Not Where You Start, But How Much You Grow: An Addendum to the Coleman Report

Allison Atteberry, PhD*
CU Boulder School of Education
249 UCB
Boulder, CO 80309
allison.atteberry@colorado.edu

Andrew McEachin, PhD
RAND Corporation
1776 Main Street
Santa Monica, CA 90407
mceachin@rand.org

Abstract

The Equality of Educational Opportunity Study (1966)—the Coleman Report—lodged an key takeaway in the minds of educators, researchers, and parents: Schools do not strongly shape students' achievement outcomes. This finding has been influential to the field, however Coleman himself suggested that—had longitudinal data been available to him—decomposing the variance in students' learning *rates*, rather than their *levels*, of achievement would have provided a clearer insight into school effects. Inspired by an intriguing finding from an earlier study conducted in 1988 by Bryk and Raudenbush, we take up Coleman's suggestion using data provided by the Northwest Evaluation Association, which has administered over 200 million vertically-scaled assessments across all 50 states since 2008. We replicate Bryk and Raudenbush's surprising finding that most of the variation in student learning rates lies between, rather than within, schools. For students moving from grades 1 through 5, we find 74% (math) to 81% (ELA) of the variance in math learning rates is at the school level. These results are intriguing since they call into question one of the dominant narratives about the extent to which schools shape students' achievement, however more research is needed. Our goal in this policy brief is to invite other scholars to conduct similar analyses in other data contexts. We delineate four key dimensions along which results need to be further probed, first and foremost with an eye toward the role of test score scaling practices, which may be of central importance.

Updated citation:
Atteberry, A., & McEachin, A. (in press, June 2020). Not Where You Start, But How Much You Grow: An Addendum to the Coleman Report. *Educational Researcher*.

ALLISON ATTEBERRY, PhD, is an assistant professor of research and evaluation methodology at the University of Colorado-Boulder School of Education. Her work addresses persistent patterns of inequality in key educational pivot points, including early childhood education, access to effective teaching, and summer learning loss.

ANDREW McEACHIN, PhD, is a policy researcher in the Economics, Statistics, and Sociology Department at the RAND Corporation and a professor at the Pardee RAND Graduate School. His research focuses on the determinants of persistent achievement gaps, as well as evaluating the effect of popular responses by policymakers and educators to reduce these gaps.

The project was supported in part by the Kingsbury Data Award funded by the Kingsbury Center at the NWEA, as well as the Smith Richardson Foundation. All errors are solely attributable to the authors.

*=Corresponding author

The Equality of Educational Opportunity Study (1966)—the Coleman Report—still undergirds the long-held understanding that schools play a limited role in shaping students' outcomes. Because Coleman found that only 10-20% of the variation in student achievement scores lies among schools, he concluded that schools were simply not a powerful lever to affect students' achievement relative to non-school factors. This "schools don't matter" narrative has long been taken up in a number of influential ways from both conservative and liberal perspectives (for a synthesis, see Hutt, 2017; Jencks, 1969).[1] Though critiques have been written[2] of the Coleman Report, this particular finding—the low proportion of variance in student achievement between (versus within) schools—has been found many times over (see Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007 for a compendium of intraclass correlations).

The Coleman Report has led to a lasting pessimism about investing in school features, because the results suggest that, by the time students first arrive to school, their achievement is largely set. Ravitch (1981) captures this sentiment well when reflecting on the Report 15 years after its release:

> *It is impossible to assess the damage done to the self-esteem of the education profession and the consequent demoralization of the very teachers dedicated enough to inform themselves about educational research. Whether students did well or poorly in schools seemed determined…little, if at all, by anything that teachers and schools did. (pg.719)*

Given subsequent research that highlighted how the Coleman Report misses pathways through which schools affect students' achievement, this pessimism was perhaps not entirely warranted.[3] Nonetheless, the Coleman Report's findings continue to influence the field: A quick Google Scholar search returns over 1,600 articles that mention the "Coleman Report" since 2017 alone.

This raises perplexing questions: If it is *really* true that schools matter very little, how do we continue to justify research and investment on school-level programs, policies, and practices? How do we reconcile the seeming contradiction that school settings are deeply unequal and yet achievement is only weakly shaped by those inequalities? Why has this takeaway from the Coleman Report remained so pervasive, even in the face of important critiques and developments? While, in

fact, the concept of school-level value added measures (Reardon & Raudenbush, 2009) offers a way to reconcile this seeming contradiction, that connection is indirect and has not been explicitly stated.

A Different Approach to Estimating School Effects

Coleman himself suggested that a better approach to capturing school effects would be to partition variation in learning *rates*—rather than *levels*—within and between schools.[4] He writes:

> *"Had a number of years been available for this survey, a quite different way of assessing effects of school characteristics would have been possible; that is, examination of the educational growth over a period of time of children in schools…This is an alternative and in some ways preferable method…Thus, the present analysis should be complemented by others that explore changes in achievement over a large span of time" (p. 292).*

Students and their entering levels of achievement are allocated to schools in ways that are outside schools' control. It makes sense, then, to not think of schools as influencing how students perform in a given grade, but rather on how quickly they grow[5] over time. In 1988, Bryk and Raudenbush used Sustaining Effects Study data to implement Coleman's recommendation. They adopted a 3-level multilevel model[6] of vertically-scaled test scores and partitioned variance within and between schools for both achievement *status* (akin to Coleman) and achievement *rates* from grades 1 to 3.

With regard to status, they first replicated Coleman's finding; only 14% of the variance in math achievement lies between schools (31% in ELA). However, their decomposition of the variance in learning *rates* showed a very different pattern. They write,

> *"…the results for learning rates, particularly in mathematics, are startling indeed. Over 80% of the variance in mathematics learning is between schools! These results constitute powerful evidence of school effects that have gone undetected in past research" (pg. 96).*

Using Coleman's own analytic recommendation, Bryk and Raudenbush's results contradicted one of the key takeaways from the Coleman Report that schools must not have much impact on achievement. These findings are striking. However, they should also be revisited, given that they were based on a small number of schools (86) with an average of only 7 students sampled per school, or also could have been idiosyncratic to that particular test score scaling.

Current Analysis

We replicate and expand upon the Bryk and Raudenbush (1988) analysis, using a dataset[7] provided by the Northwest Evaluation Association (NWEA) which has administered over 200 million assessments to nearly 18 million students in 7,500 districts across all 50 states in a very recent time period (2008 through 2016), wherein the average number of students per school is 118. Importantly, NWEA's MAP test is designed so that its scores can be expressed on a vertical scale (which NWEA calls the RIT[8] scale), and with the intent that it can be used to support equal-interval interpretations.

Before proceeding, we take a brief detour on achievement test score scaling. In theory, a vertical scale enables comparisons of student learning across grades, while the equal-interval property of the scale ensures that a unit increase in a student's score represents the same learning gain across the entire score distribution. Since we will be attempting to trace learning growth across grades, vertical scaling would be desirable, and interval scaling is essential for *any* test score comparison. However, there are many different ways of designing and calibrating a vertical scale, and there is little consensus with regard to the best methods for evaluating these properties (Briggs, 2013; Briggs & Dadey, 2015; Briggs & Domingue, 2013; Briggs & Weeks, 2009). While some scaling practices are clearly not well-suited for certain purposes, there will never be a way to identify *the* single, correct scale. Because we do not have access to NWEA item-level responses, subsequent research is needed to probe sensitivity of our results to scale development practices.

We estimate a nested model with 3 levels: (up to) 10 test scores per student from grades 1 to 5 in fall and spring (level one), students within schools (level two), and across schools (level three). Just as for Bryk and Raudenbush, this model allows us to calculate the percent of variance that lies across schools (intraclass correlations, or ICCs), both for achievement levels (like Coleman) and achievement growth (like Bryk and Raudenbush). In our primary specifications,[9] we use NWEA's RIT scores at the outcome of interest. We will see, however, that results differ when we standardize

within subject-grade-year. We also present results using a linear growth trajectory (M1) and a quadratic growth trajectory (M2). To consider the choice of functional form for growth, see Figure 1 to examine the observed RIT reading scores for 50 randomly sampled students, and see Appendix A for a full discussion of functional forms considered and robustness of results to these choices.[10]

Results

We report our ICCs alongside the relevant results from both the Coleman Report and the Bryk and Raudenbush (1988) study in Table 1 (see Appendix C for complete model results, including estimated fixed effects, variance components, 95% plausible value ranges, estimated total gains throughout the grade panel, and reliabilities). Our results replicate both the original Coleman Report ICCs, as well as those from Bryk and Raudenbush's study. In the left column of Table 1 (variation in achievement *levels*), our results are largely consistent with the Coleman Report: Only 23.7% of the variance in math achievement levels lies between schools (21.4% in ELA). When it comes to *linear growth rates* (middle panel of Table 1), our results are quite similar to the surprising findings from Bryk and Raudenbush: The majority—73.5%—of the variance in math learning rates lies between schools (80.7% in ELA). This is similar to Bryk and Raudenbush's estimate for math of 82.6% (though for ELA they find a somewhat smaller, but still sizeable, 42.5%).

We also extend this analysis to grades 6 – 8, shown in row (1b) of Table 1. Again, we find that the majority of the variation in achievement *levels* is within schools, but the majority of the variation in learning *rates* is between schools (64.9% in math, 79.5% in ELA). When using a quadratic growth model (M2) instead of a linear function, we again find that most of the variation in both instantaneous learning rates at the midpoint and acceleration of learning lies at the school level (row (2a) and (2b) of Table 1).[11] Moreover, we find this pattern holds when we consider other model specifications or analytic samples (see robustness checks[12] in Appendix C).

To visually illustrate this finding, Figure 2 presents boxplots of estimated[13] achievement levels for the students nested within 20 randomly sampled schools (left), alongside a boxplot of schools' mean achievement levels across all schools in the sample (right). Here one sees the classic Coleman pattern: Any given school seems to have a wide vertical distribution of achievement scores among its students, while the mean achievement levels (red dots) across schools are not very different from one another. In the lower panel of Figure 2, we make the same visualization for achievement *rates* and see the opposite pattern: Students in a given school exhibit similar learning rates to one another, and average learning rates vary considerably[14] from school to school.

Finally, we simulate a more common policy context in which only spring test scores are available, and scores are not vertically-scaled. When we remove fall scores and standardize RIT scores within subject-grade-year, the findings change. In these models, less than half of the variation in learning rates lies between schools (37% and 44%, see row (D) of Appendix Table C3). Of course, grade-standardized test scores are not designed to capture growth over time, so this is perhaps not entirely surprising. These results, which are consistent with results from at least one other study,[15] suggest that the practice of standardizing within subject-grade-year would likely mask our overall finding, which could explain why this result is not commonly reported. It also underscores the centrality of achievement scaling practices when examining variation in student learning trajectories.

Replication Efforts

We think the current results are intriguing since they call into question one of the dominant narratives about whether schools shape students' achievement, however more research is needed to understand them. Our results do not definitively establish that schools strongly shape growth rates, but rather they raise the possibility that our conventional wisdom needs to be revisited. Our goal in this policy brief is to invite other scholars to conduct similar analyses in other data contexts, particularly where item-level data is available. Given that the current multilevel model is not

particularly complex or novel, we anticipate that others either can or already have conducted analyses that would yield between-school ICCs in student growth rates. For instance, we are aware of at least three studies[16] that have implemented a similar multilevel model, and—while the published studies do not report the unconditional variances needed to calculate the relevant ICCs—the authors might be able to examine them retrospectively.

It is possible that our results are simply an artifact of NWEA's scaling practices. To partially address this concern, we run a similar analysis using the Early Childhood Longitudinal Study Kindergarten Cohort 1998-99 (ECLS-K:99) public use dataset. Using a range of approaches to define the analytic sample and different growth functions,[17] preliminary results are consistent with NWEA patterns: The average ICC for achievement levels was 29%, while the average ICC for linear growth rates was 59%, and the average ICC for rates of acceleration was 71% across specifications.

Researchers should probe our results along four key dimensions: First and foremost, for those who have access to item-level response data, sensitivity to scaling practices may be of paramount importance. Briggs and Domingue (2013) indeed find that their estimates of within-*district* (and residual) variance in student growth rates is nearly three times larger when using a z-score scale than when using a vertical scale, which means the vertical scale would produce larger ICCs. The degree of sensitivity to score scaling properties, they show, is a function of differential scale expansion or compression across grades. On this point, von Hippel, Workman, and Downey (2018) show that, in ECLS-K data, inferences about whether achievement disparities grow as students move through school depends on whether one uses theta scores (in which variance is constant across grades K to 2) or IRT-based scale scores (in which variance notably increases across grades). Together, this research suggests that measurement properties of achievement scores will be central to this story.

Second, we hypothesize that estimated ICCs in growth rates could be sensitive to a dataset's within- and across-school sampling frame. Most nationally-representative datasets constructed by

NCES (e.g., ECLS-K) sample a small number of students per school, which may hinder estimating reliable within-school variance parameters. Third, it will be important to continue to explore whether results are sensitive to the growth function specified and/or inclusion criteria for students in the analytic sample. Fourth, findings may in turn differ depending on the length of the longitudinal panel at hand, as well as the grade range of study.

Takeaways

The narrative that schools play a relatively small part in shaping students' achievement, which has roots in the Coleman Report, remains a powerful demotivator in research about the U.S. public education system. That finding has also managed to transcend the divide between research and public discourse, affecting how parents, the media, and policy-makers think about the value of public schooling. Yet if our results were to hold, it suggests a need to update this conventional wisdom. While it is true that students enter kindergarten with a wide range of school readiness, students' school age outcomes might not be as fixed as is widely believed. Our estimates suggest that students appear to vary more from school-to-school in terms of how fast they grow. While schools may not have much control over who enrolls, they may have an impact on how fast students' achievement improves.

The Coleman Report is often required reading for new graduate students in schools of education; as it should be, given its undeniable impact on the history of U.S. public education. However, when students learn about Coleman et al.'s finding that only 10 to 20% of the variation in student achievement exists at the school level, we should also point to subsequent research about the many school factors that do affect students' outcomes. Graduate students often walk away from the Coleman Report with a forever-damaged perception of schools as a limited lever for change. The simple analyses presented here provide one way to understand how this Coleman Report ICC statistic (and all those that followed) could be correct, but at the same time may only partially capture the role of schools in shaping students' outcomes.

Finally, while conducting these analyses, schools across the world have been shut down by the COVID-19 pandemic. If our results hold, it suggests that schools will likely play a crucial role in students' recovery from widespread school closures and other related social, economic, and health care crisis from the pandemic. Given the uncertainties of what school will look like in the upcoming school years, we will need revisit these questions using data from the post-COVID period.

Tables and Figures

Table 1. ICCs across Studies: Proportion of Variance Between Schools for Achievement Levels, Linear Growth, and Growth Curve/Acceleration

	Achievement Level: % Between Sch's		Learning Rate: % Between Sch's		Acceleration: % Between Sch's	
	Math	ELA	Math	ELA	Math	ELA
Coleman Report						
Grade 3	20.9%	20.7%	n/a	n/a		
Grade 6	15.7%	16.3%	n/a	n/a		
Grade 9	9.0%	9.0%	n/a	n/a		
Bryk & Raudenbush						
Grades 1 - 3	14.4%	31.4%	82.6%	42.5%		
Current Study, NWEA's RIT Scale						
(1a) Grades 1 - 5	23.2%	21.1%	75.4%	80.3%		
(N schools)	*(5,545)*	*(5,436)*	*(5,545)*	*(5,436)*		
(1b) Grades 6 - 8	23.5%	23.2%	64.2%	78.3%		
(N schools)	*(3,808)*	*(3,790)*	*(3,808)*	*(3,790)*		
(2a) Grades 1 - 5	23.9%	20.3%	73.0%	68.7%	65.8%	48.1%
(N schools)	*(5,545)*	*(5,436)*	*(5,545)*	*(5,436)*	*(5,545)*	*(5,436)*
(2b) Grades 6 - 8	23.8%	23.6%	68.5%	73.1%	75.0%	83.7%
(N schools)	*(3,808)*	*(3,790)*	*(3,808)*	*(3,790)*	*(3,808)*	*(3,790)*

FN: Percentages can be interpreted as the percent of total variance in achievement levels, rates, or acceleration that lies between schools (whereas the remainder lies within schools). Results for Coleman Report come from Table 3.21.5, for Mathematics Achievement and Reading Comprehension. Results for Bryk & Raudenbush (1988) come from Table 6 on page 95. The results shown in Table 1 from the current study use NWEA's achievement scores expressed in the RIT scale units, which is intended to achieve vertical and interval scaling properties (as discussed, these properties are difficult to achieve and hard to verify).

Figure 1. Reading Scores for 50 Randomly Sampled Students

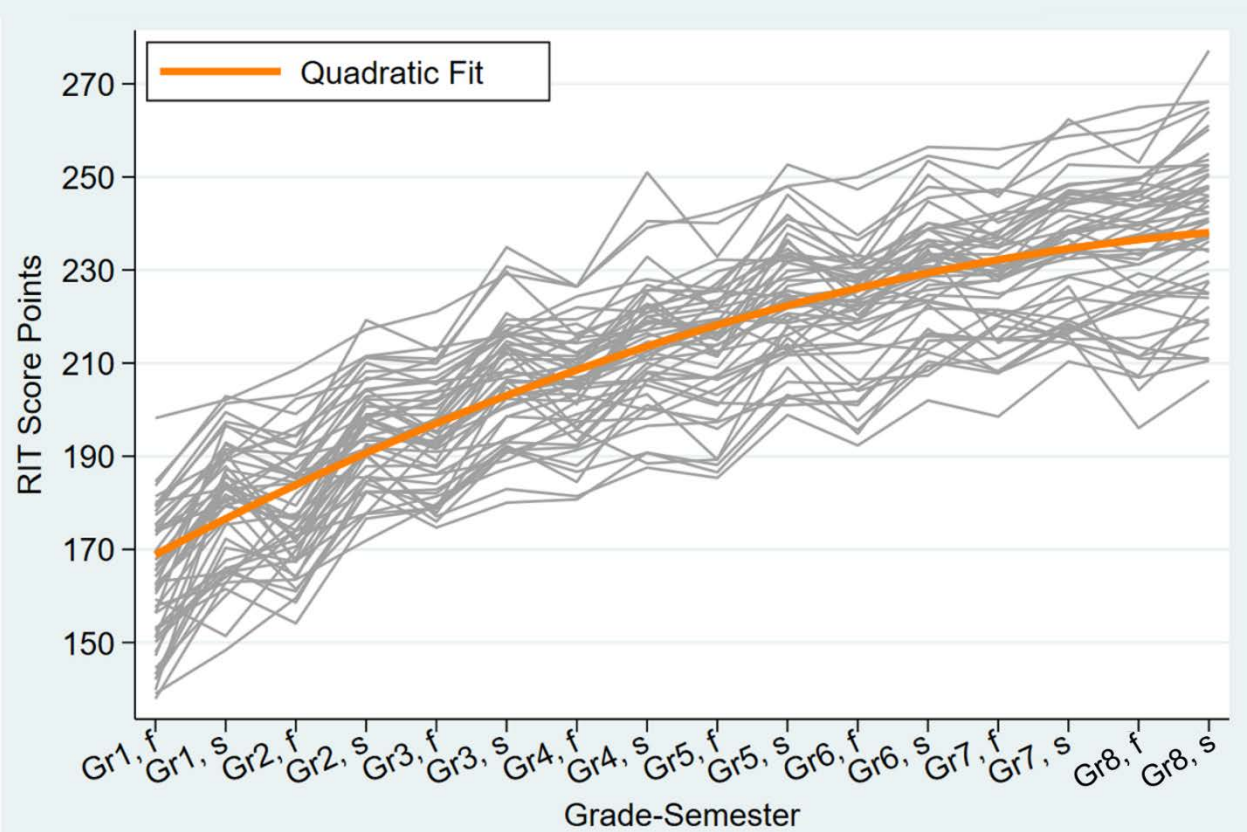
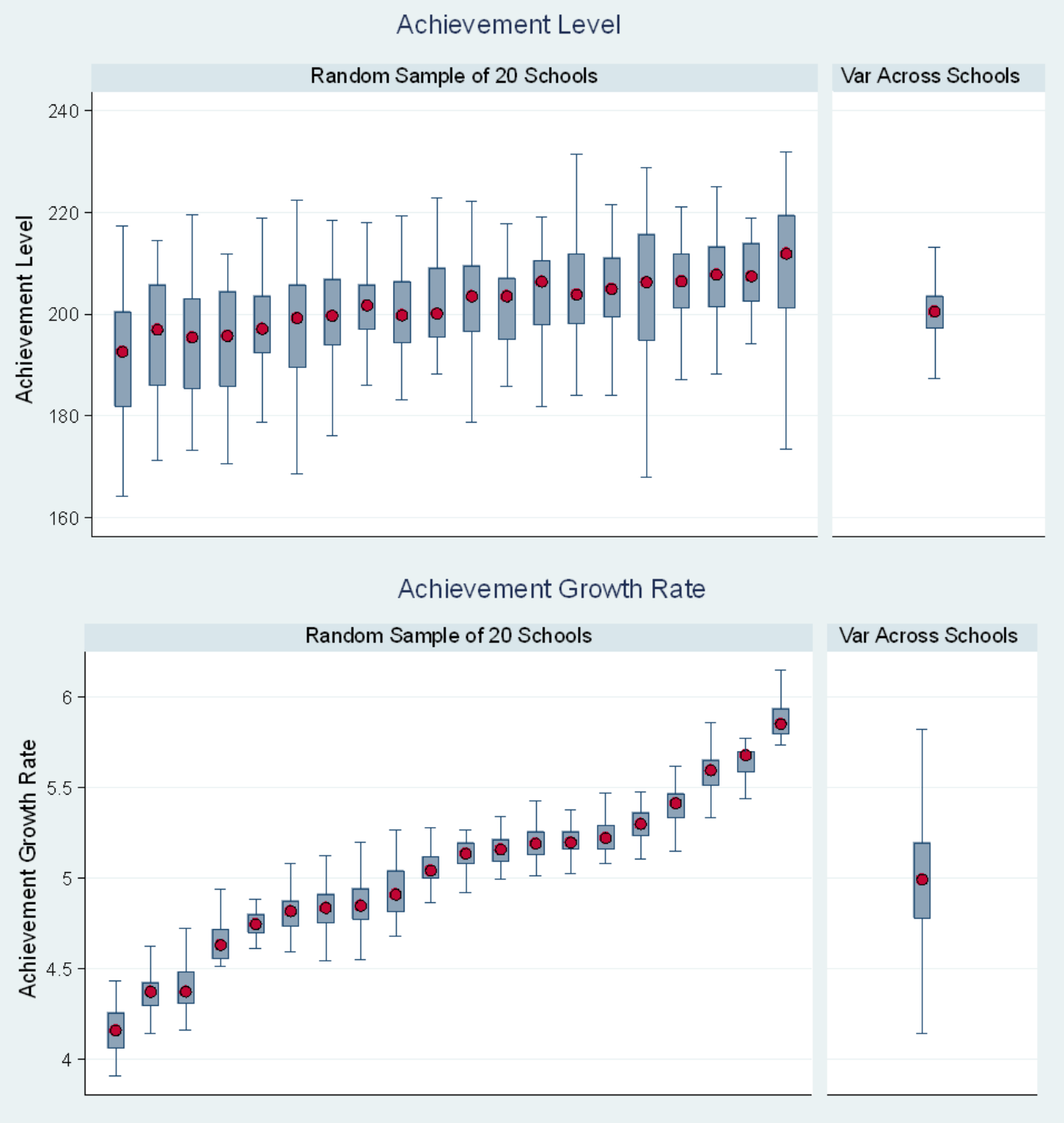


Figure 2. Boxplots of Empirical Bayes Estimates, both among Students within a Random Sample of 20 Schools (left), and across All Schools (right). Achievement Levels (upper) vs. Achievement Rates (lower)



FN: Estimates are reported in NWEA's original RIT score metric. For context, we provide a select number of mean achievement status and growth norms produced by NWEA (for all norms, see Thum & Hauser, 2015): For reading status norms, the mean is 178 in 1st grade (student SD=14.5) and 220 in 8th grade (SD=15.7). For math status, the mean is 181 in 1st grade (SD=13.6) and 231 in 8th grade (SD=19.1). For reading annual growth norms (last spring to current spring), the mean is 18.8 in 1st grade and 3.1 in 8th grade. For math annual growth norms, the mean is 21.2 in 1st grade and 4.0 in 8th grade.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1).

Alexander, K. L., & Entwisle, D. R. (1996). Early schooling and educational inequality: Socioeconomic disparities in children's learning. *James S. Coleman, 4*, 63.

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis, 23*(2), 171.

Barr, R., Dreeben, R., & Wiratchai, N. (1983). *How schools work*: University of Chicago Press.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Borman, G., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record, 112*(5), 1201-1246.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement, 50*(2), 204-226.

Briggs, D. C., & Dadey, N. (2015). Making sense of common test items that do not get easier over time: Implications for vertical scale designs. *Educational Assessment, 20*(1), 1-22.

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.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model. *American Journal of Education, November*, 65-108.

Cain, G. G., & Watts, H. W. (1970). Problems in making policy inferences from the Coleman Report. *American Sociological Review*, 228-242.

Carbonaro, W. J., & Gamoran, A. (2002). The production of achievement inequality in high school English. *American Educational Research Journal, 39*(4), 801-827.

Carver, R. P. (1975). The Coleman Report: Using inappropriately designed achievement tests. *American Educational Research Journal, 12*(1), 77-86.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. National Bureau of Economic Research.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, F., Mood, A., & Weinfeld, F. (1966). Equality of educational opportunity study. *Washington, DC: United States Department of Health, Education, and Welfare*.

Domingue, B., Thomas, S., Circi, R., & Camilli, G. (2011). Comment on "Schools and Inequality: A Multilevel Analysis of Coleman's Equality of Educational Opportunity Data". *Teachers College Record, https://www.tcrecord.org ID Number: 16544*.

Downey, D. B., & Condron, D. J. (2016). Fifty years since the Coleman Report: Rethinking the relationship between schools and inequality. *Sociology of Education, 89*(3), 207-220.

Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review, 69*(5), 613-635.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational researcher, 18*(4), 45-62.

Hanushek, E. A., & Kain, J. F. (1972). On the value of equality of educational opportunity as a guide to public policy. *On equality of educational opportunity*, 116-145.

Heckman, J. J., & Neal, D. (1996). Coleman's contributions to education: theory, research styles and empirical research. In J. Clark (Ed.), *James S. Coleman* (pp. 88-102). London and New York: Routledge Falmer.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Money does matter somewhere: A reply to Hanushek. *Educational researcher, 23*(4), 9-10.

Hutt, E. L. (2017). "Seeing Like a State" in the Postwar Era: The Coleman Report, Longitudinal Datasets, and the Measurement of Human Capital. *History of Education Quarterly, 57*(4), 615-625.

Jencks, C. (1969). *A reappraisal of the most controversial educational document of our time*: Center for Educational Policy Research, Harvard Graduate School of Education.

Lee, V. E., Smith, J. B., & Croninger, R. G. (1997). How high school organization influences the equitable distribution of learning in mathematics and science. *Sociology of Education, 70*, 128-150.

NWEA. (2011). *Technical Manual For Measures of Academic Progress® (MAP®) and Measures of Academic Progress for Primary Grades (MPG)*. Retrieved from https://www.richland2.org/RichlandDistrict/media/Richland-District/AdvancED/Standard%205/5.1/5-1-NWEA-Technical-Manual-for-MAP-and-MPG.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1): Sage Publications, Inc.

Reardon, S. F., & Raudenbush, S. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519.

Rumberger, R. W., & Palardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record, 107*(9), 1999.

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. Retrieved from https://www.nwea.org/content/uploads/2018/01/2015-MAP-Norms-for-Student-and-School-Achievement-Status-and-Growth.pdf

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Najarian, M., & Hausken, E. G. (2009). *Combined User's Manual for the ECLS-K Eighth-Grade and K–8 Full Sample Data Files and Electronic Codebooks* Retrieved from Washington, DC: https://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part1.pdf

Towers, J. M. (1992). Twenty-five Years after the Coleman Report: What Should We Have Learned? *The Clearing House, 65*(3), 138-140.

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are schools the great equalizer?". *Sociology of Education, 91*(4), 323-357.

Walberg, H. J. (1984). Families as partners in educational productivity. *Phi Delta Kappan, 65*(6), 397-400.

Whitehurst, G. J., & Croft, M. (2010). *The Harlem Children's Zone, Promise Neighborhoods, and the broader, bolder approach to education*: Brown Center on Education Policy, The Brookings Institution.

Online Appendix A: Decomposing Variance in Learning Rates Within and Between Schools: Current Study & Prior Research

Current Paper Primary Specification

The model presented below is quite similar to the unconditional model implemented by Bryk and Raudenbush (1988) in their study entitled, "Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model". As the title suggests, they use a three-level random effects framework to model (up to) five vertically-scaled test scores administered to students (N= 618) between grade 1 and 3 as a linear function of time, nested within students at level two, who are in turn nested in schools at level three.

The primary model specification used in the current paper does not differ in structure, however we run these subject-specific models separately for students with RIT scores in grades 1 through 5 and grades 6 through 8. For the sake of explication, we present a model below for projected RIT math scores between first through fifth grade. These ten repeated observations (level one) are nested within students (level two), and schools (level three):

Level One: Repeated observations across grade-semesters (t), nested in Students (i)

$$Score_{tij} = \pi_{0ij} + \pi_{1ij}(GradeSem_{tij} - M) + e_{tij} \quad \text{, where } e_{tij} \sim N_{iid}(0, \sigma)$$

Level Two: Students (i), nested within Schools (j)

$$\pi_{0ij} = \beta_{00j} + r_{0ij} \quad \text{, where } r_{0ij} \sim N_{iid}(0, \tau^{\pi}_{0,0})$$

$$\pi_{1ij} = \beta_{10j} + r_{1ij} \quad \text{, where } r_{1ij} \sim N_{iid}(0, \tau^{\pi}_{1,1})$$

Level Three: Schools (j) across the U.S.

$$\beta_{00j} = \gamma_{000} + u_{00j} \quad \text{, where } u_{00j} \sim N_{iid}(0, \tau^{\beta}_{0,0})$$

$$\beta_{10j} = \gamma_{100} + u_{10j} \quad \text{, where } u_{10j} \sim N_{iid}(0, \tau^{\beta}_{1,1})$$

The outcome of interest, $Score_{tij}$, is a projected math RIT score in semester t (for example, fall of first grade) for student i in school j. The level one predictor variable, $GradeSem_{tij} - M$, is a linear time variable centered in at a midpoint in the grade range, spring of grade 3 (i.e., -1 = fall of 3rd, 0 = spring of 3rd, 1 = fall of 4th, 2 = spring of 5th, etc.). As a result of this coding scheme, the level one parameter π_{0ij} captures student i's in school j's math RIT score in spring of grade 3, and π_{1ij} would capture the linear change in RIT scores for every passing period for student i in school j. We analyze elementary grades 1 through 5 separately from grades 6 through 8, and we nest each student within their modal school during the grade period. These parameters are allowed to vary at levels two (across students within the same school) and three (across schools). Note that students' initial status and slope parameters cannot vary at level one, because each student only has one estimate for each of these phenomena. The level two and three variance parameters here are of greatest interest. For instance, $\tau^{\pi}_{0,0}$ is the estimated variance among students within the same school in grade 1 achievement, and $\tau^{\beta}_{0,0}$ is the estimated variance across schools in terms of mean grade 1 achievement. Likewise, $\tau^{\pi}_{1,1}$ is the estimated variance among students within the same school in their linear learning rates between grades 1 and 5, and $\tau^{\beta}_{1,1}$is the estimated variance across schools in schools' mean learning rates.

To replicate the findings of both Coleman and Bryk and Raudenbush, we calculate the proportion of variance in achievement that lies between schools as $\tau^{\beta}_{0,0}$ divided by the sum of $\tau^{\pi}_{0,0}$ and $\tau^{\beta}_{0,0}$. Like Bryk and Raudenbush, we also partition the variance in students' learning rates to calculate the proportion of variance between schools as $\tau^{\beta}_{1,1}$ divided by the sum of $\tau^{\pi}_{1,1}$ and $\tau^{\beta}_{1,1}$.

Choice of Functional Form

In general, we follow growth modeling practices like those described by Raudenbush and Bryk (2002) for applications in the study of individual change. A visual inspection of Figure 1 suggests that a quadratic growth model could be a good fit for the data. We therefore also fit a model with a quadratic—rather than a linear—growth function at level one:

$$Score_{tij} = \pi_{0ij} + \pi_{1ij}\left(GradeSem_{tij} - M\right) + \pi_{2ij}\left(GradeSem_{tij} - M\right)^2 + e_{tij} \ , \qquad e_{tij} \sim N_{iid}(0, \sigma)$$

Because of the centering at the midpoint, the level one parameter π_{1ij} now captures the instantaneous learning rate in RIT scores in spring of grade 3 for student i in school j. The level one coefficient on the squared grade term, π_{2ij} captures a given student's acceleration/deceleration in their learning rate during the panel. Because some theories of learning suggest that learning may indeed slow as students move through school-age years, a quadratic growth curve may be appropriate. Moreover, other researchers who have analyzed repeated fall and spring achievement data across grades, nested within students have also adopted quadratic growth models: For instance, Alexander, Entwisle, and Olson (2001) present results from a quadratic model with seasonal achievement outcomes from grades 1–5 and find a statistically significant, negative average acceleration parameter, as do we.

As one additional investigation into sensitivity to the choice of growth function, we took a 10% random sample of the data and ran a log-linear function, which can be useful for processes that exhibit diminishing returns. The beta-coefficient on the grade-semester variable, is still a measure of change-in-scores (a one-semester increase in time is associated with a 100×beta% change in RIT scale score points). We find a pattern consistent with our primary results: For instance, in middle school reading scores, the percent of variance between schools (the ICC) is

18% for the intercept but 66% for the coefficient on grade-semester. Across both elementary and middle grades and math and reading, the ICC for intercepts was less than 20%, while the ICC for the coefficient on grade-semester was above 60%.

Other Studies that Use Similar Models

We have found one other paper by Rumberger and Palardy (2005) that *also* uses a three-level linear growth model and documents the proportion of variation in learning rates between schools (see their Appendix Table 3). Rumberger and Palardy (2005) use data from the National Education Longitudinal Study of 1988 (NELS) but find the between-school variance in growth rates is closer to 20-30% (21% for math, 20% for reading, 34% for science, and a more unusual 60% for history). However, the authors report using a t-scale achievement score with a "mean of 50, SD of 10" (pg. 2037). Vertical scaling may be essential to an analysis of growth; when we standardize RIT scores, we also no longer find large between-school variances in learning rates. Their study differs from the current study in a number of important ways: NELS largely focuses on changes in scores in high school, whereas the current study focuses on grades 1 through 8. Given that growth curves appear to decelerate across grades, it is possible that findings would genuinely be different in high school. It is also worth noting that students in NELS have at most 3 scores (spring of 8th, 10th, 12th) from which to estimate a growth trajectory. In the current study, students have between 6 and 10 test score observations, which better supports growth modeling. Finally, most of the NCES datasets have relatively small within-school samples (e.g., on average about 16 students per school). We suspect this could hinder the ability to estimate within-school variances reliably.

In addition to the paper described above, we are aware of at least three other studies that have implemented a similar multilevel model, and—while the published studies do not include the

unconditional variances needed to calculate the relevant ICCs—the authors might be able to examine them retrospectively (Briggs & Domingue, 2013; Carbonaro & Gamoran, 2002; Lee et al., 1997; Rumberger & Palardy, 2005). For instance, Lee et al. (1997) analyze vertically-scaled achievement data from NELS using a 3-level piecewise growth model (3 test scores across grades 8, 10 12, nested in students, nested in schools), however their focus is not on the random effects estimated by their models and they therefore do not report estimated variances within and between schools in growth rates. Carbonaro and Gamoran (2002) conduct a similar analysis with NELS data, instead using a simpler, linear growth model. However, because they do not report results from a totally unconditional model, they only report *residual* variances in growth rates (see their Table 3).

Briggs and Domingue (2013) conduct a particularly relevant analysis of achievement data in a medium sized state, in which up to 5 test scores observations between grades 5 and 9 are nested within students who are in turn nested in districts (rather than schools). Their study is focused on highlighting the potential impact of scaling decisions on estimating both school value-added measures (VAMs) and student growth trajectories, and they report estimated standard deviations in growth rates both among students in the same district (L2) and across districts (L3). Importantly, they compare models that use three different scalings of the achievement outcome data: (a) a z-score in which a summed score of number of items answered correctly is standardized within grade to have mean 0 and SD 1, (b) a theta score estimate of ability generated from applying an item response theory (IRT) approach with a three-parameter logistic model (3PLM) with maximum likelihood estimation, and (c) a vertical scale in which the ability estimates are linked together across grades using common items. Given their focus, they also do not report estimated variances from a totally unconditional model, and we therefore cannot partition growth within and

between districts. However, they do report estimates of residual variance in growth rates (within and between districts) for the three different approaches to scaling the outcome data. Their results suggest that scaling would have a large impact on estimated ICCs: The (residual) variance in growth rates at level two (within districts) is nearly three times smaller when using a vertical scale than when using a z-score scale. As in the current analysis, the across-cluster (here, district) ICC therefore would be much larger when using vertical scaling.

Appendix B: Full Data Description

NWEA Data

The data for this study is from the Northwest Evaluation Education Association's (NWEA) MAP assessment. The dataset contains math and reading scores based on a computer adaptive test designed to serve as part of a formative, benchmarking data system, purchased by about 7500 districts across all 50 states in the U.S. The MAP assessment is used as a supplementary tool to aid schools' in improving their instruction and meeting students' needs, not as the high-stakes test of record. Because the MAP assessment is intended to monitor students' progress throughout the school year, it is administered in both the fall and the spring. It is also administered in the winter by some districts, however the winter data is not included in the current dataset.

The MAP test is scored using a vertical and interval scale, which the NWEA calls the RIT scale. In theory, the vertical scale allows comparisons of student learning across grades and over time, while the interval scale ensures that a unit increase in a student's score represents the same learning gain across the entire distribution. The vertically-scaled nature of this outcome data is essential to our ability to examine differences in achievement disparities as students move through grade levels. However, it is worth noting that vertical scaling is difficult to achieve and hard to verify (Briggs, 2013; Briggs & Weeks, 2009). The most recent technical manual from NWEA indicates that the RIT scale is produced using a one-parameter logistic item response theory model (NWEA, 2011). The authors of that report describe a multi-stage process for generating these scales: (1) Identify the content boundaries for the measurement scale (2) Develop items that sample the content with a wide range of difficulty. (3) Identify samples of students appropriate for the items to be tested. (4) Administer the field test. (5) Estimate item difficulties. (6) Test items for model fit. (7) Test for dimensionality. (7) Apply Logit-to-RIT transformation. For more

information, refer to the technical manual. Because we do not have access to the item-level data required to explore different approaches to scaling, we cannot explore alternative approaches to producing achievement scores. Our findings regarding changes across grades rely on assuming that NWEA's vertical scale is valid.

The data for the current study comes from U.S school districts that administered the MAP assessment during the nine years between 2008 and 2016. Different districts opt to administer the MAP in different grades, however the full NWEA data includes 203,234,153 test scores for 17,955,222 million students who took a test between grades kindergarten and eleventh grade. The students in this dataset represent a large set of students, relative to the U.S. public school system. For instance, in 2012, this NWEA population a full ninth of the size of the K-12 public school student population. NWEA data is available in nearly 37 percent of all U.S. schools and in over half of all districts. The dataset includes student race and gender, their math and reading MAP scores, number of items attempted and correctly answered, duration of the test, grade of enrollment, and the date of test administration. The file does not include indicators for whether the student is an English Language Learner, belongs to the federal Free- and Reduced-Price Lunch program, or receives special education services. Since students do not take MAP tests *exactly* on the first and last day of school, we also create and use versions of the RIT scores that have been linearly projected to the first and last day of each school year. For a full description if this procedure, see (Authors, 2019). It is worth noting that the NWEA dataset is not a public-use dataset. However, the company currently has a number of research partnerships with universities (https://www.nwea.org/partnering-consulting/), and graduate students may apply to work with the data (https://www.nwea.org/the-kingsbury-research-award/).

Analytic Sample

For the current analysis, we first restrict the NWEA sample to the 89 percent of students who neither repeat or skip grades. In our preferred models, we also restrict the sample to the set of students who possess test scores for the full grade range included in the model. For instance, if we examine test score patterns between first through fifth grade in a given model, only students who have both fall and spring test scores in every grade between first and fifth grade (that is, a full vector of all ten reading test scores) will be included in the sample. While this is a restrictive sample limitation (e.g., 3% of students observed in grades 1 – 5 in the dataset meet this requirement), it ensures that our findings cannot be conflated with compositional changes from one time point to the next. In Online Appendix C, we replicate our primary findings on much less restrictive samples by only requiring that students have 75% or 50% of the 10 possible test scores between grades 1 - 5. In these larger samples, 13% of students observed in grades 1 – 5 in the dataset have at least 7 of those 10 scores, and 31% have at least 5 of the 10 scores. These samples have different advantages in terms of internal and external validity, however results are relatively consistent (see Appendix C).

Appendix C: Full Results, and ICCs across Specifications

Appendix Table C1 presents the complete model results for the linear growth model specification, including estimated fixed effects, random effects, 95% plausible value ranges at the within- and between-school levels, and reliabilities at both levels. Appendix Table C2 presents the same results for the quadratic growth model specification.

We can use results reported in Appendix Table C1 and C2 to further characterize the degree of across-school variability we would expect to observe in achievement growth: For math, the mean 1 through 5 growth rate is 6.2 RIT score points per period (recall we have fall and spring data, so there are two periods per year). Based on the magnitude of the between school variance, we can expect to see that 95% of schools will have a mean growth rate between 5.0 and 7.4 RIT points per period. At first this might not sound large, but recall there are 10 periods in grades 1 through 5 and so the differences add up: We would expect a typical student from a school at the low end of this range to gain 55 RIT points between grade 1 to 5, while a typical student in a school at the high end of the range would gain 70 RIT points. Results are similar for reading (45 versus 63 RIT points). The annual growth norms reported by NWEA vary significantly from grade to grade, making it somewhat difficult to think about the size of these differences (e.g., the mean growth is 21.2 in 1st grade, 8.6 in 5th grade, and 4.0 in 8th grade). Regardless of which grade level growth norm one uses, these benchmarks generally suggest that the magnitude of these differences in gains across schools is meaningfully large.

Table C1. Full Results from Model 1, Linear Growth Model

	MODEL 1 - LINEAR GROWTH			
	Grades 1 Through 5		Grades 6 Through 8	
	Math	Reading	Math	Reading
Fixed Effects				
Achievement Levels @ Midpoint	200.9	196.8	227.9	218.4
Achievement Rates Across Period	6.24	5.42	3.22	2.30
Rate of Deceleration Across Period	N/A	N/A	N/A	N/A
Total Gain Start to End of Period	62.4	54.2	19.3	13.8
Variance Components				
Achievement Levels @ Midpoint				
SD Among Students Within Schools	10.0	10.4	13.4	11.4
SD Across Schools	5.6	5.5	7.5	6.3
% of Variance Across Schools (ICC)	23.7%	21.4%	23.5%	23.4%
Achievement Rates Across Period				
SD Among Students Within Schools	0.37	0.23	0.46	0.22
SD Across Schools	0.61	0.48	0.62	0.43
% of Variance Across Schools (ICC)	73.5%	80.7%	64.9%	79.5%
Rate of Deceleration Across Period				
SD Among Students Within Schools	N/A	N/A	N/A	N/A
SD Across Schools	N/A	N/A	N/A	N/A
% of Variance Across Schools (ICC)	N/A	N/A	N/A	N/A
95% Plausible Value Ranges around Fixed Effect Parameters, based on Variance Parameters				
Achievement Levels @ Midpoint				
95% PVR of Means Within Schools	(181 , 221)	(176 , 217)	(202 , 254)	(196 , 241)
95% PVR of Means Across Schools	(190 , 212)	(186 , 208)	(213 , 243)	(206 , 231)
Achievement Rates Across Period				
95% PVR of Slopes Within Schools	(5.5 , 7)	(5 , 5.9)	(2.3 , 4.1)	(1.9 , 2.7)
95% PVR of Slopes Across Schools	(5 , 7.4)	(4.5 , 6.3)	(2 , 4.4)	(1.5 , 3.1)
Rate of Deceleration Across Period				
95% PVR of Decel. Within Schools	(N/A , N/A)	(N/A , N/A)	(N/A , N/A)	(N/A , N/A)
95% PVR of Decel. Across Schools	(N/A , N/A)	(N/A , N/A)	(N/A , N/A)	(N/A , N/A)
Total Gain Start to End of Period				
95% PVR Within Schools	(55 , 70)	(50 , 59)	(23 , 41)	(19 , 27)
95% PVR Across Schools	(50 , 74)	(45 , 63)	(20 , 44)	(15 , 31)
Reliability Estimates				
Achievement Levels @ Midpoint				
Reliability of Intercepts for Students	0.909	0.910	0.954	0.930
Reliability of Intercepts for Schools	0.783	0.763	0.856	0.854
Achievement Rates Across Period				
Reliability of Grade Slopes for Students	0.240	0.308	0.270	0.303
Reliability of Grade Slopes for Schools	0.793	0.712	0.771	0.671
Rate of Deceleration Across Period				
Reliability of Deceleration for Students	N/A	N/A	N/A	N/A
Reliability of Deceleration for Schools	N/A	N/A	N/A	N/A
Sample Sizes				
Repeated Observations	632,750	638,940	2,480,106	2,653,068
Students	63,275	63,894	413,351	442,178
Schools	2,305	2,277	3,798	3,780

Table C2. Full Results from Model 2, Quadratic Growth Model

	MODEL 2 - QUADRATIC GROWTH			
	Grades 1 Through 5		Grades 6 Through 8	
	Math	Reading	Math	Reading
Fixed Effects				
Achievement Levels @ Midpoint	203.3	200.3	228.3	218.6
Achievement Rates Across Period	5.93	4.98	3.05	2.21
Rate of Deceleration Across Period	-0.31	-0.44	-0.17	-0.08
Total Gain Start to End of Period	28.7	6.3	12.3	10.2
Variance Components				
Achievement Levels @ Midpoint				
SD Among Students Within Schools	9.5	11.2	13.5	11.4
SD Across Schools	5.3	5.7	7.6	6.4
% of Variance Across Schools (ICC)	23.9%	20.7%	23.9%	23.7%
Achievement Rates Across Period				
SD Among Students Within Schools	0.39	0.34	0.43	0.27
SD Across Schools	0.62	0.44	0.64	0.44
% of Variance Across Schools (ICC)	71.1%	62.0%	68.8%	73.4%
Rate of Deceleration Across Period				
SD Among Students Within Schools	0.08	0.10	0.11	0.07
SD Across Schools	0.11	0.09	0.20	0.16
% of Variance Across Schools (ICC)	64.9%	45.3%	75.3%	84.0%
95% Plausible Value Ranges around Fixed Effect Parameters, based on Variance Parameters				
Achievement Levels @ Midpoint				
95% PVR of Means Within Schools	(185 , 222)	(178 , 222)	(202 , 255)	(196 , 241)
95% PVR of Means Across Schools	(193 , 214)	(189 , 212)	(214 , 243)	(206 , 231)
Achievement Rates Across Period				
95% PVR of Slopes Within Schools	(5.2 , 6.7)	(4.3 , 5.7)	(2.2 , 3.9)	(1.7 , 2.7)
95% PVR of Slopes Across Schools	(4.7 , 7.1)	(4.1 , 5.8)	(1.8 , 4.3)	(1.3 , 3.1)
Rate of Deceleration Across Period				
95% PVR of Decel. Within Schools	(-0.5 , -0.1)	(-0.6 , -0.2)	(-0.4 , 0.1)	(-0.2 , 0.1)
95% PVR of Decel. Across Schools	(-0.5 , -0.1)	(-0.6 , -0.3)	(-0.6 , 0.2)	(-0.4 , 0.2)
Total Gain Start to End of Period				
95% PVR Within Schools	(2 , 57)	(-17 , 37)	(-18 , 49)	(-3 , 37)
95% PVR Across Schools	(-3 , 61)	(-19 , 28)	(-42 , 63)	(-27 , 51)
Reliability Estimates				
Achievement Levels @ Midpoint				
Reliability of Intercepts for Students	0.815	0.860	0.911	0.868
Reliability of Intercepts for Schools	0.770	0.749	0.854	0.850
Achievement Rates Across Period				
Reliability of Grade Slopes for Students	0.408	0.385	0.344	0.306
Reliability of Grade Slopes for Schools	0.775	0.672	0.735	0.625
Rate of Deceleration Across Period				
Reliability of Deceleration for Students	0.238	0.251	0.203	0.203
Reliability of Deceleration for Schools	0.544	0.444	0.558	0.476
Sample Sizes				
Repeated Observations	632,750	638,940	2,480,106	2,653,068
Students	63,275	63,894	413,351	442,178
Schools	2,305	2,277	3,798	3,780

Table C3. Proportion of Variance Between Schools for Achievement Levels, Linear Growth, and Growth Curve/Acceleration, Across Different Permutations of the Model and Analytic Sample

					Achievement Level: % Between Sch's		Learning Rate: % Between Sch's		Acceleration: % Between Sch's	
					Math	ELA	Math	ELA	Math	ELA
(A)	At Least 75%	**Linear Growth**	Grades 1 - 5	50% sample	23.2% *(5,545)*		75.4% *(5,545)*			
(B)	Has All Scores	**Linear Growth**	Grades 1 - 5	2% sample	23.4% *(1,250)*	21.6% *(1,244)*	77.1% *(1,250)*	84.3% *(1,244)*		
				10% sample	21.0% *(2,017)*	18.6% *(1,985)*	75.2% *(2,017)*	84.0% *(1,985)*		
			Grades 6 - 8	2% sample	19.2% *(2,717)*	17.4% *(2,748)*	53.7% *(2,717)*	70.3% *(2,748)*		
				10% sample	18.5% *(3,470)*	17.6% *(3,472)*	59.3% *(3,470)*	82.4% *(3,472)*		
(C)	Has All Scores	**Non-Linear Growth**	Grades 1 - 5	2% sample	24.1% *(1,250)*	20.9% *(1,244)*	67.6% *(1,250)*	71.0% *(1,244)*	24.1% *(1,250)*	33.4% *(1,244)*
				10% sample	21.2% *(2,017)*	18.1% *(1,985)*	68.4% *(2,017)*	68.1% *(1,985)*	37.0% *(2,017)*	38.0% *(1,985)*
			Grades 6 - 8	2% sample	19.4% *(2,717)*	17.8% *(2,748)*	58.7% *(2,717)*	63.3% *(2,748)*	53.3% *(2,717)*	38.3% *(2,748)*
				10% sample	18.8% *(3,470)*	18.1% *(3,472)*	63.0% *(3,470)*	81.0% *(3,472)*	70.1% *(3,470)*	74.4% *(3,472)*
(D)	Has All Scores	**Policy Model**	Grades 1 - 5	50% sample	24.2% *(2,306)*	22.4% *(2,280)*	43.5% *(2,306)*	38.9% *(2,280)*		
			Grades 6 - 8	50% sample	23.4% *(3,800)*	23.4% *(3,783)*	45.3% *(3,800)*	36.8% *(3,783)*		

FN: %'s reported ICC's-- % of variation in the scores, linear rates, or acceleration between schools (relative to within schools). N = (# of schools). (A) Examines results with a less stringent requirement for minimum number of scores required per student to be included in the analytic sample (grades 1 - 5). (B) Examines results when using a 2% vs. 10% sample of the full analytic sample for a linear growth model (grades 1 - 5, 6 - 8). (C) Is that same as (B) but for a non-linear (quadratic) growth model. (D) "Policy model" examines results when only spring scores are included, and those scores are standardized within subject-grade-year.

Endnotes

[1] This includes, for example, debates about whether the U.S. over-invests funding in schools (Hanushek, 1989; Hedges, Laine, & Greenwald, 1994), as a reason to support school choice (Towers, 1992), or arguments for greater focus on out-of-school learning (Whitehurst & Croft, 2010).

[2] (Alexander & Entwisle, 1996; Barr, Dreeben, & Wiratchai, 1983; Borman & Dowling, 2010; Cain & Watts, 1970; Carver, 1975; Domingue, Thomas, Circi, & Camilli, 2011; Hanushek & Kain, 1972; Heckman & Neal, 1996).

[3] For instance, Coleman could not consider the role of teachers, which have been shown to influence within-school achievement (see e.g., Aaronson, Barrow, & Sander, 2007; Chetty, Friedman, & Rockoff, 2011). Summer learning loss researchers often point out that the average U.S. high school graduate spends less than 15% of their waking hours in school (Downey, von Hippel, & Broh, 2004; Walberg, 1984), which implies that the Coleman Report conflates school effects with a great deal of out-of-school time (Downey & Condron, 2016).

[4] However, data for the Report was collected within one year to meet the timeline established by the Civil Rights Act.

[5] We use the term "growth" to refer to the annual rate of change of achievement test scores over time.

[6] Bryk and Raudenbush (1988) use a three-level hierarchical framework to model (up to) five vertically-scaled test scores administered between grades 1 and 3 as a linear function of time, nested within students at level two (N= 618), who are in turn nested in schools at level three (N=86). This model is discussed in greater depth in Appendix A, along with an overview of several other studies that have implemented similar models but did not report ICCs in growth rates.

[7] For a full description of this dataset and the analytic sample, see Appendix B.

[8] NWEA technical manuals indicate that the RIT scale is produced using a one-parameter logistic item response theory model. For more information on the scale development, see Appendix B or NWEA (2011).

[9] We make the following analytic choices for our primary specification: (1) We use RIT scale scores that have been projected to the first day of school in the fall semester and the last day of school in the spring semester. These projected scores are correlated at 0.996 with the RIT scores observed on the actual test date from the same semester (see Authors, 2020). (2) We run separate models for grades 1-5 and 6-8. (3) Only students with a full vector of fall and spring test scores across those grades are included in the model, to eliminate concerns about sample compositional changes over time.

[10] For instance, results are similar when we use a diminishing returns, log-linear function.

[11] Because we possess a longer panel of test scores (e.g., up to 10 scores from grade 1 to 5), a quadratic growth model (M2) allows us to also consider the extent to which students' learning rates *accelerate/decelerate* while in school. This too seems like a phenomenon that is more under the direct purview of schools than achievement levels.

[12] We rerun these models with a variety of different analytic choices. For example, we estimate models using observed RIT scores instead of projected scores; using a linear, quadratic, or log-linear time function; placing a less stringent requirement on the number of RIT scores a student must have to be included, centering time at level one in the earliest grade of the panel, rather than a midpoint, or limiting the analysis to large schools. The basic results hold. In none of these permutations is the between-school variance in achievement growth rates close to the conventional 15% - 20% mark that we have come to think of as the amount of variation in achievement across schools.

[13] These student- and school-specific estimates of linear learning rates are the empirical Bayes residual estimates of the model intercept and coefficient on grade/time.

[14] The model-based estimates presented in Appendix Table C1 suggest that 95% of schools would exhibit a mean grade 1-to-5 score gain between 55 and 70 RIT score points (math). That is, we would expect a typical student from a school at the low end of this range to gain 55 RIT points between grade 1 to 5, while a typical student in a school at the high end of the range would gain 70 RIT points during the same time frame. For context, the annual growth norm reported by NWEA is 8.6 points in grade 5. See Appendix C for additional discussion of the magnitude of across-school growth rates.

[15] Rumberger and Palardy (2005) *also* uses a three-level linear growth model and documents the proportion of variation in learning rates between schools (see their Appendix Table 3). As in the current study, the between-school variance in growth rates is closer to 20-30% when using standardized scores. See a detailed description of this study in Appendix A.

[16] See Appendix A for a thorough description of the following relevant studies: Briggs and Domingue (2013); Carbonaro and Gamoran (2002); Lee, Smith, and Croninger (1997); Rumberger and Palardy (2005).

[17] We examined students' IRT-based theta scores from the end of K, 1, 3, 5, and 8 as our outcomes first as a linear and then as a quadratic function of grade. We opted to use the theta scores ECLS-K:99 provides (rather than scale scores or standardized T-scores) because, like NWEA's RIT scores, theta scores are IRT-based and ECLS-K:99

documentation states that "theta scores are ideally suited for measuring growth from kindergarten through eighth grade" (Tourangeau et al., 2009, p. 3—9). We emphasize that these results are preliminary, because the sampling procedures used to construct the ECLS-K:99 dataset means that there is an average of only about 15 students per school. Estimating within-school variances may be more difficult in this setting. In addition, we did not apply relevant sampling weights. We ran models using three different sets of inclusion criteria: (1) a *least* restrictive approach, in which we included any students with any number of scores nested in any schools (33,447 students in 6590 schools); (2) a *most* restrictive approach, in which we only included students who had at least 4 of 5 possible theta scores and who never move schools and only in schools with at least 10 of these students (1,011 students in 74 schools); and (3) an approach that maximizes the amount of within-school data, in which we included any students in schools with at least 20 students (5,398 students in 248 schools).