measures: simple gain scores, residual gain scores, and the three CSPR measures.

3.2 Empirical Data Comparison

A longitudinal data set containing English test scores of about 240,000 students from two different grades was used for the analysis. The operational scale scores (Scale 1) of the two grades had similar standard deviations ($k = 1.024$) and the correlation of scores between grades was around .73. To demonstrate the impact of vertical scales, the operational scale scores were manipulated through linearly transforming the upper grade scores into scores on two fake scales. One manipulation resulted in a vertical scale (Scale 2) with scale expansion ($k = 1.638$) and the other resulted in a vertical scale (Scale 3) with scale shrinkage ($k = 0.617$). Simple gain scores, residual gain scores, SGPs, PRRs, and ECPRs were calculated for each student. Then the correlations among these growth measures were calculated and their scatter plots were examined.

Table 1 shows the correlations of the selected growth measures on the operational scale and on the two manipulated scales. The correlations between simple gain scores and residual gain scores varied depending on the score variability across grades. The correlation between these two measures was high (.94) on the operational scale (Scale 1) on which score variability were similar across grades. On the scale with the larger scale expansion (Scale 2), the correlation was even higher (.99), but on the scale with scale shrinkage (Scale 3), the correlation decreased to .61. These correlations indicated that the rank ordering of students or schools based on simple gain scores or residual gain scores were expected to be mostly consistent on the first two scales, but expected to vary considerably on Scale 3. All these observed correlations were consistent with what would have been expected by applying Eq.

Table 1 Observed correlations among five growth measures on three alternative scales

	Gain	Residual	ECPR	PRR	SGP
Scale 1 (Operational, $k = 1.024$)					
Gain	1.00	0.94	0.90	0.90	0.90
Residual	0.94	1.00	0.95	0.96	0.96
ECPR	0.90	0.95	1.00	0.99	0.99
PRR	0.90	0.96	0.99	1.00	0.99
SGP	0.90	0.96	0.99	0.99	1.00
Scale 2 ($k = 1.638$)					
Gain	1.00	0.99	0.94	0.95	0.94
Residual	0.99	1.00	0.95	0.96	0.95
ECPR	0.94	0.95	1.00	0.99	0.99
PRR	0.95	0.96	0.99	1.00	0.99
SGP	0.94	0.95	0.99	0.99	1.00
Scale 3 ($k = 0.617$)					
Gain	1.00	0.61	0.60	0.61	0.58
Residual	0.61	1.00	0.94	0.96	0.94
ECPR	0.60	0.94	1.00	0.98	0.96
PRR	0.61	0.96	0.98	1.00	0.97
SGP	0.58	0.94	0.96	0.97	1.00

(1), with the different values of k and the known between-grade score correlation of about .73. Consistent with findings from other research (Castellano & Ho 2013), the correlations among the three CSPR measures were very high on all three scales. The lowest was .96 between ECPR and SGP for the manipulated scale with severe scale shrinkage.

The correlations between residual gain scores and PRRs were consistently high on all three scales. As pointed out earlier, PRRs are the percentile ranks of residual gain scores, therefore these two measures are perfectly related. The less-than-perfect correlation between residual gain scores and PRRs was due to the non linearity of their relationship. The values of the correlations were similar across all three scales. These results were consistent with expectations in Hypothesis (1). The correlations between residual gain scores and each of the CSPR measures in Table 1 were also high (.94–.96). These correlations were also similar across the three scales. These results were consistent with expectations in Hypothesis (2).

The observed correlations between simple gain scores and the CSPRs were consistent with expectations in Hypothesis (3). Due to the high correlations between residual gain scores and the CSPRs, it was expected that the correlations between simple gain scores and the CSPRs would be similar to the correlations between simple gain and residual gain scores. As shown in Table 1, the correlations between simple gain scores and the CSPR measures were slightly lower than but close to the correlations between simple gain scores and residual gain scores. When simple gain scores and residual gain scores were highly correlated, the correlations between simple gain scores and the CSPRs were also high, as demonstrated in Scale 1 and Scale 2; when the correlation between simple gain and residual gain score dropped to .61 on Scale 3, the correlations between simple gain scores and the three CSPR measures also dropped to .61 or lower.

Scatter plots of these measures on the three scales are presented in Fig. 2, in which Gain and Residual refer to simple gain scores and residual gain scores, respectively. The scatterplots provided additional information regarding the relationships between these measures beyond the correlations. First, the plots showed nonlinear relationships not only between residual gain scores and PRRs but also between each of the three CSPR measures and the simple gain or residual gain scores. The nonlinear relationships led to the under estimation of the association between these measures when using the Pearson correlation statistic. These plots also showed that compared with the gain score metrics, the percentile rank metrics enlarged the differences between individuals in the middle of the score scale where there were more examinees and reduced differences between individuals at the extremes of the scales where there were fewer examinees. Second, even when the correlations among the CSPR measures were very high (e.g., .99 on Scale 2), there were still many outliers. Further investigations revealed that these outliers were mainly examinees at the extremes of the scale. When extreme scores with few examinees were excluded, the number of outliers decreased. Third, a closer comparison of the scatter plots among the CSPR measures showed that SGPs and ECPRs were more closely related to each other than with PRRs, which could be explained by the greater conceptual similarity between the first two. The scatter

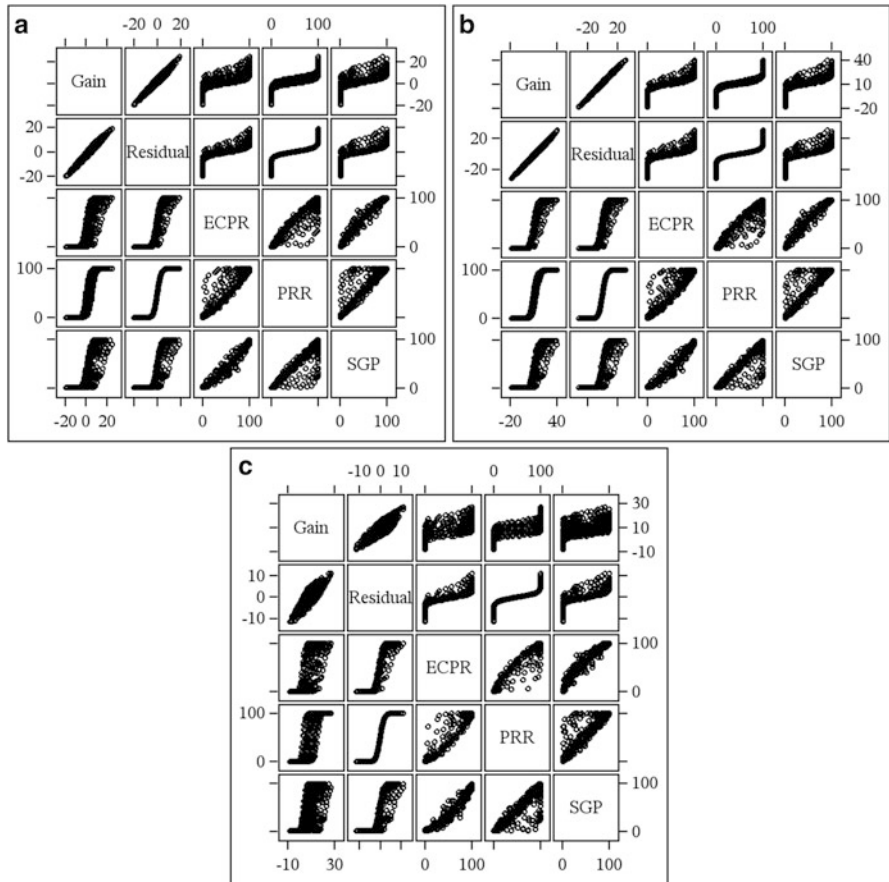


Fig. 2 Scatter plots of growth measures on (a) Scale 1, (b) Scale 2, and (c) Scale 3

plots also showed that for the extreme scores, students' PRRs tended to be higher than their SGPs or ECPRs on all the three scales. This observation was not expected, but might be related to the heterogeneity of the conditional score distributions.

4 Conclusion and Discussion

This study demonstrated the usefulness of a mathematical relationship as described in Li and Kolen (2011) in predicting the relationships between simple gain scores and residual gain scores on different vertical scales. Based on the conceptual similarities between some of the growth measures and findings from other research, it was hypothesized that the mathematical relationship was also useful in predicting the

approximate relationships between simple gain scores and a few CSPR measures. An empirical comparison of the growth measures was conducted which confirmed the predictions based on the mathematical relationship and other hypotheses. In addition, the empirical study also revealed some information that was not expected based on the mathematical relationship or the theoretical comparisons among the growth measures, including the many outliers that differed substantially between the highly correlated CSPR measures and the relatively higher estimates of PRRs compared to the other two CSPR measures. Though these findings might be explained by the scarcity of data for some score points or the heterogeneity of conditional score distributions, further investigations may be needed to see if the results generalize beyond the specific tests used for these analyses.

The mathematical relationship and the empirical comparison suggested that simple gain scores tended to rank order students more consistently with a few other measures, such as residual gain scores or CSPR measures, when the score variability increased from a lower grade to a higher grade, as opposed to when score variability decreased across grades. The extent to which these measures agreed also depended on the correlation of the scores between grade levels. This between-grade score correlation should remain constant among linearly related vertical scales but might differ across different grade pairs or different tests. This study was limited in that the growth measures were all based on data from two years. More research is needed to reveal the effect of vertical scales on relationships between measures based on modeling student growth trajectories over multiple years and residual gain scores or CSPRs that condition on scores from multiple prior years. Findings from this study and further investigations could provide useful information to both test developers and test users by revealing why different growth results could be highly consistent on one vertical scale but differ significantly on another vertical scale. Specifically, such information could help test developers to determine how different choices in constructing a vertical scale might impact the consistency of results from various growth measures and allow test users to understand to a greater extent how the growth measures used for decision making might have been affected by features of the particular vertical scales in use.

References

- Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York, NY: Taylor & Francis.
- Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51.
- Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*, 38(6), 551–576.
- Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3–14.
- Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practice*, 3(4), 15–16.

- Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student “growth” percentiles. *Journal of Educational and Behavioral Statistics*, 38(2), 190–215.
- Dunn, J., & Allen, J. (2009). Holding schools accountable for the growth of nonproficient students: Coordinating measurement and accountability. *Educational Measurement: Issues and Practice*, 28(4), 27–41.
- Goldschmidt, P., Choi K., & Beaudon, J. P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Retrieved from http://www.ccsso.org/Documents/2012/Growth_Model_Comparison_Study_Evaluating_School_Performance_2012.pdf.
- Grady, M., Lewis, D., & Gao, F. (2010). *The effect of sample size on student growth percentiles*. Paper presented at the 2010 annual meeting of the National Council on Measurement in Education. May 1–3, Denver, CO.
- Harris, D. W. (Ed.). (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.
- Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GEs. *Educational Measurement: Issues and Practice*, 3(4), 8–14.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.
- Lei, P., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement*, 36(1), 21–39.
- Li, D. (2015a). Investigations of the Thurstone scaling method in the ACT Aspire vertical scaling study. In W. Tao (Chair), *Constructing a vertical scale under linked scaling tests design*. Symposium conducted at the annual meeting of the National Council on Measurement in Education. April 15–19, Chicago, IL.
- Li, D. (2015b). *Relationships of growth measures from different plausible vertical scales*. Paper presented at the annual meeting of the National Council on Measurement in Education. April 15–19, Chicago, IL.
- Li, D., & Kolen, M. J. (2008). *Models of individual growth for school accountability—An empirical comparison*. Paper presented at the annual meeting of the American Educational Research Association. March 24–28, New York City, NY.
- Li, D., & Kolen, M. J. (2011). *Relationships between status, simple gain, residual gain, and linear growth*. Paper presented at the annual meeting of the National Council on Measurement in Education. April 7–11, New Orleans, LA.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68(5), 304–305.
- Maris, E. (1998). Covariance adjusted versus gain scores—Revisited. *Psychological Methods*, 3, 309–327.
- Paek, I., Young, M. J., & Yi, Q. (2008). The impact of data collection design, linking method, and sample size on vertical scaling using the Rasch model. *Journal of Applied Measurement*, 9(3), 229–248.
- Patz, R. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: The Council of Chief State School Officers.
- Phillips, S. E., & Clarizio, H. F. (1988). Conflicting growth expectations cannot both be real: A rejoinder to Yen. *Educational Measurement: Issues and Practice*, 7(4), 18–19.
- Roberts, D. M., & Burrill, D. F. (1995). Gain score grading revisited. *Educational Measurement: Issues and Practice*, 14(1), 29–33.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227–253.