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 the *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 17, Number 14, November 2012 ISSN 1531-7714

A Meta-Analysis of Growth Trends from Vertically Scaled Assessments

Nathan Dadey & Derek C. Briggs
University of Colorado at Boulder

A vertical scale, in principle, provides a common metric across tests with differing difficulties (e.g., spanning multiple grades) so that statements of *absolute* growth can be made. This paper compares 16 states' 2007-2008 effect size growth trends on vertically scaled reading and math assessments across grades 3 to 8. Two patterns common in past research on vertical scales, score deceleration (grade-to-grade growth that decreases over time) and scale shrinkage (variability in scale scores that decreases from lower to higher grades), are investigated. Pervasive, but modest, patterns of score deceleration are found for both math and reading. Limited evidence of scale shrinkage was found for reading, and virtually no evidence was found for math. In addition, linear regression was used to show that little of the considerable variability in the growth effect sizes across states could be explained by readily identifiable characteristics of the vertical scales. However, many scale characteristics were not well documented in available technical reports. The most important of these characteristics, along with their implications for interpretations of growth, are discussed. The results serve both as a normative baseline against which other scaling efforts can be compared.

American states and their school districts are increasingly implementing accountability policies that focus not only on levels of student achievement, but also on growth. To some extent this represents a reaction to a flaw in the No Child Left Behind legislation, something that seems to have been at least tacitly acknowledged by the federal government when it initiated the Growth Model Pilot Project in 2005 (U.S. Department of Education, 2005). The requirement for "clear approaches to measuring student growth" (U.S. Department of Education, 2009, p. 9) in the Race to the Top (RTTT) competition suggests that growth modeling will play a prominent role when the Elementary and Secondary Education Act is eventually reauthorized. This move towards growth modeling implies a desire to make *absolute* statements about how much any given student has learned in the subject domains of math and/or reading from one grade to the next. Such statements can only be directly supported when test scores have been vertically scaled.

A vertical scale places the scores of different tests onto a common metric so that, in principle, comparisons can be made between scores that span multiple grade levels. At present, two of the assessment consortia that were funded through RTTT grants are either considering (Partnership for the Assessment of Readiness for College and Careers), or are committed to (Smarter-Balanced) the development of vertical scales. One of the main motivations for to the development of these vertical scales is to support direct inferences about student growth. Although many states have already applied the methodology of vertical scaling to their assessment systems, it is not always clear whether this is being done with an eye toward modeling growth. To date there have been no efforts made to compare, across states, the patterns of growth in math and reading that are implied by preexisting vertical scales. This stands in stark contrast to the periodic efforts made to rank and compare states with regard to the obtained levels of student academic achievement (e.g., state NAEP results).

In this paper, we use vertically scaled test scores from students in 16 states during the 2007-08 school year to compare trends in grade to grade growth, in an effect size metric. As part of this comparison, we examine the extent to which the scales exhibit two related patterns that have previously been identified in the literature: score deceleration (grade-to-grade growth that decreases over time; Yen, 1986) and scale shrinkage (variability in scale scores that decreases from lower to higher grades; Camilli, 1988; Camilli, Yamamoto, & Wang, 1993; Camilli, 1999; Yen, 1986; Yen & Burkett, 1997). We also use a regression-based approach to examine the amount of variability in the growth trajectories that is explained by readily identifiable characteristics of the vertical scales. By conducting what amounts to a meta-analysis of growth trajectories, we provide a normative baseline against which both contemporary and future vertical scaling efforts can be compared. A key point we will emphasize is that there is considerable variability in the growth effect sizes that are observed across states. This can complicate the use of "average" growth trends to evaluate whether the estimated effect of an educational intervention is practically significant (Hill et al, 2008).

Methods

Data

Between the Fall of 2008 and the Fall of 2009 we visited the web sites for 24 states that had been reported to have vertical scales in the annual "Quality Counts" issue produced by *Education Week* in 2008. For five of these states we found no information to support the assertion that any of their tests had been vertically scaled. As part of our search process we also examined the websites of the 26 states reported as not having vertical scales by *Education Week*, and found that two other states did in fact have vertical scales. This left us with a total population of 21 states with vertical scales spanning a minimum of grades 3 through 8 in math and reading.

We subsequently reviewed the 2007-08[1] technical manual and/or interpretive guide associated with each state's criterion-referenced assessment. From these publicly available documents, we compiled the mean scale scores and standard deviations for each state's math and reading assessments in grades 3 through 8. There were seven states for which the descriptive statistics were not publicly and electronically available; we made formal requests for this information from each state's department of education, and, in some cases, their test vendors. Despite our best efforts, there were five states for whom we were unable to obtain any descriptive statistics related to their vertical scales. Ultimately, we were able to collect data for 16 states: Arizona, Colorado, Delaware, Florida, Idaho, Illinois, Indiana, Missouri, New Mexico, North Carolina, North Dakota, Oregon, South Dakota, West Virginia, Wisconsin and Wyoming. In addition to the descriptive statistics, we also gathered information on variables that are generally considered relevant to the creation and maintenance of a vertical scale (Briggs & Weeks, 2009; Kolen & Brennan, 2004): the response model used for the item calibration (Rasch Model/Partial Credit Model vs. Three Parameter Logistic Model/Generalized Partial Credit Model), test administration date (Fall vs. Spring), whether the test reported a single score for reading or both reading and writing, and the age of the vertical scale (i.e., the number of years over which the scale has been maintained). However, there are a variety of variables that are also very relevant to the vertical scaling process that we were not able to obtain, a limitation to which we return in our discussion section.

In this paper we standardize gains from grade-to-grade such that for a given test subject (math or reading), the growth for state i (i = 1, 2, …, 16) from grade g to g+1 (g = 3, 4, …, 7) is characterized by the effect size

$$Y_{ip} = \frac{\bar{\theta}_{i(g+1)} - \bar{\theta}_{ig}}{\sqrt{\frac{\hat{\sigma}^2_{i(g+1)} + \hat{\sigma}^2_{ig}}{2}}}$$

where the subscript p (p = 1, 2, …, 5) indicates one of five adjacent grade pairs between grades 3 and 8, $\overline{\theta}_{i(g+1)}$ is the mean scale score reported for the higher grade for grade pair p, $\overline{\theta}_{ig}$ is the mean scale score reported for the lower grade, and $\hat{\sigma}^2$ is the reported variance of the scale scores. These effect sizes are nested within states, so for each of the 16 states there are five effect size statistics for a total of 80 grade-pair effect sizes in each subject.

We later use Y_{ip} as the dependent variable in a meta-analytic regression, with effect sizes within states as the units of analysis and state-specific design factors as

[1] If the test was administered in the Fall we used the 2007 technical manual and if it was administered in the Spring we used the 2008 technical manual, so that all information came from the 2007/2008 school year. Due to issues of availability, we used data from 2007, the 2006/2007 school year, for West Virginia (Spring).

predictor variables. These predictor variables are as follows[2]:

1. **Time**. The variable "Time" in our regressions takes on values {0, 1, 2, 3, 4} with 0 representing the effect sizes, or growth, from grades 3 to 4, 1 representing growth from grades 4 to 5, and so on up until the Time variable takes on a value of 4 for growth from grades 7 to 8.

2. **Fall vs. Spring Administration**. Growth based on grade to grade comparisons from tests that are administered during the spring has a different interpretation relative to tests that are administered in the fall. For example, the effect size computed from a fourth grade and fifth grade test that are both administered in the spring represents growth that occurred mostly in fifth grade, while fall administration in the same grades represents growth the occurred mostly in the fourth grade. Three out of the 16 states in our samples tested students in the fall rather than the spring. For the regression analysis the dummy variable, the variable "Fall", takes a value of 1 when the test was administered in the fall and 0 otherwise.

3. **Tests that Combine Reading and Writing.** Scales created from tests that combine items assessing both reading and writing might represent a different construct than scales created from tests that only assess reading comprehension. We have designated these combined reading and writing tests as measuring English Language Arts (the variable "ELA" in our subsequent regression analysis, which takes a value of 1 when reading and writing tests have been combined and 0 otherwise). There were three states for whom this designation applied.

4. **IRT Model.** Our choice to include the IRT model used to create the vertical scale stems from work which suggests that systematic differences between Rasch and 3PL scalings may exist (Briggs & Weeks, 2009; Yen & Burket, 1997). For the regression analysis we created a variable, 3PL, which takes the value of 1 when the 3PL model was used for scale calibration and a 0 when the state used the Rasch model[3]. Nine out of the 16 states considered here used the 3PL model.

Results

Growth in Effect Size Units

Figure 1 below provides a graphical representation of growth patterns across the 16 states for reading and math respectively; Table 1 presents the corresponding descriptive statistics. As is evident in the plots and tables, there is substantial variability in effect sizes both within and between adjacent grade-pairs.

There is evidence of growth deceleration in both subjects. In reading, there is a significant drop in growth (0.21 effect size units) from grades 3-4 (mean effect size = 0.55) to growth in grades 4-5 (0.34). However, the size of this drop is partially due to the influence of Wyoming, which has a pattern of effect sizes that is a good deal more variable than the patterns found in other states. If this state is excluded, the sharp drop in growth from grade 4-5 to 3-4 becomes less pronounced (from a decrease of 0.21 to a decrease of 0.12) while the decrease in growth from grades 4-5 to 5-6 becomes more pronounced (0.06 to 0.10). In the remaining grades mean growth is mostly constant, regardless of the inclusion or exclusion of Wyoming. In math, there is no pronounced drop in growth from grades 3-4 to grades 4-5, however, in contrast to reading, there are gradual decreases from grades 5 to 8. To summarize growth trends with a single statistic, we regressed, for each state, effect sizes on time. The resulting slopes show clear evidence of downward trends (in reading 15 of the 16 slopes are negative and in math 14 of the slopes are negative), although the magnitudes of these slopes are small (see Appendix B for plots of these regressions). The average slope in reading was -0.05 with an SD of 0.04, which translates into an average decrease of 0.20 effect size units across grades 3 to 8. In math the average slope was -0.06 with an SD of 0.05, for a total effect size decrease of 0.24 across grades 3 to 8. The slopes and intercepts from these within-state regressions have a strong negative relationship ($r = -0.79$ in reading and -0.87 in math), indicating that states with vertical scales with above average growth from grades 3 to 4 are also those with above average amounts of growth declines. Another finding worth noting is that the slopes

[2] We also examined another continuous variable, scale age, which is the number of years since the scale was initially established. We had hypothesized that scale age might be a proxy for item parameter drift (e.g., Bock, Muraki, & Pfeiffenberger, 1988), however this variable lacked predictive power in our regression. Because of this finding, combined with the complexity added to interpretation of the regression results, we excluded this variable from our analysis.

[3] In many of the states considered, the underlying tests consisted of mixed format items. In such instances, a combination of either the 3PL and Generalized Partial Credit Model (GPCM) or Rasch and Partial Credit Model (PCM) were used to calibrate the scale. We used "3PL" and "Rasch" as shorthand for such scenarios.

are rather strongly correlated across subjects (r = 0.73), while the intercepts are not (r = 0.23). Thus the increase or decrease in growth for states between reading and math is associated, but the amount of initial growth is not Overall, the reading and math effect sizes were equally variable, with average SDs across grade-pairs of 0.18 and 0.17 respectively. However, the trends in these SDs differ. In the 3 to 4 and 4 to 5 grade pairs, the variability in effect sizes is much larger for reading than it is for math; yet as of the grade 5 to 6 pairing, the SD in math is much larger than the SD in reading (0.25 in math and 0.14 in reading). In the final two grade pairings (6 to 7 and 7 to 8) there is no real difference in the variability by test subject.

Figure 1. Effect Size Trajectories over 16 States for Reading and Math. Note: The large red dots represent the mean ES across the 16 states within each grade pair. The horizontal bars represent +/- 1 SD of the ES.

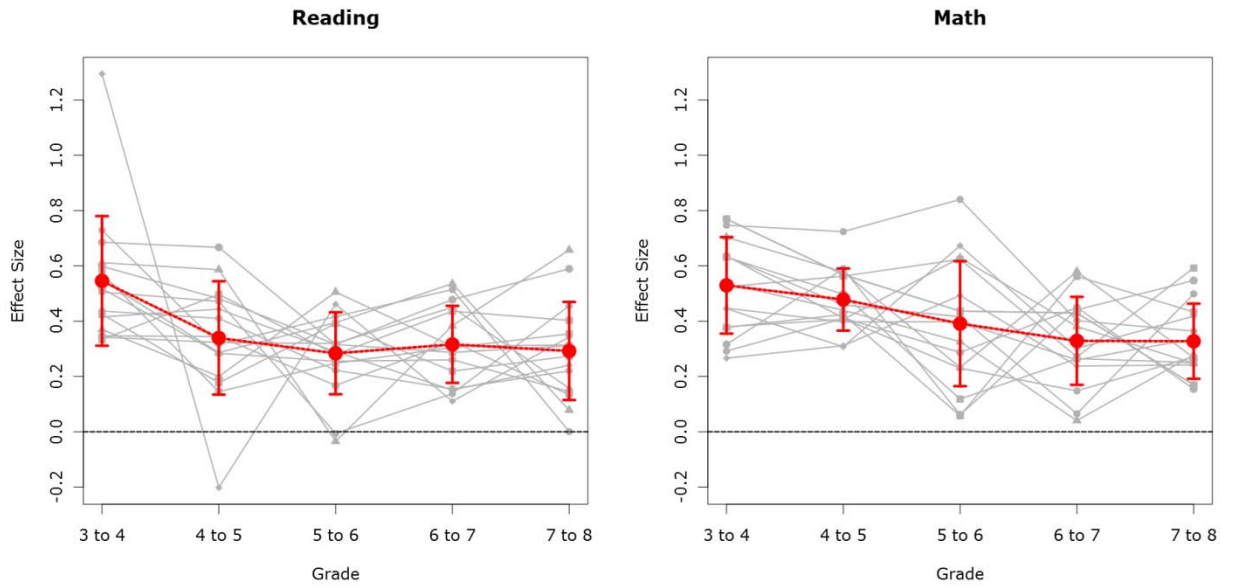


Table 1. Descriptive Statistics for Effect Sizes in Reading and Math

	Grade-Pair				
	3 to 4	**4 to 5**	**5 to 6**	**6 to 7**	**7 to 8**
Reading					
Min	0.33	-0.20	-0.04	0.11	0.00
Mean[1]	0.55	0.34	0.28	0.32	0.29
Median	0.51	0.33	0.30	0.31	0.29
Max	1.29	0.67	0.51	0.54	0.66
SD	0.23	0.21	0.15	0.14	0.18
Math					
Min	0.27	0.31	0.06	0.04	0.15
Mean	0.53	0.48	0.39	0.33	0.33
Median	0.53	0.45	0.41	0.34	0.28
Max	0.77	0.72	0.84	0.58	0.59
SD	0.17	0.11	0.23	0.16	0.14

Notes: N = 16. Excluding Wyoming, which has a highly variable effect size pattern, the mean effect sizes in grade pairs 3-4 to 7-8 are 0.50, 0.38, 0.27, 0.33 and 0.29, respectively.

Scale Shrinkage

To characterize the extent to which scale shrinkage is evident across grades in a given state, we express the grade 4 to 8 SDs as a proportion of the grade 3 SDs. Figure 2 plots the results in reading and math for each state. Several trends are evident. First, from visual inspection, one would be hard-pressed to conclude that there are signs of widespread scale shrinkage. While many states have SDs that decrease over time, the decreases tend to be very small (the average decrease in SD from grades 4 to 8 is -0.08 for both reading and math). Second, scale shrinkage occurs more often in reading than math. Visual inspection of Figure 2 bears this finding out, as well as analyses we conducted by regressing each state's SD trend on time. In reading, 9 of the 16 slopes were negative while in math 5 slopes were negative. Third, all of the standard deviation patterns are non-monotonic – displaying "spikes" or "dips" for certain grades.

Explaining Variability in Effect Sizes

We now examine how much of the between state variability in the effect sizes can be accounted for by variables that capture certain methods used by each state to establish and maintain their vertical scales. The results from regressing the grade-pair effect size statistics Y_{ip} on the predictors above are shown in Tables 2 and 3.

Figure 2. Scale Score Standard Deviations (expressed as ratio grade 3 SD). States with strong visual evidence of scale shrinkage are shown in red, while the remaining states are shown in grey

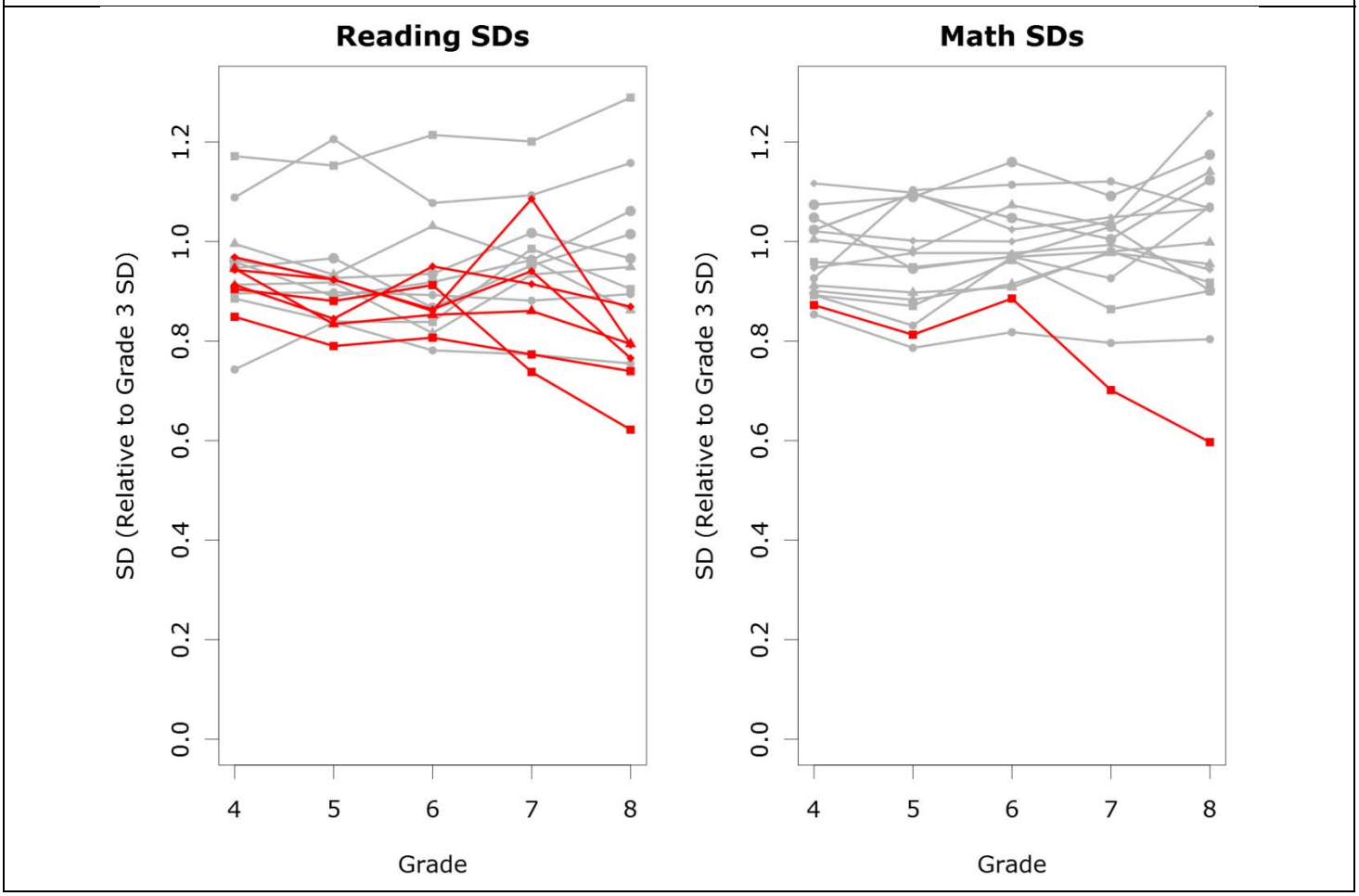


Table 2. Regressions for Reading.

	Model				
Predictor	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Intercept	0.462 *	0.452 *	0.475 *	0.483 *	0.477 *
Time	-0.053 *	-0.048 *	-0.054 *	-0.045 +	-0.042 +
Fall		0.049			0.041
ELA			-0.070		-0.048
3PL				-0.038	-0.025
Time*Fall		-0.025			-0.021
Time*ELA			0.006		0.010
Time*3PL				-0.014	-0.016
Statistic					
RMSE	0.191	0.193	0.192	0.190	0.195
R^2	0.136	0.141	0.149	0.165	0.172
$R^2\Delta$ from Base Model		0.005	0.013	0.029	0.036

Notes: + $p < 0.10$, * $p < .05$, N = 80

Table 3. Regressions for Math.

	Model			
Predictor	**(1)**	**(2)**	**(3)**	**(4)**
Intercept	0.522 *	0.471 *	0.486 *	0.448 *
Time	-0.055 *	-0.043 *	-0.042 *	-0.033 +
Fall		0.269 *		0.264 *
3PL			0.064	0.043
Time*Fall		-0.067 *		-0.065 *
Time*3PL			-0.023	-0.018
Statistic				
RMSE	0.164	0.153	0.165	0.154
R^2	0.189	0.317	0.200	0.322
$R^2\Delta$ from Base Model		0.128	0.011	0.133

Notes: + $p < 0.10$, * $p < .05$, N = 80

The base model in each table (model 1) only includes the time variable as a predictor. These models accounts for 14% and 19% of the variability in effect sizes in reading and math respectively. For reading (Table 2), the intercept of 0.462 represents the average grade 3 to 4 growth across the 16 states, and the slope coefficient of -0.053 represents the average change in growth across adjacent grade pairings from grades 4 to 8. In other words, without knowing anything else about a state's vertical scale, one would predict that growth in reading from grades 3 to 4 would be about 0.462 effect size units and growth from grades 4 to 5 would be 0.409. Cumulatively, the model predicts that growth in reading will decline to 0.250 effect size units by grades 7-8. For math, the intercept and slope for model 1 are 0.522 and -0.055, which translates into a predicted decline to 0.302 by grades 7-8. The two base models suggest the same basic trend of linear growth declines in reading and math. However, there is a large degree of imprecision in these predictions: the root mean square error for the base regression in is 0.191 effect size units in reading and 0.164 in math.

Subsequent models in reading (models 2-4) show the marginal impact of adding an additional predictor variable to the base model. In reading these variables are Fall, ELA and 3PL that were described above. Each variable is added to the base model as both a main effect (influencing the interpretation of average grade 3 to 4 growth, i.e., the intercept) and as an interaction with the time variable (influencing the interpretation of the grade to grade growth trend, i.e., the slope). Finally, model 5 includes all covariates together in a full specification. In reading, none of covariates have a statistically significant[4] impact on either the intercept of the slope, and the full model only increases R^2 from the base model from 0.136 to 0.172.

We followed a similar process for our math regressions, in this case adding two new covariates to the model, Fall and 3PL. In contrast to reading, the Fall variable has a significant impact on the interpretation of the intercept and the slope when added to the base model. Under model 2, the 13 states that test their students in the spring have base growth of 0.471 effect size units; the 3 states that test their students in the fall have base growth of 0.740 effect size units. In addition, because they start with higher base growth, the deceleration trend for these 3 states is much stronger at -0.110. This is about two and half times larger than the trend of states testing students in the spring (-0.043). As we noted earlier, states testing students in the fall are really testing growth that had occurred a year prior than is indicated by the grade pair. So for these states, the intercept is more properly understood as growth from grade 2 to 3 rather than growth from grade 3 to 4. The results here are consistent with the notion that growth deceleration in math is strongest in the early elementary grades. For math, going from the base model to the full model increases R^2 from 0.189 to 0.322, primarily due to the impact of including the Fall variable.

[4] The reader will note that in presenting our results we do not emphasize tests of homogeneity and statistical significance as is typical in other meta-analytic contexts (i.e., Hedges & Olkin, 1985). This is because in the present context, the effect sizes being computed within a given state are based on the entire population of test-takers, so there is little sense in characterizing this as a source of sampling variability. At the state level, one might imagine a hypothetical population of states that could have developed vertical scales, but it is quite a stretch to suggest that the 16 included in the present study represent a random sample from this hypothetical population. Hence while we do flag predictor variables with conventional p-values less than 0.10 or 0.05 when presenting our results, this should be taken with a grain of salt. Our emphasis is on statistical description rather than statistical inference (Briggs, 2005).

Discussion

It is not entirely clear how one should interpret the large degree of unexplained variability in growth patterns between and within states evident in our meta-analysis. On the one hand, we might assume that scores along each state's vertical scale are intended as measures of the same general reading and math "constructs." After all, many of these vertical scales have been developed by the same testing contractors, often using the same anchor items from nationally normed vertical scale batteries. If the constructs being measured were comparable, then it would be natural and desirable to speculate about possible reasons that growth across grades in one state is larger in magnitude than growth across grades in another. Perhaps one state has stronger curricula than the other, better professional development for its teachers, etc. On the other hand, although the higher order subject standards for students (i.e., number sense, algebra, geometry & measurement, etc.) are often very similar from state to state, the specific indicators used to design test items may well differ dramatically, as can the alignment of each state's test to the enacted curricula in the elementary and middle school grades.

To the extent that the reading and math constructs measured by each state test are operationally defined through the alignment of test items to the content standards, then growth will only have a similar operational definition with respect to the way that linking items have been selected to overlap across adjacent grades. The growth that is observed from grade to grade in a given state will depend largely upon the design principles that were used to select linking items (e.g., are items chosen to be representative of the content domains for each of two adjacent grades, or a common domain that overlaps across multiple grades?), and the extent to which these linking items are representative of the content domain and also instructionally sensitive (e.g., are items chosen to represent content that teachers emphasize in the enacted curriculum?). Defining and coding state-level variables that capture these sorts of design differences was not possible when conducting our secondary data analysis of publicly available reports. This is an obvious limitation to our study, but also reflection of the quantity and quality of information provided by states in the publically available documentation of their vertical scaling process. It is worth keeping in mind that in many cases, states did not publicly report the grade to grade means and SDs that were needed for us to compute growth in effect size units, and even after repeated requests, five states with vertical scales were unable to provide us with these summary statistics.

Although a priori design decisions are quite likely to explain a considerable amount of the variability in grade to grade growth patterns across vertical scales, so do the decisions that are made by psychometricians when calibrating and scaling vertically linked tests (Schafer, 2009). Briggs & Weeks (2009) show that choice of IRT model, linking approach, estimation method, and the interactions between these factors can have significant impacts on the magnitude of growth in effect size units when the factors are used in combination on the same longitudinal set of item responses. Weeks (2011) demonstrates that growth interpretations can be distorted when the dimensional composition of a construct, that shifts over time, is modeled as though it were unidimensional (see also Martineau, 2004). And Harris (2007) points out that the maintenance of a vertical scale over time through horizontal equating can lead to shifts in grade to grade growth that are at least in part a reflection of equating errors. To this list we add another psychometric practice that has not received as much attention in the literature, but which figured prominently in the scaling approaches described for two of the states in our sample: nonlinear transformations of the vertical scale.

Kolen & Brennan (2004) have argued that it is reasonable to nonlinearly transform a vertical scale so long as the state has developed a "conceptual definition of growth" and communicated this to the test developer:

The theta scale also can be nonlinearly transformed to provide for growth patterns that *reflect the kind of patterns that are expected* [emphasis added]. Consider a situation in which a test developer believes that the variability of scale scores should increase over grades. If the variability of the theta estimates is not found to increase over grades, a nonlinear transformation of the ability scale might be used that leads to increased variability. (p. 393)

Instances of these sorts of practices were readily found in the technical manuals of two of the states in our sample. When vertical scales were being established in each of these two states it was found empirically that the mean scale scores in a higher grade were *lower* than those found in the immediately adjacent lower grade after the tests were vertically linked. Rather than report these results, the states—in consultation with their test contractors—decided to adjust the upper grade scale scores so that the reported mean was that which would have been observed if successive grade means followed a polynomial trend. In other words, the vertical scales were nonlinearly transformed. If, in fact, nonlinear transformations are deemed admissible when vertical scales are being established, then it follows that it would be possible to engineer *any* pattern of effect sizes that would be desired. This suggests that the underlying scales have only ordinal properties, making it potentially meaningless to compare grade to grade growth trends in terms of magnitude, e.g. in effect size units, across different scales. There is considerable confusion in the psychometric literature as to whether the use of IRT methods will produce a score scale with interval properties (Ballou, 2009; Briggs, 2010; in press; Michell, 1990; 1999; Yen, 1986). As it turns out, interval properties are quite critical if ones wishes to make comparisons of growth across states in terms of absolute differences in magnitudes.

Implications for Practice

In summary, the trends found in this analysis are consistent with the assertion that, on average, growth on a vertical scale in math and reading appears to decrease as students move from the early grades of elementary school to the last grade of middle school. An nonparametric examination of effect size trends indicates a roughly constant deceleration of effect sizes in math, while in reading there is a more rapid deceleration between grades 3-4 to grades 4-5 followed by a flat trend of no acceleration or declaration. However, part of this rapid deceleration can be explained by a single outlying state (Wyoming), with a grade 3 to 4 effect size of 1.3 followed by a grade 4 to 5 effect size of -0.20. When we use a regression analysis to summarize grade to grade growth trends we find evidence for cumulative effect size declines in math and reading of about 0.25 from grades 3-4 to 7-8. A very small amount of the total variability in effect sizes could be explained by our state-level variables (up to an additional 4% in reading and 13% in math). Interestingly, we found that controlling for whether or not a state tested its students in the fall or spring of a given grade had a significant impact on growth interpretations for math, but not for reading. This finding of no impact when controlling for fall testing for reading vertical scales is surprising as previous research has found strong evidence of growth deceleration occurring in the earliest grades of schooling (Hoover, 1984). Finally, our analysis finds limited evidence of scale shrinkage in reading, and almost none in math.

At a minimum, our results provide a normative context that any state with a vertical scale could use to compare a given growth pattern with the patterns that have been observed in other states. For example, a quick perusal of Figure 1 makes clear that Wyoming has a growth pattern between grades 3 and 5 that is well outside

the range of 15 other states. A result such as this might be grounds for a state's director of assessment and testing contractor to engage in some detective work to better understand why their growth appears so unusual.

However, beyond this normative baseline, caution must be used when interpreting and generalizing these results. Hill et al. (2008) have recommended the use of grade to grade gains (in effect size units) from vertically scaled assessments as a basis for evaluating the practical significance of an educational intervention. The logic here is that the average grade to grade gains along a vertical scale can be interpreted as the magnitude of achievement growth that would be observed as a consequence of all the different factors that cause students to learn. Given this, one would expect any single intervention to have an effect on achievement that is some proportion of this average. If the proportion is large, one would conclude that the intervention has an effect that is practically significant. Our meta-analysis points to a potential flaw in taking this approach, in that it uses national averages which indicate more stability in grade to grade growth than is warranted empirically in any given state. Consider a hypothetical reading intervention that produces an effect size of 0.20 from grade 7 to 8. For Arizona, this effect size is almost twice the state's grade 7 to 8 growth of 0.13, hence we one might conclude that the impact of the intervention is not just practically significant, but dramatically so. However, the same results would be given a much different interpretation in New Mexico, where the state's average growth from grade 7 to 8 growth is 0.59. Furthermore, the use of grade to grade effect sizes for assessing practical significance assumes that these magnitudes have an absolute interpretation. As we have noted in the previous section, such an interpretation becomes problematic if the scales have been manipulated in a matter (i.e., subjected to nonlinear transformations) that presumes they only communicate ordinal information.

There are some clear advantages, in principal, to having a vertical score scale. First, only a vertical scale makes it possible to directly compare student growth in terms of criterion referenced changes in magnitude. Second, in a computer adaptive testing context a vertical scale facilitates out of level testing. Third, having item difficulty estimates across grades located on a single continuum makes it easier to set proficiency cutpoints coherently during standard-setting. And fourth, the biggest potential advantage is that grade to grade gains from a vertical scale can serve as a basis for evaluating whether standards, curriculum and instruction, and assessment appear to be properly aligned across grades. When low or even negative mean growth is observed, it provides a clear signal that something is amiss. In the absence of a vertical scale, such misalignment may be harder to detect.

However, it is also important to note that a vertical scale is not necessary for many common uses of test scores, including purposes of educational accountability. For example, a vertical scale is generally not necessary when test scores are being used to evaluate teachers and/or schools with a value-added model (Briggs & Domingue, in press). In these contexts, a variety of alternative approaches can also be employed to make *normative* statements about student growth. A prominent example is the student growth percentile approach popularized by Betebenner (2009). Another alternative to vertical scales, growth scales (Schafer & Twing, 2006; Schafer, 2006), relies on vertically articulated content standards (Ferrara, Johnson & Chen, 2005) to make statements about student growth.

References

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275–285.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*(4), 351–383. MIT Press.

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

Braun, H. I. (2005). Value-added modeling: What does due diligence require? In R. E. Lissitz (Ed.), *Value added models in education: Theory and application* (p. 19-39). Maple Grove, MN: JAM Press.

Briggs, D. C. (2005) Meta-analysis: a case study. *Evaluation Review*, 29(2), 87-127.

Briggs, D. C. (2010). The problem with vertical scales. Paper presented at the 2010 Annual Meeting of the American Educational Research Association, Denver, CO, May 3, 2010.

Briggs, D. C. (in press). Measuring growth with vertical scales. *Journal of Educational Measurement.*

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.

Briggs, D. C. & Domingue, B. (forthcoming). The gains from vertical scaling. *Applied Measurement in Education.*

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics, 13*(3), 227-241.

Camilli, G. (1999). Measurement error, multidimensionality and scale shrinkage: A reply to Yen and Burkett. *Journal of Educational Measurement, 36*(1), 73-78.

Camilli, G., Yamamoto, K., &Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17* (4), 379-388.

Education Weekly. (2008). *Standards, Assessments, and Accountability*. Retrieved From: http://www.edweek.org/media/ew/qc/2008/18sos.h27.saa.pdf

Ferrara, S. F., Johnson, E. & Chen, W. H. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education*, 18(1), 35-59.

Harris, D. (2007). Practical issues in vertical scaling. In N. Dorans, M. Pommerich & P.. Holland (eds) *Linking and aligning scores and scales*, 233–251, Springer.

Hedges, L. V. & I. Olkin. (1985). *Statistical methods for meta-analysis*. New York, Academic Press, Inc.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

Hill, R., Marion, S., DePascale, C., Dunn, J., Simpson, M. A. (2006). Using value tables to explicitly value student growth. In R. W. Lissitz (2006), *Longitudinal and value added models of student performance* (255-282). Maple Grove, Minnesota: JAM 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. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Martineau, J. A. (2004). *The effects of construct shift on growth and accountability models.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

McCaffrey, D. F., Lockwood, J.R., Koretz, D., Thomas, L. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Schafer, W. D. (2006). Growth Scales as an Alternative to Vertical Scales. *Practical Assessment Research & Evaluation*, 11(4). Available online: http://pareonline.net/getvn.asp?v=11&n=4

Schafer, W. D., & Twing, J. S. (2006). Growth scales and pathways. In R. W. Lissitz (2006), *Longitudinal and value added models of student performance* (321-344). Maple Grove, Minnesota: JAM Press.

U.S. Department of Education. (2005, November 18). *Secretary spellings announces growth model pilot, addresses Chief State School Officers' annual policy forum in Richmond* [Press Release]. Available from http://www.ed.gov/news/pressreleases/2005/11/11182005.html

U.S. Department of Education. (2009). Race to the Top Executive Summary. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf on March 3rd, 2011.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In Jason Millman (Ed.). Grading teachers, grading schools, Is student achievement a valid evaluation measure? (pp. 137-162). Thousand Oaks, CA:Corwin Press.

Weeks, J. P. (2011). *Is math always math? Examining achievement growth in multiple dimensions.* Unpublished doctoral dissertation, University of Colorado, Boulder, CO.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*, 399-410.

Yen, W. M. (1986). The choice of scale for educational measurement: an IRT perspective. *Journal of Educational Measurement, 23*, 299-325.

Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34* (4), 293-313.

Appendix A: Scale Score Effect Sizes and Standard Deviations, with Corresponding Predictor Variables

Table 1A. Effect Sizes and Standard Deviations for Reading Scale Scores with Corresponding Predictor Variables, Ordered From Largest to Smallest Mean Effect Size Across Grades.

	Effect Size							Standard Deviation (as ratio of grade 3 SD)					Predictor Variable*			
	3 to 4	**4 to 5**	**5 to 6**	**6 to 7**	**7 to 8**	**SD**	**Mean**	**4**	**5**	**6**	**7**	**8**	**3PL**	**Fall**	**Age**	**ELA**
New Mexico	0.685	0.667	0.316	0.478	0.685	0.153	0.547	0.947	0.966	0.867	0.952	1.015	0	0	3	0
Delaware	0.611	0.586	-0.035	0.383	0.611	0.286	0.441	0.913	0.918	0.816	0.934	0.949	0	0	9	0
Wyoming	1.294	-0.202	0.461	0.110	1.294	0.560	0.401	0.913	0.845	0.950	0.914	0.869	0	0	3	0
North Carolina	0.598	0.482	0.316	0.287	0.598	0.136	0.399	0.849	0.790	0.807	0.773	0.739	1	0	2	0
North Dakota	0.729	0.316	0.417	0.514	0.729	0.268	0.395	1.088	1.206	1.077	1.093	1.158	1	1	3	0
Idaho	0.333	0.497	0.280	0.436	0.333	0.085	0.390	0.885	0.839	0.838	0.985	0.905	0	0	--	0
Oregon	0.547	0.286	0.395	0.536	0.547	0.195	0.368	0.946	0.834	0.853	0.860	0.794	0	1	2	0
South Dakota	0.370	0.200	0.505	0.309	0.370	0.111	0.340	0.995	0.933	1.031	0.962	0.862	0	0	6	0
Florida	0.573	0.145	0.248	0.304	0.573	0.159	0.325	0.905	0.880	0.912	0.738	0.622	1	0	7	0
Indiana	0.342	0.323	0.320	0.450	0.342	0.105	0.318	0.943	0.924	0.861	1.086	0.792	1	0	6	1
Missouri	0.505	0.471	-0.006	0.138	0.505	0.232	0.313	0.896	0.898	0.892	0.881	0.894	1	0	2	1
Illinois	0.416	0.443	0.289	0.151	0.416	0.122	0.308	0.969	0.923	0.864	0.941	0.766	1	0	2	0
Wisconsin	0.425	0.177	0.391	0.219	0.425	0.108	0.297	1.171	1.153	1.214	1.201	1.289	1	1	3	0
West Virginia	0.518	0.284	0.255	0.260	0.518	0.137	0.292	0.963	0.927	0.935	1.017	0.966	1	0	4	1
Colorado	0.437	0.410	0.223	0.155	0.437	0.126	0.289	0.743	0.838	0.781	0.773	0.755	1	0	6	0
Arizona	0.347	0.345	0.168	0.326	0.347	0.104	0.264	0.959	0.890	0.918	0.963	1.061	0	0	3	0

*Notes: 3PL = 1 for the 3PL/GPCM, 0 otherwise; Fall = 1 for a Fall Test Administration, 0 otherwise; Age is the scale age (in years); ELA = 1 if the assessments were a combination of reading and writing, 0 if the assessment tested reading only.

Table 2A. Effect Sizes and Standard Deviations for Math Scale Scores with Corresponding Predictor Variables, Ordered From Largest to Smallest Mean Effect Size Across Grades.

	Effect Size					Standard Deviation (as ratio of grade 3 SD)						Predictor Variable*		
	3 to 4	**4 to 5**	**5 to 6**	**6 to 7**	**7 to 8**	**SD**	**Mean**	**4**	**5**	**6**	**7**	**8**	**3PL**	**Fall**	**Age**
North Dakota	0.747	0.724	0.840	0.402	0.364	0.268	0.615	0.892	0.831	0.969	0.926	1.070	1	1	3
Idaho	0.524	0.563	0.624	0.271	0.592	0.085	0.515	0.892	0.870	0.963	0.864	0.901	0	0	--
Wisconsin	0.770	0.568	0.436	0.430	0.169	0.108	0.475	1.023	1.095	1.047	1.005	1.123	1	1	3
West Virginia	0.770	0.568	0.436	0.430	0.169	0.137	0.475	1.023	1.095	1.047	1.005	1.123	1	0	4
Oregon	0.703	0.578	0.233	0.579	0.268	0.195	0.472	1.004	0.981	1.073	1.031	1.140	0	1	2
Indiana	0.445	0.307	0.673	0.301	0.418	0.105	0.429	1.021	1.002	1.000	1.041	1.257	1	0	6
Missouri	0.633	0.459	0.416	0.065	0.499	0.232	0.414	0.926	1.103	1.114	1.121	1.067	1	0	2
South Dakota	0.380	0.406	0.629	0.380	0.248	0.111	0.409	0.901	0.883	0.914	0.977	0.955	0	0	6
New Mexico	0.316	0.589	0.061	0.441	0.547	0.153	0.391	1.048	0.945	0.969	1.030	0.901	0	0	3
Arizona	0.636	0.411	0.287	0.450	0.154	0.104	0.388	1.074	1.089	1.160	1.091	1.174	0	0	3
Florida	0.377	0.424	0.058	0.562	0.433	0.159	0.371	0.872	0.813	0.885	0.702	0.597	1	0	7
Illinois	0.447	0.397	0.398	0.260	0.336	0.122	0.368	0.947	0.977	0.977	0.993	0.944	1	0	2
North Carolina	0.631	0.497	0.118	0.265	0.250	0.136	0.352	0.959	0.948	0.969	0.979	0.918	1	0	2
Delaware	0.536	0.439	0.326	0.040	0.278	0.286	0.324	0.912	0.897	0.908	0.980	0.998	0	0	9
Wyoming	0.266	0.311	0.493	0.238	0.242	0.560	0.310	1.117	1.098	1.024	1.049	1.066	0	0	3
Colorado	0.292	0.410	0.230	0.148	0.271	0.126	0.270	0.854	0.786	0.818	0.796	0.804	1	0	6

*Notes: 3PL = 1 for the 3PL/GPCM, 0 otherwise; Fall = 1 for a Fall Test Administration, 0 otherwise; Age is the scale age (in years).

Appendix B: Plots of Within State Regressions

Figure 1B. Wainer Plots - Within State Effect Size Regressions on Time for ELA.

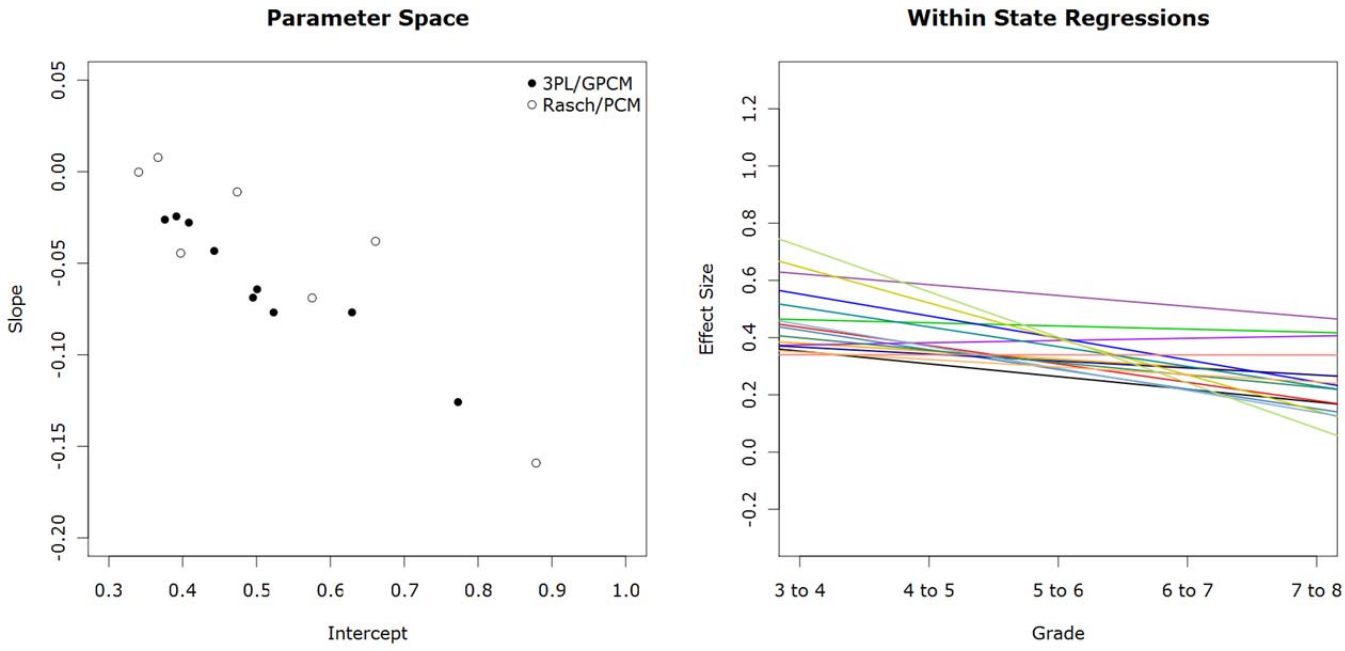
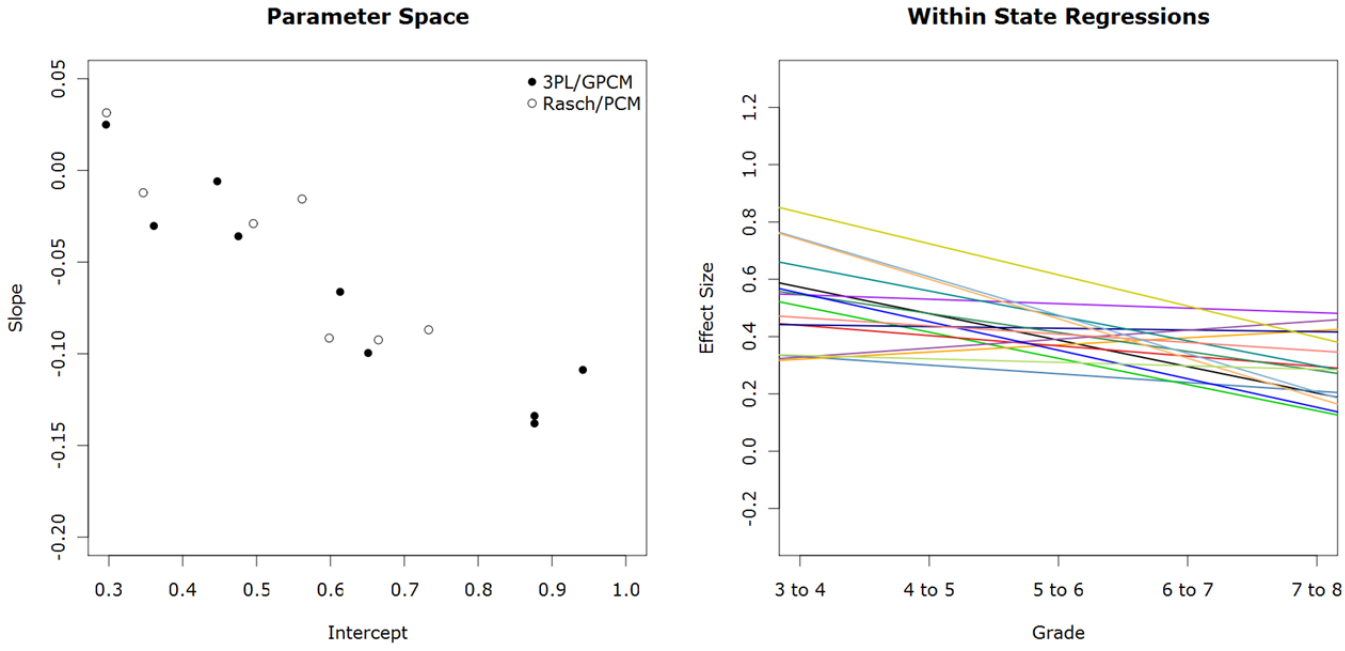


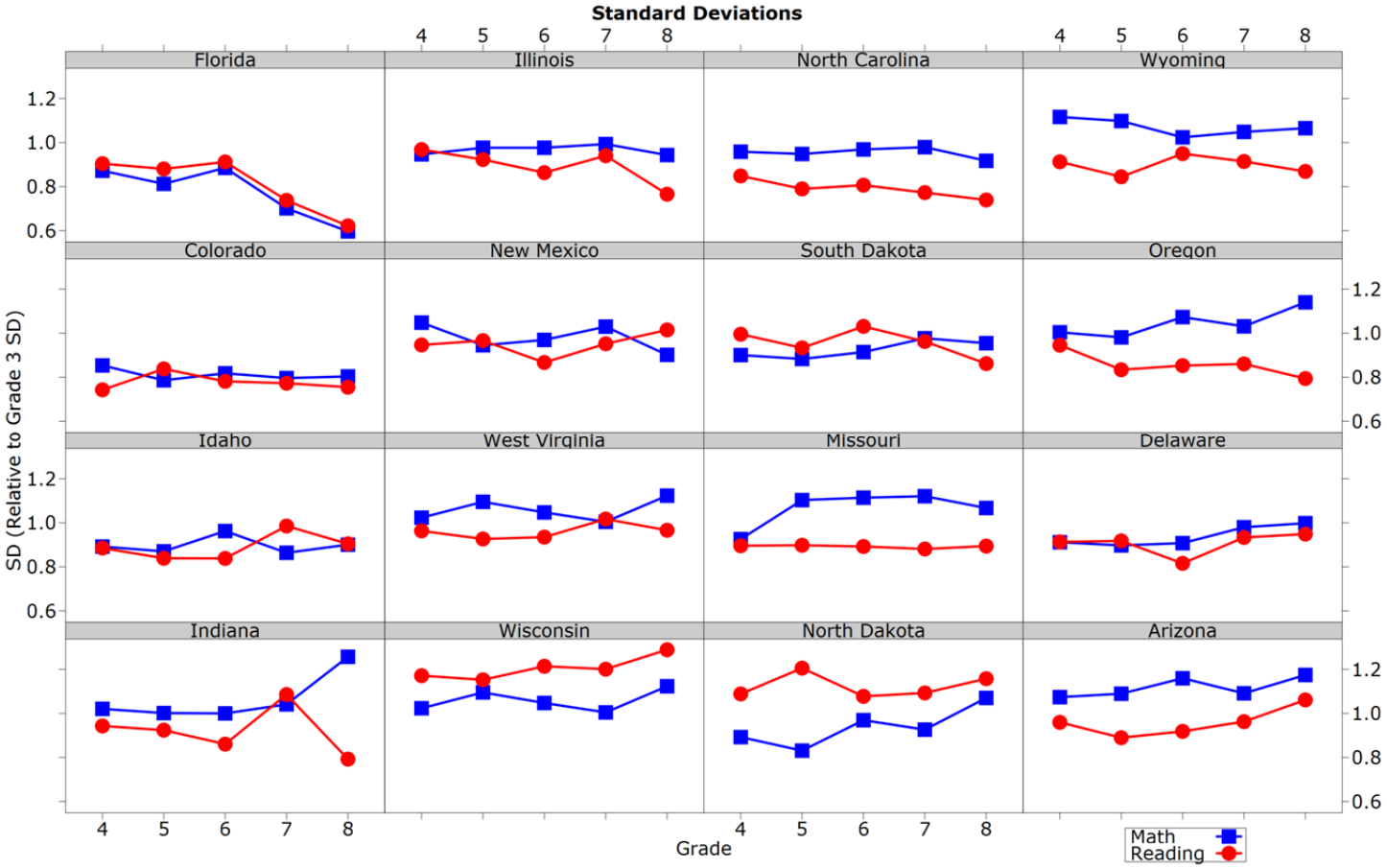
Figure 2B. Wainer Plots - Within State Effect Size Regressions on Time for Math (Note, Wyoming and West Virginia's slopes have been jittered slightly to prevent overplotting).



Note: These plots were inspired by a suggestion from Howard Wainer, so we call them "Wainer Plots." On the left hand side of the panel we plot the results from within state regression of time (1 to 5) on grade to grade effect sizes. The x-axis represents the estimated intercept, and the y-axis represents the estimated slope. The negative slope indicates that states with higher intercepts tend to have stronger growth deceleration. The solid dots represent states using the 3PL/GPCM to calibrate their vertical scale; the empty dots represent states that used the Rasch Model/Partial Credit Model. There appears to be a slightly steeper negative relationship for the 3PL/GPCM. The right hand side of the panel displays the within state growth trend lines.

Appendix C: Scale Score Standard Deviations (Relative to Grade 3 SD)

Figure 1C. Standard deviations in Math and Reading, Ordered by Evidence of Scale Shrinkage, From Most to Least (in Terms of Magnitude of Within State Regression of SD on Time).



Citation:

Dadey, N.& Briggs, D. C. (2012). A Meta-Analysis of Growth Trends from Vertically Scaled Assessments. *Practical Assessment, Research & Evaluation*, 17(14). Available online: http://pareonline.net/getvn.asp?v=17&n=14

Acknowledgment:

This research was supported by grants from the Spencer Foundation and the Carnegie Corporation.

Authors:

Nathan Dadey (nathan.dadey [at] colorado.edu)
Derek C. Briggs (Derek.Briggs [at] colorado.edu]
School of Education, Room 211
University of Colorado at Boulder
249 UCB
Boulder, CO 80309