School Effectiveness and School Improvement
Vol. 20, No. 2, June 2009, 187–213

Routledge
Taylor & Francis Group

Achievement, learning, and seasonal impact as measures of school effectiveness: it's better to be valid than reliable

Paul T. von Hippel*

Department of Sociology, The Ohio State University, Columbus, OH, USA

(*Received ????; final version received ????*)

When evaluating schools' effectiveness, is it better to use absolute *achievement levels* or to use *learning rates* over a 9- or 12-month period? Or is it better to use a new measure, seasonal *impact*, which is defined as the acceleration in learning rates that occurs when students finish summer vacation and start the school year? Answering this question involves a tradeoff between validity and reliability, since unfortunately the most reliable measure (achievement) is the least-valid measure as well. In this paper, we evaluate the tradeoff using a simple covariance-structure model. The results suggest that, when achievement is compared to learning and impact, reliability is a minor issue compared to validity. Although achievement is more reliable than learning or impact, achievement's advantage in reliability is much smaller than its disadvantage in validity. Our findings support the view that reliability is not a sufficient reason to evaluate schools using achievement levels.

Keywords: seasonal impact; school effectiveness; evaluation; learning gain; learning growth

Introduction

Most US schools today are evaluated on the basis of their achievement levels – either the average score of their students on an achievement test or the percentage of students scoring above some threshold of proficiency. Yet, a school's achievement level is strongly influenced by contextual or intake variables that are beyond the school's control, most notably the socioeconomic composition of the student body and the prior achievement of students before the school year begins (e.g., Coleman et al., 1966; Dumay & Dupriez, 2007; Fitz-Gibbon, 1996; Opdenakker & Van Damme, 2006, 2007; Sammons, Mortimore, & Thomas, 1996; Thrupp, 1999; Van de Grift, this issue; Willms, 1992). Recognizing that different schools enroll profoundly different types of students, school effectiveness researchers have long criticized raw achievement levels as a measure of school effectiveness (e.g., Teddlie, Reynolds, & Sammons, 2000).

As an alternative to achievement-based evaluations, many school effectiveness researchers advocate evaluating schools using longitudinal data in which achievement is measured at several points in time (e.g., Willett, 1988; Willms, 1992). If achievement is measured on two occasions, evaluators can subtract the first score

*Email: von-hippel.1@osu.edu

ISSN 0924-3453 print/ISSN 1744-5124 online

DOI: 10.1080/09243450902883888
http://www.informaworld.com

from the second to estimate the amount learned between tests. Alternatively, especially if achievement is measured on more than two occasions, evaluators can estimate learning rates using a multilevel growth model (Raudenbush & Bryk, 2002). Another alternative is to control for prior achievement by regressing a later achievement score on an earlier one. Such a lagged-score analysis is usually similar to an analysis based on subtracting successive scores, although subtle differences in the assumptions can occasionally lead to noticeably different results (Holland & Rubin, 1983; Johnson, 2005; Lord, 1967).

Compared to achievement levels, learning rates are more susceptible to schools' control and are therefore a more valid measure of school effectiveness. On the other hand, learning rates are also a less *reliable* measure than achievement levels, in the sense that a school's learning rates vary more than its achievement levels from one year to the next (e.g., Kane & Staiger, 2002).

Although learning-based evaluation is an improvement on achievement-based analysis, learning rates are still affected by inputs over which schools have little control. Children who learn quickly during the school year may do so because of effective schools, or they may learn quickly because of advantages in their nonschool environments. In an attempt to overcome these limitations, Downey, von Hippel, and Hughes (2008) proposed a new measure of school effectiveness that they call seasonal *impact*. Impact is the difference between school-year and summer learning rates and is interpreted as the degree to which schools increase children's rate of learning above the rate that they learn when they are not in school. The impact measure is conceptually similar to measures obtained when children are deprived of schooling – for example, in countries where schooling is not compulsory (e.g., Heynemann & Loxley, 1983), or when schools are shut down by strikes or political events (e.g., Miller, 1983) – but the impact measure has the advantage that it can be applied under normal conditions when schooling has not been disrupted.

Unfortunately, although impact may be a more valid index of school effectiveness than learning or achievement, impact is also less reliable than those alternatives.

In short, competing measures of school effectiveness differ both in validity and in reliability. This paper outlines these differences and develops a statistical approach to evaluating the validity-reliability tradeoff. In general, the analysis shows that the benefits of switching to a more valid measure of school effectiveness outweigh the costs in reliability. The results also show that achievement levels, despite their high reliability, are not a good proxy for more valid measures of school effectiveness.

Three ways to measure school effectiveness

Achievement levels

Most schools today are evaluated in terms of their achievement levels. In the USA, under the federal No Child Left Behind Act (NCLB), schools are labeled as "failing" or "in need of improvement" if an unacceptable fraction of students score below a state-defined standard on proficiency tests. Even before NCLB, most state and city accountability systems were already based on achievement levels, and much research focused on achievement levels as well. Parents and homeowners also seem to judge schools on the basis of achievement levels; near a school with high achievement levels, the price of housing tends to be high as well (e.g., Black, 1999).

Achievement-based evaluation relies on the assumption that student achievement is a direct measure of school quality. But this assumption is not correct. Although student achievement is affected by schools, achievement is strongly influenced by contextual or intake variables that are beyond schools' control, most notably student socioeconomic status and the prior achievement of students before the school year begins (e.g., Coleman et al., 1966; Fitz-Gibbon, 1996; Sammons et al., 1996; Willms, 1992). The vast majority of the variation in test scores lies within rather than between schools (Coleman et al., 1966; Scheerens, 1992), and even between-school score differences are not necessarily the result of differences in school quality. Instead, many differences between schools' average test scores come from the fact that different schools serve different kinds of students.

On the first day of kindergarten, we can already see clearly that school achievement levels are not a pure index of school quality. If school quality were the only important influence on achievement levels, we would expect different schools to have similar average scores on the first day of kindergarten, before school quality has had a chance to make it mark. But this is not what we see. Instead, on the first day of kindergarten, almost a quarter of the variation in reading and math scores already lies between rather than within schools (Downey, von Hippel, & Broh, 2004; Lee & Burkham, 2002). The between-school gaps observed on the first day of kindergarten do not disappear with time, nor are they swamped by achievement gains in the later grades. In fact, even at the end of ninth grade, almost a third of the gap between poor and middle-class children's achievement scores can be traced to gaps that were present at the beginning of first grade (Alexander, Entwisle, & Olson, 2007).

It does not make sense to give schools credit for what students learn before the start of kindergarten. Yet this is exactly what an achievement-based system does.

Learning rates

Twelfth-month learning

One way to avoid some of the bias of achievement-based evaluation is to base evaluations on *learning rates* – also called growth, gains, or progress – over a calendar year. The advantage of learning rates is that they do not obviously credit or penalize a school for the knowledge that its students accumulated before school began. For example, a school whose test scores average 60 at the end of kindergarten and 80 at the end of first grade would get credit for 20 points of learning, and so would a school kindergarten and first-grade scores averaged 70 and 90. The latter school would not get extra credit for the fact that its students started first grade with a 10-point head start.

Compared to raw achievement levels, learning is clearly a better measure of school effectiveness. As longitudinal data become more widely available, academic researchers have shifted some of their focus from achievement to learning (Raudenbush & Bryk, 2002, chapter 6; Scheerens & Bosker, 1997; Teddlie & Reynolds, 2000) In the USA, the Tennessee Value Added Assessment System (TVAAS) focuses on annual learning or gains,[1] and so do evaluation systems in Dallas and North and South Carolina (Chatterji, 2002; Kupermintz, 2002). Although the federal NLCB Act requires that schools be evaluated in terms of achievement levels, 14 states have asked federal permission to evaluate schools on learning instead (Schemo, 2004). Under the NCLB's testing schedule, measuring

annual learning is quite practical, since NCLB requires that students be tested at the end of every school year from third to ninth grade.

Unfortunately, although annual learning rates are far more valid than achievement levels, 12-month learning is still not a clean measure of school quality. The most obvious problem with 12-month learning is that it holds schools accountable for learning that occurs over the summer, when school is out of session. Children's summer learning rates are very unequal – much more unequal than they are during the school year – and the children who make big summer gains tend to be relatively affluent, the same kinds of children who start with high scores on the first day of kindergarten (Downey et al., 2004; Entwisle & Alexander, 1992).

It makes no more sense to evaluate schools on summer learning than it does to evaluate them on learning before the start of kindergarten. Yet this is just what a system based on 12-month learning does.

Nine-month learning

To avoid holding schools responsible for summer learning, a simple refinement is to focus on *9-month* rather than 12-month learning rates. In a 9-month evaluation system, students would be tested at the beginning and end of selected academic years, so that only learning during the academic year would be viewed as an index of school quality. Under NCLB, such a system could be implemented by doubling the number of achievement tests, pairing each spring test with a test given the previous fall. Or, if doubling the number of tests seems onerous,[2] selected spring tests could simply be moved to the following fall. For example, instead of testing at the end of third grade, schools could test students at the beginning of fourth grade. Since NCLB already requires a test at the end of fourth grade, it would be straightforward to estimate learning for the fourth-grade academic year.

Covariate adjustment

Nine- or even 12-month learning would be a much better basis for school evaluation than most of the systems currently in place – and if this paper simply helps to encourage the spread of learning-based evaluation, we will be more than satisfied.

But school-year learning is still not an entirely valid measure of school quality. Even though school-year learning excludes the summer, the variation in summer-learning rates draws our attention to the fact that children's learning is subject to many non-school influences. Children's non-school environments differ tremendously in how much they encourage the development of academic skills. The summer learning gap between poor and middle-class children attests to these out-of-school differences (Alexander, Entwisle, & Olson, 2001; Downey et al., 2004; Heyns, 1978; Verachtert, Van Damme, Onghena, & Ghesquière, this issue), and so does detailed observation of poor and middle-class families' parenting styles (Lareau, 2000; Linver, Brooks-Dunn, & Kohen, 2002). As the *Seven Up* documentary showed us 40 years ago (Apted, 1963), some children's after-school hours are packed with structured learning activities such as music and dance lessons and supervised homework, while other children are left more or less on their own. The effect of these non-school influences is muted during the school year but still must have substantial influence, since even during the academic year, students spend more than two thirds

of their waking hours outside of school (Hofferth & Sandberg, 2001; Downey et al., 2008).

One approach to the problem of non-school influences is to "adjust" learning rates (or achievement levels), using student characteristics such as poverty and ethnicity (e.g., Clotfelter & Ladd, 1996; Ladd & Walsh, 2002; Van de Grift, this issue). But this approach has three problems. First, it is politically contentious to adjust for poverty and ethnicity, since such adjustments seem to hold poor and minority children to lower standards.[3] Second, the usual crude measures of class and ethnicity do not come close to fully capturing non-school influences on learning (Meyer, 1996, p. 210). For example, during the summer, when only non-school influences are operative, measures of race, gender, and socioeconomic status explain only 1% of the variation in learning rates (Downey et al., 2004). A final problem is that, since school quality is probably correlated with student characteristics (Dumay & Dupriez, 2007; Opdenakker & Van Damme, 2006, 2007; Thrupp, 1999), it is almost impossible to remove non-school influences on learning without removing some school influences as well. For example, if schools serving Black children have lower average quality than schools serving White children, then removing the effect associated with observed race will simultaneously remove the effect of unobserved quality differences between Black and White schools (cf. Rubenstein, Stiefel, Schwartz, & Amor, 2004, p. 59).

In short, using student covariates to adjust estimates of school quality is politically controversial and not as methodologically attractive as it may first appear.

Seasonal "impact"

Since summer learning provides a window into children's non-school environments, Downey et al. (2008) have suggested that schools be evaluated by comparing summer learning rates to learning rates during the school year. During the school year, learning is shaped by both school and non-school influences, while during the summer, learning is shaped by non-school influences alone (cf. Heyns, 1978).

The difference between school-year and summer learning rates is an estimate of the marginal effect, or *impact*, of schooling (Downey et al., 2008). Evaluating schools on the basis of seasonal impact can radically change our picture of which schools are effective; in particular, schools serving disadvantaged children who learn little over the summer are more likely to seem effective under an impact-based system than under a system based on achievement or learning (Downey et al., 2008). For example, suppose that we are comparing two schools where children gain an average of 3 achievement points per month during the academic year. At first, these schools might seem equally effective. But suppose one school serves affluent children with a summer learning rate of 2 points per month, while the other school serves relatively disadvantaged children with a summer learning rate of just 1 point per month. Clearly the school that serves slow summer learners is having more *impact*; it is doing more to accelerate its students' learning rates. Specifically, the school with slow summer learners has an impact of $3-1=2$ points per month, while the school with fast summer learners has an impact of just $3-2=1$ points per month.

Figure 1 illustrates the relationship between achievement, learning, and impact.

By accounting for non-school influences on learning, seasonal impact may provide a more valid measure of school effectiveness than do competing methods.

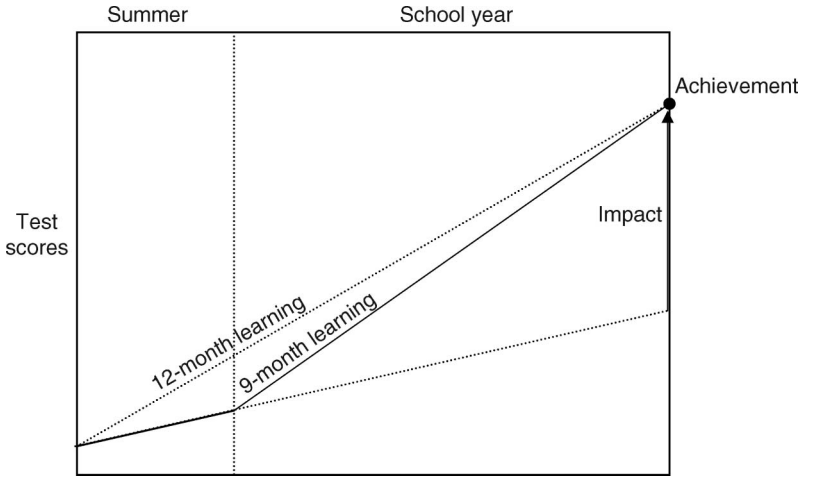

Figure 1. Achievement, learning, and impact: schematic diagram.

Or it may not. On close examination, the impact measure makes a number of assumptions that may or may not be more plausible than the assumptions of other evaluation methods. For one thing, subtracting summer learning from school-year learning implies that non-school influences have the same effect during the school year as they do during the summer. This may be false: Perhaps non-school influences matter *less* during the school year, when children are spending less time in their non-school environment. In that case, it would be better to subtract just a fraction of the summer learning rate, but the correct fraction is hard to know.[4] The use of subtraction also implies that school and non-school influences are additive in their effects, but it may be that the influences interact as well. For example, it may be that parents aid learning not so much by teaching children directly, but by providing breakfast, monitoring homework, and generally preparing children to get the most out of school. This raises the further possibility that the non-school influences that matter during the school year may be not just smaller but *different* than the influences that matter during the summer. In that case, measures of summer learning may not provide a clear window onto the non-school influences that matter during the school year. Finally, it may be that the most effective schools increase learning not just during the school year but during the summer as well, by, for example, assigning summer reading. As it happens, the data that we analyze in this paper identify schools that assign summer book lists; surprisingly, summer learning is no faster in those schools than elsewhere. This finding is disappointing with respect to the effectiveness of summer book lists but reassuring with regard to the validity of impact-based evaluation.

While the assumptions of impact-based evaluation are nontrivial, we should bear in mind that *every* school evaluation measure makes assumptions. The assumptions behind impact-based evaluation should be compared to those required for evaluation based on achievement or learning. As previously noted, evaluation systems based on achievement levels or learning rates assume that non-school factors play a minor role in shaping student outcomes. This assumption is badly wrong for achievement levels and somewhat wrong for learning rates.

Validity versus reliability

To this point, we have compared measures of school effectiveness with respect to *validity*. As measures of school effectiveness, learning rates are clearly more valid than achievement levels, and school-year learning is clearly more valid than learning observed over a calendar year. It may also be that impact is a more valid measure than learning – but the advantages of impact are more arguable.

These statements are based on considerations of *content* validity. That is, we maintain that the *contents* of a school evaluation measure should exclude sources of variation that schools cannot control. For example, achievement scores have less content validity than learning rates, because achievement scores *contain* knowledge that children accumulate in the years before schooling begins. Likewise, 12-month learning has less content validity than school-year learning, because 12-month learning contains summer learning over which schools have little control. Even school-year learning has imperfect content validity, because during the school year, children spend two thirds of their waking hours in non-school environments (Hofferth & Sandberg, 2001). Impact, then, is an attempt to further improve content validity by emptying school-year learning of some non-school contents.

Validity, however, is not the only desirable quality of a school evaluation measure. *Reliability* is also an important consideration. A school evaluation measure may be quite valid and yet highly unreliable in the sense that it cannot be estimated without substantial sampling error and measurement error. In school evaluation research, the role of reliability is sometimes addressed by reporting confidence intervals as well as point estimates of school effectiveness (Goldstein, 1995).

Unfortunately, there is often a tradeoff between validity and reliability. For example, learning rates, although more valid than achievement levels, also tend to be less reliable in the sense that learning cannot be estimated as accurately as achievement levels (Kane & Staiger, 2002). To put the issue heuristically, the difference between two achievement scores is usually more variable than either score is by itself. Since impact involves comparing *three* test scores, it turns out that impact is even less reliable than learning.

Evaluating the tradeoff

We face a tradeoff, then. Some measures of school effectiveness are more valid than others, but the more valid measures are often less reliable as well. School evaluation researchers have debated the tradeoff between validity and reliability (e.g., Kane & Staiger, 2002; Ladd, 2002) but have not offered a way to evaluate the tradeoff and decide which measure is, on balance, the most accurate measure of school quality.

In this paper, we evaluate the tradeoff between validity and reliability using a simple variance-components model. To foreshadow the conclusions, we find that the benefits of switching to a more valid measure of school effectiveness outweigh the costs in reliability. Not only is learning a more valid criterion than achievement levels, but the gain in validity is greater than the loss in reliability. Likewise, if we take the position that impact is a more valid criterion than learning, then, again, the cost in reliability is offset by the benefit in validity.

Data and methods

Observed data

To evaluate the validity and reliability of competing measures of school effectiveness, we use data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) – the only national US survey that permits separate estimation of school-year and summer learning rates (http://nces.ed.gov/ecls/Kindergarten.asp). In the ECLS-K, 992 schools were visited for testing in the fall of kindergarten (Time 1), the spring of kindergarten (Time 2), and the spring of first grade (Time 4). Among these 992 schools, 309 were randomly selected for an extra test in the fall of first grade (Time 3). Since this extra test is essential for estimating summer and first-grade learning rates – and therefore essential for measuring impact – we focus on the 309 schools where the extra test was given. Since the summer learning rate was meant to be a window into the non-school environment, we excluded schools that used year-round calendars and children who attended summer school; we also deleted any scores collected after a child had transferred from one school to another.[5] In the end, our analysis focused on 4,217 children in 287 schools.

Excellent tests of reading and mathematics skill were designed especially for the ECLS-K (Rock & Pollack, 2002). The tests were over 90% reliable and tried to measure the full range of young children's abilities – from rudimentary skills, such as recognizing isolated letters and numbers, to advanced skills, such as reading words in context and solving multiplication and division problems. Using Item Response Theory, the ECLS-K scored reading on a 92-point scale and scored mathematics on a 64-point scale. We can think of each point on the scale as the amount learned in 2 to 4 weeks of school, since our later estimates suggest that most children gained between 1 and 2.5 points per month when school was in session.

Missing data

Missing values were filled in using a multiple imputation strategy (Rubin, 1987). We did not need to impute missing test scores, since imputed values of the dependent variable contribute no information to the analysis (Little, 1992; von Hippel, 2007). Likewise, we did not need to impute missing test dates, since test dates are only missing when test scores are missing as well (i.e., when no test was given). The only variables that we needed to impute were the dates for the beginning and end of the school year, and those dates are very predictable since they vary little from one school to another. Our results are therefore quite robust to our treatment of missing data.

Since school dates vary only at the school level, we created a school-level dataset containing one line per school, with school dates[6] alongside test dates and test scores averaged up to the school level. By putting these variables on a single line, we were able to account for the serial correlation among successive tests given in the same school (Allison, 2002, p. 74). In this school-level dataset, we imputed 10 sets of values under a multivariate normal model.[7] Following imputation, we merged the imputed school dates with the observed test scores and test dates.

Again, though we used test scores and test dates to improve the imputation of school dates, we did not use imputed test scores and test dates in the analysis. This strategy is known as *multiple imputation, then deletion* (MID); the dependent variable Y is used to impute independent variables X, but imputed Ys are not used in the analysis (von Hippel, 2007). MID is an especially appropriate strategy for the

ECLS-K, since it is difficult to impute Y in a way that properly accounts for the multilevel variance structure. On the other hand, since only 7% of Y values were missing from the sampled schools, the handling of imputed Ys makes little difference, and an ordinary multiple imputation strategy would yield very similar results. We have analyzed these data using a variety of different imputation strategies, without material effects on the results.

Statistical methods

If each child were tested on the first and last day of each school year, then monthly learning rates could be estimated simply by subtracting successive test scores and dividing by the months elapsed between tests. In the ECLS-K, however, schools were visited on a staggered schedule, so that, depending on the school, fall and spring measurements were taken anywhere from 1 to 3 months from the beginning or end of the school year. To compensate for the varied timing of achievement tests, we must adjust for the time that children had been in school and on vacation before each test was given. In addition, to evaluate the reliability of different school effectiveness criteria, we must break the variation of each criterion into within- and between-school components.

To achieve these goals, we estimated schools' achievement levels, learning rates, and impact using a multilevel growth model (Raudenbush & Bryk, 2002). Briefly, we fit a 3-level model in which test scores (Level 1) were nested within children, and children (Level 2) were nested within schools (Level 3). The model adjusted for the time elapsed between test dates and school dates and produced estimates of how much achievement levels and learning rates vary within and between schools. A detailed specification of the model is given in Appendix 1.

Results

In this section, we compare school evaluation methods based on achievement levels, learning rates, and impact. We focus on the results for reading. Results for mathematics, which were generally similar, are given in Appendix 2.

Reliability

Table 1 averages achievement, learning, and impact across the ECLS-K's random sample of US schools. At the end of first grade, the average achievement level in reading was 59.33 out of 92 possible points. Children reached this achievement level by learning at an average rate of 1.70 points per month during kindergarten, then losing an average of 0.08 points per month during summer vacation, and then gaining an average of 2.57 points per month during first grade. So school impact – the difference between first-grade and summer learning rates – had an average value of 2.64 points per month; and 12-month learning – the average monthly learning rate across the 12 months of summer and first grade – had an average value of 1.99 points per month. Each measure of school effectiveness varied substantially across schools; for every measure – achievement, 9-month and 12-month learning, and impact – the between-school standard deviation was at least 12% of the mean.

Our ability to reliably detect between-school differences depends in part on how much of the variation lies between rather than within schools. The variance of

Table 1. Reading achievement, learning, and impact, as measured on a 92-point scale.

			Achievement, end of first grade	Monthly learning rates				
				Kindergarten	Summer	first grade	12-month[a]	Impact[b]
Fixed effects	Mean		59.33***	1.70***	−0.08	2.57***	1.99***	2.64***
Random effects	Variances	Between school (school-level)	7.07^2 (23%)	0.39^2 (16%)	0.57^2 (7%)	0.45^2 (16%)	0.36^2 (17%)	0.78^2 (8%)
		Within school: Child-level	12.55^2 (74%)	0.83^2 (71%)	1.54^2 (52%)	0.97^2 (73%)	0.73^2 (72%)	2.05^2 (58%)
		Within school: Test-level	2.45^2 (3%)	0.35^2 (13%)	1.35^2 (40%)	0.38^2 (11%)	0.29^2 (11%)	1.58^2 (34%)
	Total variance		14.61^2	0.98^2	2.13^2	1.13^2	0.86^2	2.70^2
Reliability of school mean if 20 children are sampled per school[c]			86%	79%	61%	79%	81%	65%

$^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

[a]Twelve-month learning is reckoned from the end of kindergarten to the end of first-grade.

[b]Impact is the difference between the first-grade and summer learning rates.

[c]If n children are sampled per school, then the reliability of the school mean is $p/[p + (1–p)/n]$, where p is the proportion of variance that lies at the school level.

a school mean is equal to the within-school variance divided by the within-school sample size, so if the within-school variance is large compared to the between-school variance, then between-school differences may be hard to detect in a reliable fashion.

To evaluate reliability, Table 1 partitions achievement, learning, and impact into between-school and within-school components and further partitions the within-school variance into child-level and test-level variance. The child-level variance is true within-school variation that reflects skill differences among children attending the same school. The test-level variation comes from random measurement error[8] – the difference between a test score and the true achievement level of the child who took the test. Note that test-level measurement error accounts for more of the variance in learning and impact than in achievement; this is because estimates of learning are affected by measurement errors on two tests, while estimates of impact are affected by measurement errors on three tests.

Compared to learning and impact, achievement levels have a larger percentage of variance lying between rather than within schools. Specifically, the school level accounts for 23% of the variance in achievement levels, just 16–17% of the variance in 9- or 12-month learning rates, and only 8% of the variance in impact.

The fraction of variance that lies between schools affects the reliability of school means.[9] Specifically, if n children are sampled per school, then the reliability of the school mean is $p/[p + (1–p)/n]$, where p is the proportion of variance that lies at the school level. For example, if the school achievement levels are estimated by averaging test scores for $n = 20$ students per school, then, since $p = 23\%$ of the variance in test scores lies between rather than within schools, the school means will be 86% reliable in the sense that 86% of the variation in school means reflects true between-school variation, whereas 14% of the variation reflects contamination by within-school variation and the child and test level. Similarly, with $n = 20$ students per school, school means for 9- and 12-month learning rates would be 79–81% reliable, and school means for impact would be 65% reliable.

Note that differences in reliability matter less if the number of students sampled per school is reasonably large. As the within-school sample size n increases, all measures converge toward 100% reliability, so that the differences between measures matter less and less (Figure 2). A school accountability system would rarely rely on fewer than $n = 20$ children per school,[10] and at $n = 20$, average learning rates are just 5–7% less reliable than average achievement levels.

Validity versus reliability: evaluating the tradeoff

As we remarked earlier, we face a tradeoff. As a measure of school effectiveness, learning rates are more valid than achievement levels, and impact may (or may not) be more valid than learning rates. Yet learning rates are also less reliable than achievement levels, and impact is less reliable still.

As we showed in Figure 2, differences in reliability are small when a reasonable number of children are sampled in each school. This in itself suggests that reliability is a relatively minor issue compared to validity. To evaluate the tradeoff between validity and reliability more formally, though, it would be ideal if we could compare the available measures to a "gold standard" that perfectly reflects true school quality. The measure that correlates best with the gold standard would be the winning measure of school effectiveness. Unfortunately, no such gold standard exists.

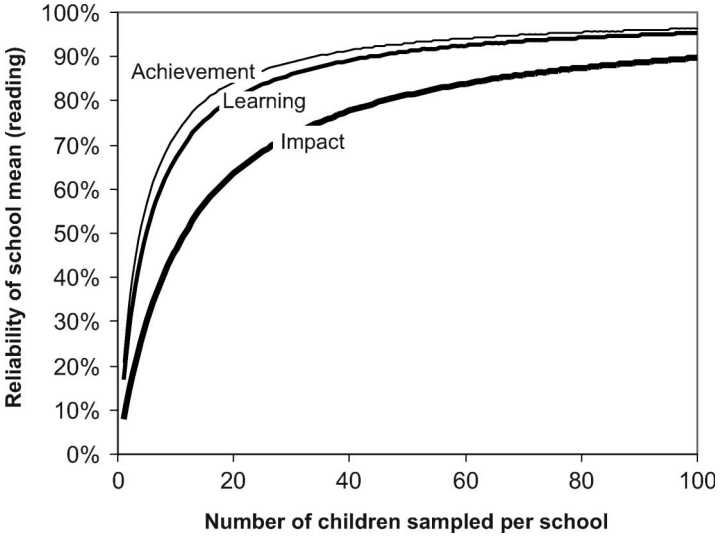

Figure 2. Reading: reliability of achievement, learning, and impact.

Although there is no gold standard, we can nevertheless estimate the correlation between the available measures and some kind of "bronze standard". A bronze standard is not a perfect index of school quality, but it is still better than the available measures.

For example, we might propose that a school's average learning rate would be a bronze standard, if only learning could be estimated without error. That is, there is no doubt we would prefer learning over raw achievement levels if it were not for the issue of reliability. Alternatively, if it were not for concerns about reliability, we might hold up impact as a bronze standard.

Although we cannot observe learning or impact without error, we can estimate the correlations among such error-free quantities. Our multilevel growth model provides an estimate of the school-level correlations between a school's achievement levels, learning rates, and impact, and these correlations are *latent*, in the sense that they are the correlations that *would* be observed if only achievement, learning, and impact could be measured with perfect reliability (Raudenbush & Bryk, 2002, chapter 11). Since we know the reliability of each measure, we can also estimate the extent to which the correlations will be attenuated when we have to use unreliable observed measures in place of their perfectly reliable latent counterparts. Details of the calculations are given in Appendix 1.

To start with a simple example, suppose we want the correlation between the true value and the observed value of a school's average first-grade learning rate. The correlation between the true and estimated values of a variable is just the square root of that variable's reliability (e.g., Bollen, 1989). With 20 children per school, the reliability of a school's average first-grade learning rate is 79% (Table 1), so the school-level correlation between true learning and estimated learning is $\sqrt{.79} = .89$. Likewise, the school-level correlation between true and estimated achievement levels is. 93, and the school-level correlation between true impact and estimated impact is .80.

By extending this calculation (as described in Appendix 1), we cannot only estimate the correlation between true and estimated values of the same variable – we can also estimate the correlation between true and estimated values of *different*

variables. For example, we can estimate the school-level correlation between true learning and estimated achievement level. Tables 2a and 2b give school-level correlations between the true and estimated values for different measures of school effectiveness. Table 2a gives the correlations when $n = 20$ children are sampled per school, and Table 2b gives the asymptotic correlations that would be observed if the number of children per school were infinite. The asymptotic correlations are the same as the latent school-level correlations that would be observed if estimates of school-level achievement, learning, and impact were not disturbed by within-school variation. Since the asymptotic correlations are not attenuated by random disturbances, they are slightly larger than the observed correlations when there are $n = 20$ children per school. But the general pattern of results is the same.

Using the information in Tables 2a and 2b, if we take a school's *true* first-grade learning rate as a bronze standard, then the measure that correlates best with this bronze standard is the school's *estimated* first-grade learning rate. Specifically, when $n = 20$ children are sampled per school the estimated learning rate has a .89 school-level correlation with the true learning rate. No other estimate has more than a .84 correlation with true learning. Alternatively, if we take a school's true *impact* as a bronze standard, then the measure that correlates best with the bronze standard is the *estimated* impact. The school-level correlation between true impact and estimated impact is .80, whereas no other measure has more than a .61 correlation with true impact. Similar reasoning shows that the best indicator of a school's true achievement level is its estimated achievement level – but this is a less useful result since, even if achievement could be observed without error, it would not be a bronze standard since it is clearly a far-from-valid measure of school effectiveness.

In short, the best indicator of a school's true learning rate is its estimated learning rate, and the best indicator of a school's true impact is its estimated impact. This may sound like common sense, but it was not a foreordained conclusion. For example, if achievement levels were *much* more reliable than learning rates, and the latent school-level correlation achievement and learning was *very* strong, it might have turned out that the best indicator of a school's true learning rate is its estimated achievement level. In fact, though, learning rates are not terribly unreliable and the latent correlation between achievement and learning is not overwhelmingly strong. So estimated achievement turns out to be a poor indicator of learning.

A secondary question is this: If the *best* indicator of school effectiveness is for some reason unavailable, then what would be the second- or third-best indicator? For example, if we lack the seasonal data to estimate 9-month school-year learning, would 12-month learning be a reasonable proxy? Table 2a provides an answer, showing that, when $n = 20$ children are sampled per school, estimated 12-month learning has a .84 school-level correlation with true first-grade learning – and no other measure correlates with true first-grade learning nearly as well. So if first-grade learning were unavailable, the best available proxy would be 12-month learning. Similar reasoning shows that if impact were considered the best measure of school quality, but impact were unavailable, the best proxy would be first-grade learning. However, estimated first-grade learning has just a .61 correlation with true impact, so even though first-grade learning is the best proxy available for impact, learning is not a good proxy for impact in an absolute sense.

Note that estimated achievement levels are not even a second-best indicator of true learning or true impact. With $n = 20$ children per school, estimated end-of-first-grade achievement levels have just a .54 school-level correlation with true 12-month learning,

Table 2a. Reading: correlations between true school means and estimated school means when 20 children are sampled per school.

	True school mean					
Estimated school mean	Achievement, end of first grade	Kindergarten learning	Summer learning	First-grade learning	12-month learning	Impact
Achievement	**0.93*** (0.91,0.94)**	0.37*** (0.23,0.50)	0.18^ (−0.02,0.38)	0.48*** (0.37,0.60)	0.54*** (0.44,0.64)	0.15 (−0.04,0.33)
Kindergarten learning	0.35*** (0.22,0.48)	**0.89*** (0.86,0.91)**	−0.27** (−0.46, −0.08)	−0.17* (−0.33, −0.01)	−0.26*** (−0.41, −0.11)	0.10 (−0.10,0.29)
Summer learning	0.15^ (−0.01,0.32)	−0.24** (−0.41, −0.07)	**0.78*** (0.72,0.85)**	−0.11 (−0.29,0.07)	0.16^ (−0.01,0.33)	−0.64*** (−0.74, −0.53)
First grade learning	0.46*** (0.35,0.57)	−0.17* (−0.33, −0.01)	−0.12 (−0.32, 0.07)	**0.89*** (0.87,0.91)**	0.84*** (0.80,0.88)	0.61*** (0.50,0.72)
12-month learning	0.52*** (0.42,0.62)	−0.26*** (−0.41, −0.11)	0.19^ (−0.01,0.38)	0.84*** (0.81,0.88)	**0.90*** (0.88,0.92)**	0.35*** (0.18,0.53)
Impact	0.13 (−0.03,0.29)	0.09 (−0.09,0.27)	−0.66*** (−0.76, −0.55)	0.55*** (0.44,0.66)	0.32*** (0.16,0.47)	**0.80*** (0.75,0.86)**

Table 2b. Reading: correlations between true school means and estimated school means when an infinite number of children are sampled per school (asymptotic correlations).

	True school mean					
Estimated school mean	Achievement, end of first grade	Kindergarten learning	Summer learning	First grade learning	12-month learning	Impact
Achievement	**1.00**	0.40***(0.25,0.54)	0.19† (−0.02,0.40)	0.52*** (0.40,0.64)	0.58*** (0.48,0.69)	0.16 (−0.04,0.36)
Kindergarten learning	0.40*** (0.25,0.54)	**1.00**	−0.30** (−0.52, −0.09)	−0.19* (−0.37, −0.01)	−0.29*** (−0.46, −0.13)	0.11 (−0.11,0.33)
Summer learning	0.19† (−0.02,0.40)	−0.30** (−0.52, −0.09)	**1.00**	−0.14 (−0.36,0.08)	0.21† (−0.01,0.42)	−0.82*** (−0.90, −0.74)
First grade learning	0.52*** (0.40,0.64)	−0.19* (−0.37, −0.01)	−0.14 (−0.36,0.08)	**1.00**	0.94*** (0.91,0.97)	0.68*** (0.57,0.80)
12-month learning	0.58*** (0.48,0.69)	−0.29*** (−0.46, −0.13)	0.21† (−0.01,0.42)	0.94*** (0.91,0.97)	**1.00**	0.39*** (0.20,0.58)
Impact	0.16 (−0.04,0.36)	0.11 (−0.11,0.33)	−0.82*** (−0.90, −0.74)	0.68*** (0.57,0.80)	0.39*** (0.20,0.58)	**1.00**

$^{\dagger}p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$. Parentheses enclose 95% confidence intervals.

Note: For the calculations that produced this correlation matrix, see Appendix 1. In Table 2a, note that the diagonal elements are the square roots of the reliabilities in Table 1. Note also that Table 2a is not symmetric, since the correlation between (say) true impact and estimated achievement is not the same as the correlation between *estimated* impact and *true* achievement. Table 2b, by contrast, is symmetric since, with an infinite number of students per school, there is no difference between the true school means and the estimated school mean.

730 725 720 715 710 705 700 695 690

just a .48 correlation with true first-grade learning, and a mere .15 correlation with true impact. These correlations are not large at all; with correlations of this magnitude, a school from the top achievement quintile has less than a 50% chance of being in the top quintile on learning or impact (Downey et al., 2008). In short, if we are really interested in learning or impact, achievement levels are not even a decent proxy.

External validity

Although this paper has focused primarily on the question of content validity, the validity of the impact measure may also be evaluated by comparing impact to other measures of school effectiveness. We have already seen that impact is positively correlated with school-year and 12-month learning rates (Tables 2a and 2b), and it is also comforting to note that a school's impact on reading is positively correlated with its impact on mathematics. (The observed correlation between estimated reading impact and estimated mathematics impact is 0.4, and the latent correlation between the underlying constructs is 0.6.) Future research should compare the impact measure to alternatives such as learning or achievement corrected for socioeconomic status, current achievement corrected for past achievement, or regression-discontinuity measures that compare the achievement of same-aged students in different grades (Cahan & Davis, 1987; Luyten, 2006). Even if some of these measures are less valid than others, any measure with a reasonable amount of validity ought to have a positive correlation with the others.

Conclusion: why it is better to be valid than reliable

When choosing a school evaluation criterion, it appears that validity trumps reliability. Not only is learning a more valid criterion than achievement, but the increase in validity more than compensates for the loss in reliability. Likewise, if we believe that impact is a more valid criterion than learning, then, again, the increase in validity trumps the loss in reliability.

The results support the view that there is no good reason to evaluate schools using average achievement levels. Although average achievement levels are reliable, they are not valid and they are not even good proxies for measures that are valid. School evaluation systems should rely on measures of learning, favoring 9-month learning, when available, over 12-month learning that includes summer gains and losses over which the school has little control. In addition, when both summer and school-year learning rates are available, evaluators should consider taking the difference between school-year and summer learning to evaluate schools on the basis of seasonal impact.

Acknowledgements

This research was supported in part by a grant from the Spencer Foundation to Douglas B. Downey and Paul T. von Hippel. I thank Doug Downey and Pam Paxton for helpful comments.

Notes

1. Sanders and Horn (1998) suggest that the TVAAS is not based on achievement gains. What they mean, though, is that the statistical method used in TVAAS does more than simply subtract one score from another. Although the method of estimation is more sophisticated under TVAAS, however, the quantity being estimated is still the number of points gained in a calendar year. (The analyses in this paper, incidentally, use a mixed model very much like the one used under TVAAS).

2. We are a little puzzled by the objection that giving two achievement tests a year is an excessive burden. Most children take dozens of tests every year as part of their ordinary coursework.
3. Some of the popularity of the TVAAS system may stem from the claim that learning rates estimated by TVAAS do not need adjustment since they are unrelated to race and socioeconomic status (Sanders, 1998; Sanders & Horn, 1998). Unfortunately, this claim is incorrect (Downey et al., 2008; Kupermintz, 2002).
4. As the fraction approaches zero, the impact measure becomes a simple measure of 9-month learning. It is hard to know the right fraction to subtract, so school evaluators might try different fractions in a sensitivity analysis of school effectiveness. An initially attractive possibility is to estimate the fraction by regressing the school-year learning rate on the summer learning rate. But since the correlation between school-year and summer learning can be *negative* (see Tables 2a and 2b), the fraction to subtract would be negative as well, yielding an impact measure that is a weighted sum rather than a weighted difference of school and summer learning rates.
5. Our analytic technique, multilevel modeling, requires that each unit from the lower level (each child) remains nested within a single unit from the higher level (a school). Data that violate this assumption may be modeled using a cross-classified model, but such models are very hard to fit when the data are large and the model is complicated.
6. Imputation of school dates was improved by a survey question where parents were asked to recall dates for the end of kindergarten and the beginning of first grade. Although parents' memories were fallible, their guesses, averaged up to the school level, helped to improve the imputation of school dates. Technically, what we imputed was not the school dates per se but a set of exposure variables measuring the time elapsed between various school dates (e.g., the first day of kindergarten) and the date of each test. Since the exposure variables are a linear combination of school dates and test dates, imputing exposures is equivalent to imputing school dates. The exposures were more directly useful, though, since they were used directly in the multilevel model described in Appendix 1.
7. Dummy variables such as school location and school sector are clearly nonnormal, but this is not important, since those variables have no missing values (Schafer, 1997).
8. The ECLS-K uses exceptionally well-designed tests in which random measurement error accounts for only about 5% of the variation in test scores (Rock & Pollack, 2002). Some of the tests used in state and municipal accountability systems may be less reliable than this. However, our basic results hold even for tests with 40% measurement error.
9. This discussion uses the following formula: If n children are sampled per school, then the reliability of the school mean is $p/[p + (1–p)/n]$, where p is the proportion of variance that lies at the school level.
10. The No Child Left Behind law requires that averages be reported for disadvantaged subgroups. Thus, in some instances, the number of students evaluated by a measure may be smaller than 20. However, states are able to set the threshold below which evaluating a subgroup would be unreliable, and this threshold is often set to 30 students or more.

Notes on contributor

Paul von Hippel has published research demonstrating that school attendance prevents childhood obesity (*American Journal of Public Health*, 2007, with Brian Powell, Douglas Downey, and Nicholas Rowland), that schools reduce inequality in cognitive skills (*American Sociological Review*, 2004, with Douglas Downey and Beckett Broh), and that year-round school calendars do not raise achievement or shrink achievement gaps (American Sociological Association, 2006).

References

Agresti, Al. (2002). *Categorical data analysis*. New York: Wiley.

Alexander, K., Entwisle, D.R., & Olson, L.S. (2001). Schools, achievement, and inequality: A seasonal perspective. In G. Bohrman & M. Boulay (Eds.), *Summer learning: Research, policies, and programs* (pp. 25–52). Mahwah, NJ: Erlbaum.

Alexander, K.L., Entwisle, D.R., & Olson, L.S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review*, *72*, 167–180.

Allison, P.D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Apted, M. (1963). *Seven Up!* [Film documentary]. London, UK: Granada Television.
Black, S. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics*, *114*(2), 577–599.
Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.
Cahan, S., & Davis, D. (1987). A between-grades-level approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal*, *24*(1), 1–12.
Chatterji, M. (2002). Models and methods for examining standards-based reforms and accountability initiatives: Have the tools of inquiry answered pressing questions on improving schools? *Review of Educational Research*, *72*(3), 345–386.
Clotfelter, C.T., & Ladd, H.F. (1996). Recognizing and rewarding success in public schools. In H.F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 23–64). Washington, DC: Brookings Institution.
Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., et al. (1966). *Equality of educational opportunity*. Washington, DC: US Government Printing Office.
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*, 613–635.
Downey, D.B., von Hippel, P.T., & Hughes, M. (2008). Are ''failing'' schools really failing? Using seasonal comparisons to evaluate school effectiveness. *Sociology of Education*, *81*(3), 242–270.
Dumay, X., & Dupriez, V. (2007). Accounting for class effect using the TIMSS 2003 eighth-grade database: Net effect of group composition, net effect of class process, and joint effect. *School Effectiveness and School Improvement*, *18*, 383–408.
Entwisle, D.R., & Alexander, K.L. (1992). Summer setback: Race, poverty, school composition and math achievement in the first two years of school. *American Sociological Review*, *57*, 72–84.
Fitz-Gibbon, C.T. (1996). *Monitoring education: Indicators, quality, and effectiveness*. London/New York: Cassell.
Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.
Harville, D. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer.
Heyneman, S.P., & Loxley, W.A. (1983). The effect of primary school quality on academic achievement scores across twenty-nine low and high income countries. *American Journal of Sociology*, *88*, 1162–1194.
Heyns, B. (1978). *Summer learning and the effects of schooling*. New York: Academic Press.
Hofferth, S.L., & Sandberg, J.F. (2001). How American children spend their time. *Journal of Marriage and the Family*, *63*(2), 295–308.
Holland, P.W., & Rubin, D.B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 3–25). Hillsdale, NJ: Lawrence Erlbaum Associates.
Johnson, D. (2005). Two-wave panel analysis: Comparing statistical methods for studying the effects of transitions. *Journal of Marriage and Family*, *67*(4), 1061–1075.
Johnson, R.A., & Wichern, D.W. (1997). *Applied multivariate statistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
Kane, T.J., & Staiger, D.O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 235–269). Washington, DC: Brookings Institution.
Kupermintz, H. (2002). Value-added assessment of teachers: The empirical evidence. In A. Molar (Ed.), *School reform proposals: The research evidence* (pp. 217–234). Greenwich, CT: Information Age Publishing.
Ladd, H.F. (2002). Comment [on Kane and Staiger 2002]. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 273–283). Washington, DC: Brookings Institution.
Ladd, H.F., & Walsh, R.P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review*, *21*, 1–27.
Lareau, A. (2000). *Home advantage: Social class and parental intervention in elementary education*. Oxford, UK: Rowman and Littlefield.
Lee, V.E., & Burkam, D.T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.
Linver, M.R., Brooks-Dunn, J., & Kohen, D.E. (2002). Family processes as pathways from income to young children's development. *Developmental Psychology*, *38*, 719–734.

Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*(420), 1227–1237.
Lord, F.M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305.
Luyten, H. (2006). An empirical assesment of the absolute effect of schooling: Regression discontinuity applied to TIMSS 95. *Oxford Review of Education*, *32*(2), 397–429.
Meyer, R.H. (1996). Value-added indicators of school performance. In E.A. Hanushek & D.W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197–223). Washington, DC: National Academy Press.
Miller, S.K. (1983, April). *The history of effective schools research: A critical overview*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada. (ERIC Document Reproduction Service No. ED231818.).
Opdenakker, M.-C., & Van Damme, J. (2006). Differences between secondary schools: A study about school context, group composition, school practice, and school effects with special attention to public and Catholic schools and types of schools. *School Effectiveness and School Improvement*, *17*, 87–117.
Opdenakker, M.-C., & Van Damme, J. (2007). Do school context, student composition and school leadership affect school practice and outcomes in secondary education? *British Educational Research Journal*, *33*, 179–206.
Raudenbush, S.W., & Bryk, A.S. (2002). *Hierachical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
Rock, D.A., & Pollack, J.M. (2002). *Early childhood longitudinal study – Kindergarten class of 1998–99 (ECLS-K), psychometric report for kindergarten through first grade*. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved May 31, 2007, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200205
Rubenstein, R., Stiefel, L., Schwartz, A.E., & Amor, H.B.H. (2004). Distinguishing good schools from bad in principle and practice: A comparison of four methods. In W.J. Fowler, Jr. (Ed.), *Developments in school finance: Fiscal proceedings from the Annual State Data Conference of July 2003* (pp. 55–70). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Government Printing Office (NCES 2004–325).
Rubin, D.B. (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.
Sammons, P., Mortimore, P., & Thomas, S. (1996). Do schools perform consistently across outcomes and areas? In J. Gray, D. Reynolds, C. Fitz-Gibbon, & D. Jesson (Eds.), *Merging traditions: The future of research on school effectiveness and school improvement* (pp. 3–29). London: Cassell.
Sanders, W.L. (1998). Value-added assessment. *The School Administrator*, *55*(11), 24–32.
Sanders, W.L., & Horn, J.P. (1998). Research findings from the Tennessee Value Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247–256.
Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
Scheerens, J. (1992). *Effective schooling: Research, theory, and practice*. London: Cassell.
Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.
Schemo, D.J. (2004, March 25). 14 states ask U.S. to revise some education law rules. *New York Times*.
Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research: An international survey of research on school effectiveness*. London: Falmer Press.
Teddlie, C., Reynolds, D., & Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research: An international survey of research on school effectiveness* (pp. 55–133). London: Falmer Press.
Thrupp, M. (1999). *Schools making a difference. Let's be realistic*. Buckingham, UK: Open University Press.
von Hippel, P.T. (2007). Regression with missing ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*(1), 83–117.
Willett, J.B. (1988). Questions and answers in the measurement of change. In E.Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–422). Washington, DC: American Education Research Association.
Willms, J.D. (1992). *Monitoring school performance: A guide for educators*. London: Falmer Press.

Appendix 1. Statistical methods

Multilevel growth model: specification and estimation

As mentioned in the text, we modeled achievement and learning rates by fitting a multilevel growth model, in which tests (Level 1) were nested within children (Level 2) and children were nested within schools (Level 3). Specifically, at Level 1, we modeled each test score Y_{ics} as a linear[1] function of the months that child c in school s had been exposed to KINDERGARTEN, SUMMER, and FIRST GRADEat the time of test i:[2]

$$Y_{ics} = \alpha_{0cs} + \alpha_{1cs}\,\text{KINDERGARTEN}_{ics} + \alpha_{2cs}\,\text{SUMMER}_{ics} + \alpha_{3cs}\ \text{FIRST GRADE}_{ics} + e_{ics}$$

or more concisely

$$Y_{ics} = \mathbf{EXPOSURES}_{ics}\,\alpha_{cs} + e_{ics} \tag{1}$$

where $\alpha_{cs} = [\alpha_{0cs}\,\alpha_{1cs}\,\alpha_{2cs}\,\alpha_{3cs}]^{\mathrm{T}}$ and $\mathbf{EXPOSURES}_{ics} = [1\ \text{KINDERGARTEN}_{ics}\,\text{SUMMER}_{ics}\,\text{FIRST GRADE}_{ics}]$. The EXPOSURES variables are maximum-centered[3] so that the intercept α_{0cs} represents the child's achievement level on the last day of first grade.[4] The slopes α_{1cs}, α_{2cs}, and α_{3cs} represent monthly learning rates during kindergarten, summer, and first grade. The residual term e_{ics} is measurement error, or the difference between the test score Y_{ics} and the child's true achievement level at the time of the test. The variance of the measurement error can be calculated from test-reliability estimates in Rock and Pollack (2002); for calculations, see Table 1.

At Level 2, the child-level coefficient vector α_{cs} is modeled as the sum of a school-level coefficient vector $\beta_s = [\beta_{0s}\,\beta_{1s}\,\beta_{2s}\,\beta_{3s}]^{\mathrm{T}}$ plus a child-level random effect vector $\boldsymbol{a}_{\mathrm{c}} = [a_{0c}\ a_{1c}\ a_{2c}\ a_{3c}]^{\mathrm{T}}$.

$$\alpha_{cs} = \beta_s + a_c \tag{2}$$

At Level 3, the school-level coefficient vector β_s is modeled as the sum of a fixed effect $\gamma_0 = [\gamma_{00}\ \gamma_{10}\ \gamma_{20}\ \gamma_{30}]^{\mathrm{T}}$ plus a school-level random effect $\boldsymbol{b}_{\mathrm{s}} = [b_{0s}\ b_{1s}\ b_{2s}\ b_{3s}]^{\mathrm{T}}$:

$$\beta_s = \gamma_0 + \boldsymbol{b}_s \tag{3}$$

Note that Equations (2) and (3) may be combined into a single mixed-level equation:

$$\alpha_{cs} = \gamma_0 + \boldsymbol{b}_s + \boldsymbol{a}_c \tag{4}$$

where γ_0 is a fixed effect representing the grand mean for achievement and learning rates; $\boldsymbol{b}_{\mathrm{s}}$ is a school-level random effect representing the departure of school s from the grand mean; and $\boldsymbol{a}_{\mathrm{c}}$ is a child-level random effect representing child c's departure from the mean for school s. The

Table 1. Measurement error variance on four reading tests and four mathematics tests.

	Reading			Mathematics		
Occasion (i)	Total variance	Reliability	Measurement error variance	Total variance	Reliability	Measurement error variance
1. Fall 1998	73.62	0.93	5.15	50.55	0.92	4.04
2. Spring 1999	117.72	0.95	5.89	76.39	0.94	4.58
3. Fall 1999	160.53	0.96	6.42	92.35	0.94	5.54
4. Spring 2000	200.79	0.97	6.02	90.25	0.94	5.42

Note: Reliabilities were calculated by Rock and Pollack (2002) using Item Response Theory. In psychometric terms, the reliability of a test is the fraction of test-score variance that represents true variation in skill; any additional variance is just measurement error. So, if the reliability is r and the total variance of a test is $Var\ (Y_{sct})$, then the measurement error variance is $(1–r)\ Var\ (Y_{sct})$. Note that the variance changes (though not by much) from one measurement occasion to the next. Our analyses account for this heteroscedasticity, but ignoring it would yield very similar results.

random effects $\boldsymbol{a}_c$ and $\boldsymbol{b}_s$ are assumed to be independent multinormal variables with covariance matrices of Σ_a and Σ_b. For certain purposes, it will be convenient to work with $vech(\Sigma_a)$ and $vech(\Sigma_b)$, which are vectors containing all the nonredundant elements of Σ_a and Σ_b – for example, $vech(\Sigma_a)$ is a vector containing the lower triangle of Σ_b, beginning with the first column[5] (Harville 1997).

Multilevel modeling software (such as the MIXED procedure in SAS) provides point estimates $\hat{\gamma}_0, \hat{\gamma}_1$ and $\hat{\Sigma}_b$, as well as asymptotic covariance matrices $V(\hat{\gamma}_0)$, $V(\hat{\gamma}_1)$, and $V(vech(\hat{\Sigma}_b))$ that represent the uncertainty in the point estimates. (The diagonal elements of these covariance matrices are squared standard errors.)

Combining these estimates to obtain estimates of 12-month learning and impact requires some transformation. As remarked in the main text, the impact of school s is the difference between the first-grade learning rate and the summer learning rate; that is, impact is $\beta_{4s} = \beta_{3s} - \beta_{2s}$, or $\beta_{5s} = \mathbf{c}_{impact}\ \beta_s$, where $\mathbf{c}_{impact} = [0\ 0\ -1\ 1]$. Likewise, the 12-month learning rate in school s is the average monthly learning rate over a 12-month period consisting (on average) of 2.4 months of summer followed by 9.6 months of first grade; that is, 12-month learning is $\beta_{5s} = \frac{1}{12}(2.4\beta_{2s} + 9.6\beta_{3s})$ or, in vector form, $\beta_{5s} = \mathbf{c}_{12month}\ \beta_s$ where $\mathrm{c}_{12month} = \frac{1}{12}[0 \quad 0 \quad 2.4 \quad 9.6]$. So, if we let

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_4 \\ \mathbf{c}_{12\mathrm{month}} \\ \mathbf{c}_{\mathrm{impact}} \end{bmatrix} \tag{5}$$

where $\mathbf{I}_4$ is the 4-by-4 identity matrix, then $\beta_s^* = \mathbf{C}\,\beta_s = [\beta_{0s}\,\beta_{1s}\,\beta_{2s}\,\beta_{3s}\,\beta_{4s}\,\beta_{5s}]^{\mathrm{T}}$ is an expanded school-level vector that includes impact and 12-month learning as well as achievement, kindergarten learning, summer learning, and first grade learning. The following equation represents how this vector varies across schools:

$$\beta_s^* = \gamma_0^* + \boldsymbol{b}_s^* \tag{6}$$

where $\gamma_0^* = \mathbf{C}\gamma_0$ and $\gamma_1^* = \mathbf{C}\gamma_1$ are the fixed intercept and slope, and the random effect $\mathbf{b}_s^*$ has a covariance matrix of $\Sigma_b^* = \mathbf{C}\Sigma_b\,\mathbf{C}^{\mathrm{T}}$. Estimated parameters for this expanded Equation (6) can be derived from the estimates for the basic Equation (3), as follows: $\hat{\gamma}_0^* = \mathbf{C}\hat{\gamma}_0, \hat{\gamma}_1^* = \mathbf{C}\hat{\gamma}_1$, and $\hat{\Sigma}_b^* = \mathbf{C}\hat{\Sigma}_b\mathbf{C}^{\mathrm{T}}$, or $vech(\hat{\Sigma}_b^*) = \mathbf{F}vech(\hat{\Sigma}_b)$, where $\mathbf{F} = \mathbf{H}_6(\mathbf{C}\otimes\mathbf{C})\mathbf{G}_4$, with $\mathbf{G}_4$ a duplication matrix and $\mathbf{H}_6$ an inverse duplication matrix.[6] The asymptotic covariance matrices for these transformed parameter estimates are $V(\hat{\gamma}_0^*) = \mathbf{C}V(\hat{\gamma}_0)\mathbf{C}^{\mathbf{T}}$, $V(\hat{\gamma}_1^*) = \mathbf{C}V(\hat{\gamma}_1)\mathbf{C}^{\mathbf{T}}$, and $V(vech(\hat{\Sigma}_b^*)) = \mathbf{F}V(vech(\hat{\Sigma}_b))\mathbf{F}^{\mathrm{T}}$.[7]

The final step in our calculations is to convert the variances and covariances in Σ_b^* into standard deviations and correlations, which are easier to interpret. This is straightforward; a standard deviation σ is just the square root of the corresponding variance σ^2, and there is a simple matrix formula $\mathbf{R}_b^* = \mathbf{R}(\Sigma_b^*)$ for converting a covariance matrix such as Σ_b^* into a correlation matrix $\mathbf{R}_b^*$ (Johnson & Wichern, 1997). Again, it will be convenient to work with *vecp* ($\mathbf{R}_b^*$), which is a vector containing the nonredundant elements of $\mathbf{R}_b^*$ – that is, the lower triangle of $\mathbf{R}_b^*$ excluding the diagonal, starting with the first column (Harville, 1997).

Standard errors for the standard deviations and correlations that result from these calculations can be obtained using the delta rule (e.g., Agresti, 2002, section 14.1.3). For example, if $\hat{V}(\hat{\sigma}^2)$ is the squared standard error for the variance estimate $\hat{\sigma}^2$, then $\hat{V}(\hat{\sigma}) = (\frac{d\hat{\sigma}}{d\hat{\sigma}^2})^2\hat{V}(\hat{\sigma}^2) = \frac{1}{4\hat{\sigma}^2}\hat{V}(\hat{\sigma}^2)$ is the squared standard error for the standard deviation estimate $\hat{\sigma}$. Likewise, if $V(vech(\hat{\Sigma}_b^*))$ is the asymptotic covariance matrix of $vech(\hat{\Sigma}_b^*)$, then

$$V(vecp(\hat{\mathbf{R}}_b^*)) = \left[\frac{dvecp(\mathbf{R}(\hat{\Sigma}_b^*))}{dvech(\hat{\Sigma}_b^*)}\right] V(vecp(\hat{\Sigma}_b)) \left[\frac{dvecp(\mathbf{R}(\hat{\Sigma}_b^*))}{dvech(\hat{\Sigma}_b^*)}\right] \tag{7}$$

is the asymptotic covariance matrix of $vecp(\mathbf{R}_b^*)$.

Calculating reliability

To calculate reliability, we will need not only the school-level and child-level covariance matrices, but the test-level covariance matrix as well. To simplify the calculations, let us assume that, in a

well-designed school-evaluation system, children would be tested on the first and last day of kindergarten and first grade – that is, not on the staggered schedule used by the Early Childhood Longitudinal Study. Under those assumptions, let the kindergarten and first-grade test scores be $\mathbf{Y}^*_{cs} = [Y_{1cs}\ Y_{2cs}\ Y_{3cs}\ Y_{4cs}]^{\mathrm{T}}$ with independent measurement errors $\mathbf{e}_{cs} = [e_{1cs}\ e_{2cs}\ e_{3cs}\ e_{4cs}]^{\mathrm{T}}$. Then $\mathbf{e}_{cs}$ has a diagonal covariance matrix Σ_e with diagonal elements from Table 1.

Now a child's end-of-first-grade achievement can be estimated by the end-of-first-grade test score Y_{4cs}, or

$$Y_{4cs} = \mathbf{d}_{ach}\mathbf{Y}^*_{cs}, \quad \text{where} \quad \mathbf{d}_{ach} = [\,0 \quad 0 \quad 0 \quad 1\,] \tag{8}$$

The child's monthly learning rate during kindergarten can be estimated by subtracting Test 1 from Test 2, and dividing by the length of the school year (9.4 months):

$$\tfrac{1}{9.4}(Y_{2cs} - Y_{1cs}) \text{ or } \mathbf{d}_{kind}\mathbf{Y}^*_{cs}, \text{ where } \mathbf{d}_{kind} = \tfrac{1}{9.4}[\,-1 \quad 1 \quad 0 \quad 0\,] \tag{9}$$

Likewise, the first-grade learning rate can be estimated by subtracting Test 3 from Test 4, and dividing by the length of the school year,

$$\tfrac{1}{9.4}(Y_{4cs} - Y_{3cs}) \text{ or } \mathbf{d}_{first}\mathbf{Y}^*_{cs}, \text{ where } \mathbf{d}_{first} = \tfrac{1}{9.4}[\,0 \quad 0 \quad -1 \quad 1\,] \tag{10}$$

and the summer learning rate can be estimated by subtracting Test 2 from Test 3 and dividing by the length of summer vacation (2.6 months):

$$\mathbf{d}_{summer}\mathbf{Y}^*_{cs}, \text{ where } \mathbf{d}_{summer} = \tfrac{1}{2.6}[\,0 \quad -1 \quad 1 \quad 0\,] \tag{11}$$

Combining these results, we let

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_{ach} \\ \mathbf{d}_{kind} \\ \mathbf{d}_{summer} \\ \mathbf{d}_{first} \end{bmatrix} \tag{12}$$

so that the 4-element vector $\mathbf{DY}^*_{cs}$ contains estimates for child c's end-of-first-grade achievement level, as well as seasonal learning rates for kindergarten, summer, and first grade. Then the 6-element vector $\mathbf{CDY}^*_{cs}$ contains the elements of $\mathbf{DY}^*_{cs}$, as well as estimates for 12-month learning and impact. So, if we want to estimate achievement, seasonal learning, 12-month learning, and impact, the test-level covariance matrix for these matrices is

$$\Sigma^*_e = (\mathbf{CD})\Sigma_e(\mathbf{CD})^{\mathrm{T}} \tag{13}$$

and the overall covariance matrix, including school, child, and test levels is

$$\Sigma^* = \begin{bmatrix} \Sigma^*_a & & \\ & \Sigma^*_b & \\ & & \Sigma^*_e \end{bmatrix} \tag{14}$$

Now, for a given measure of school performance, reliability depends on how much of the measure's variance lies between rather than within schools. For a single test score, the between-school variance is Σ^*_a and the within-school variance is $\Sigma^*_b + \Sigma^*_e$, so that the total variance is $\Sigma^*_b + \Sigma^*_b + \Sigma^*_e$. When test scores are averaged across n children per school, the between-school variance is Σ^*_a and the within-school variance is $\frac{1}{n}(\Sigma^*_b + \Sigma^*_e)$, so that the total variance is $\Sigma^*_a + \frac{1}{n}(\Sigma^*_b + \Sigma^*_e)$. So the reliability of a school average can be obtained by dividing the true variance in the true school-level mean (a diagonal element of Σ^*_a) by the variance in the estimated mean (the corresponding diagonal element of $\Sigma^*_a + \frac{1}{n}(\Sigma^*_b + \Sigma^*_e)$).

An equivalent definition is that reliability is the squared correlation between a true school-level value and its estimate. Again, the covariance matrix for the true school-level values is Σ^*_a,

while the covariance matrix for the estimated values is $\Sigma_a^* + \frac{1}{n}(\Sigma_b^* + \Sigma_e^*)$. The latter expression can be rewritten as

$$\Sigma_a^* + \tfrac{1}{n}(\Sigma_b^* + \Sigma_e^*) = \begin{bmatrix} \mathbf{I}_6 & \frac{1}{\sqrt{n}}\mathbf{I}_6 & \frac{1}{\sqrt{n}}\mathbf{I}_6 \end{bmatrix} \Sigma^* \begin{bmatrix} \mathbf{I}_6 \\ \frac{1}{\sqrt{n}}\mathbf{I}_6 \\ \frac{1}{\sqrt{n}}\mathbf{I}_6 \end{bmatrix} \tag{15}$$

where $\mathbf{I}_6$ is the 6-by-6 identity matrix. So the covariances between the true and estimated levels of achievement, learning, and impact are given by

$$\mathbf{K}\Sigma * \mathbf{K}^{\mathrm{T}}, \text{ where } \mathbf{K} = \begin{bmatrix} \mathbf{I}_6 & & \\ \mathbf{I}_6 & \frac{1}{\sqrt{n}}\mathbf{I}_6 & \frac{1}{\sqrt{n}}\mathbf{I}_6 \end{bmatrix} \tag{16}$$

Displayed in correlation form, part of this covariance matrix is given in Table 2.

Notes

1. The linearity assumption can be tested by taking advantage of the fact that different schools are tested at different times. At one school, the fall test might be given in October and the spring test in May, while at another school the fall test might be given in November and the spring test in April. In general, the points gained between fall and spring tests is proportional to the months elapsed between them – that is, school-year learning is approximately linear.
2. These exposures are estimated by comparing the test date to the first and last date of kindergarten and first grade. Test dates are part of the public data release; the first and last dates of the school year are available to researchers with a restricted-use data license.
3. Centering is often used to clarify the interpretation of the intercept (Raudenbush & Bryk, 2002). The most popular form of centering is mean-centering, but in this context we use maximum-centering so that the intercept α_{0cs} will represent the child's achievement level on the last day of first grade. To understand maximum-centering, let $\text{KINDERGARTEN}^*_{ics}$ be the number of months that child c in school s has spent in kindergarten at the time of test i. The maximum possible value of $\text{KINDERGARTEN}^*_{ics}$ is KINDLENGTH_s, which is the length of the kindergarten year in school s. (An average value would be $\text{KINDLENGTH}_s = 9.4$ months.) Then $\text{KINDERGARTEN}_{ics} = \text{KINDERGARTEN}^*_{ics} - \text{KINDLENGTH}_s$ is the maximum-centered value of kindergarten exposure. SUMMER_{ics} and FIRST GRADE_{ics} are maximum-centered as well. When $\text{KINDERGARTEN}_{ics}$, SUMMER_{ics}, and FIRST GRADE_{ics} are maximum-centered, the intercept α_{0cs} represents the child's score on the last day of first grade, when $\text{KINDERGARTEN}_{ics}$, SUMMER_{ics}, and FIRST GRADE_{ics} have all reached their maximum values of 0.
4. This is not the same as the final test score, because the final test was typically given one to three months before the end of first grade. Instead, the last-day achievement level is in effect an extrapolation to the score that would have been received had the test been given on the last day of first grade.
5. In SAS software, the vector form of a symmetric matrix is called SYMSQR and begins with the first row rather than the first column. The elements of SYMSQR (Σ_b) must be rearranged to obtain *vech* (Σ_b).
6. Duplication and inverse duplication matrices are defined in section 16.4*b* of Harville (1997). The relationship between $vech(\hat{\Sigma}_a)$ and $vech(\mathbf{C}\hat{\Sigma}_a\mathbf{C}^{\mathrm{T}})$ is given, using different notation, by formula 4.25 in Chapter 16. Formula 4.25 is restricted to the case where $\mathbf{C}$ is a square matrix; we use a generalization appropriate to the case where $\mathbf{C}$ is not square. We thank David Harville for suggesting this generalization (personal communication, September 27, 2005).
7. These formulas make use of the general formula that, if the vector $\boldsymbol{X}$ has mean μ and covariance matrix Σ, then the vector $\mathbf{A}\boldsymbol{X}$, where $\mathbf{A}$ is a matrix, has mean $\mathbf{A}\mu$ and covariance matrix $\mathbf{A}\Sigma\mathbf{A}^{\mathrm{T}}$ (Johnson & Wichern, 1997, p. 79).

Appendix 2. Mathematics tables

The main text of this paper focuses on results for reading. Results for mathematics, which were generally similar, are tabled below.

Table 1. Mathematics achievement, learning, and impact, as measured on a 64-point scale.

				Monthly learning rates				
			Achievement end of first grade	Kindergarten	Summer	first grade	12-month[a]	Impact[b]
Fixed effects	Mean		45.58***	1.34***	0.47***	1.57***	1.33***	1.10***
Random effects	Variances	Between school (school-level)	4.26^2 (21%)	0.28^2 (14%)	0.58^2 (8%)	0.26^2 (13%)	0.20^2 (13%)	0.71^2 (9%)
		Within school: Child-level	8.01^2 (73%)	0.61^2 (68%)	1.50^2 (55%)	0.59^2 (64%)	0.46^2 (65%)	1.80^2 (56%)
		Within school: Test-level	2.33^2 (6%)	0.31^2 (18%)	1.22^2 (37%)	0.35^2 (23%)	0.26^2 (22%)	1.44^2 (36%)
	Total variance		9.36^2	0.74^2	2.02^2	0.73^2	0.56^2	2.41^2
Reliability of school mean if 20 children are sampled per school[c]			84%	76%	64%	74%	75%	65%

$^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

[a]Twelve-month learning is reckoned from the end of kindergarten to the end of first grade.

[b]Impact is the difference between the first-grade and summer learning rates.

[c]If n children are sampled per school, then the reliability of the school mean is $p/[p + (1–p)/n]$, where p is the proportion of variance that lies at the school level.

Table 2a. Mathematics: correlations between true school means and estimates school means when 20 children are sampled per school.

	True school mean					
Estimated school mean	Achievement, end of first grade	Kindergarten learning	Summer learning	First-grade learning	12-month learning	Impact
Achievement	**0.92*** (0.90,0.93)**	0.40*** (0.26,0.54)	0.06 (−0.13,0.25)	0.10 (−0.06,0.26)	0.13^ (−0.02,0.29)	−0.01 (−0.20,0.17)
Kindergarten learning	0.38*** (0.25,0.52)	**0.87*** (0.84,0.91)**	−0.39*** (−0.55,−0.22)	−0.20* (−0.37,−0.02)	−0.44*** (−0.58,−0.30)	0.24** (0.06,0.42)
Summer learning	0.05 (−0.12,0.22)	−0.35*** (−0.50,−0.20)	**0.80*** (0.75,0.85)**	−0.25** (−0.42,−0.09)	0.24** (0.08,0.40)	−0.75*** (−0.81,−0.69)
First grade learning	0.09 (−0.06,0.24)	−0.19* (−0.37,−0.02)	−0.27** (−0.45,−0.10)	**0.86*** (0.83,0.89)**	0.70*** (0.62,0.78)	0.54*** (0.42,0.67)
12-month learning	0.13^ (−0.02,0.27)	−0.43*** (−0.57,−0.30)	0.26** (0.08,0.43)	0.70*** (0.62,0.78)	**0.86*** (0.83,0.90)**	0.05 (−0.14,0.23)
Impact	−0.01 (−0.17,0.15)	0.22** (0.06,0.39)	−0.76*** (−0.82,−0.69)	0.51*** (0.39,0.63)	0.04 (−0.13,0.22)	**0.81*** (0.76,0.86)**

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$. Parentheses enclose 95% confidence intervals.

Table 2b. Mathematics: correlations between true school means and estimates school means when an infinite number of children are sampled per school (asymptotic correlations).

	True school mean					
Estimated school mean	Achievement, end of first grade	Kindergarten learning	Summer learning	First-grade learning	12-month learning	Impact
Achievement	**1.00**	0.44*** (0.29,0.59)	0.07 (−0.12,0.27)	0.11 (−0.06,0.28)	0.15† (−0.01,0.32)	−0.02 (−0.21,0.17)
Kindergarten learning	0.44*** (0.29,0.59)	**1.00**	−0.44*** (−0.62,−0.26)	−0.22* (−0.42,−0.03)	−0.50*** (−0.65,−0.34)	0.28** (0.07,0.48)
Summer learning	0.07 (−0.12,0.27)	−0.44*** (−0.62,−0.26)	**1.00**	−0.31** (−0.51,−0.12)	0.30** (0.11,0.50)	−0.94*** (−0.96,−0.91)
First grade learning	0.11 (−0.06,0.28)	−0.22* (−0.42,−0.03)	−0.31** (−0.51,−0.12)	**1.00**	0.81*** (0.73,0.88)	0.63*** (0.49,0.76)
12-month learning	0.15† (−0.01,0.32)	−0.50*** (−0.65,−0.34)	0.30** (0.11,0.50)	0.81*** (0.73,0.88)	**1.00**	0.05 (−0.17,0.26)
Impact	−0.02 (−0.21,0.17)	0.28** (0.07,0.48)	−0.94*** (−0.96,−0.91)	0.63*** (0.49,0.76)	0.05 (−0.17,0.26)	**1.00**

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$. Parentheses enclose 95% confidence intervals.

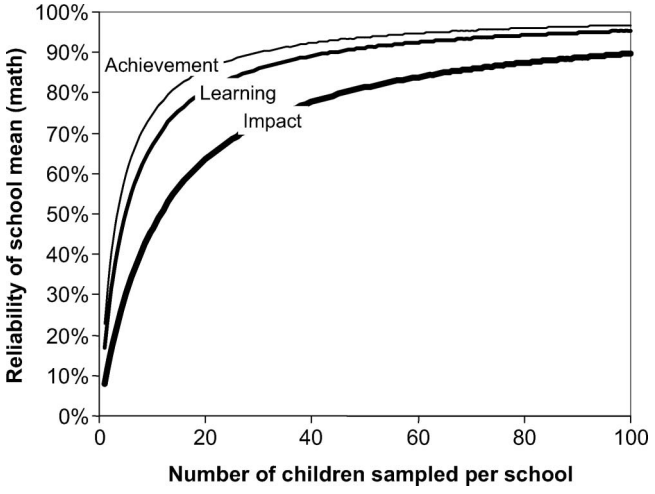

Figure 1. Mathematics: reliability of achievement, learning, and impact.