Translating Standardized Effects of Education Programs Into More Interpretable Metrics

Matthew D. Baird[1] and John F. Pane[1]

Evaluators report effects of education initiatives as standardized effect sizes, a scale that has merits but obscures interpretation of the effects' practical importance. Consequently, educators and policymakers seek more readily interpretable translations of evaluation results. One popular metric is the number of years of learning necessary to induce the effect. We compare years of learning to three other translation options: benchmarking against other effect sizes, converting to percentile growth, and estimating the probability of scoring above a proficiency threshold. After enumerating the desirable properties of translations, we examine each option's strengths and weaknesses. We conclude that years of learning performs worst, and percentile gains performs best, making it our recommended choice for more interpretable translations of standardized effects.

Keywords: education program effects; education policy; program evaluation; research utilization; standardized effect sizes; translation of research results; years of learning

When evaluators report the effects of education programs and policies on student outcomes, they typically use standardized effect sizes. This facilitates the comparability of results across studies, across programs or policies, and across outcome measures. But it hinders another important goal of research, to enable practitioners and policymakers to understand the real-world implications of the research results to their own work. Because standardized effect sizes are measured on an abstract scale—standard deviation units of the outcome measure—magnitudes are difficult to interpret. As an example, it may not be evident to consumers of research whether a program effect of 0.13 is meaningful enough that they should act to increase adoption of the program.

Consequently, there have been efforts to improve the usability of research results by translating them into more readily interpretable metrics. Research by Lipsey et al. (2012) is dedicated to this topic, offering several options for translating effects and discussing some advantages and disadvantages of each. However, the scope of discussion is more broad than deep—without a thorough exploration of tradeoffs—and refrains from normative statements about which translation is preferred. This article addresses that gap.

Translating to units of time, such as years of learning, has become a popular choice. We are often asked to perform this translation on our own study results, which we have resisted because of its undesirable properties. Nonetheless, others have gone ahead and translated our results (e.g., Childress & Amrofell, 2017, translating Pane, Steiner, Baird, & Hamilton, 2015). A goal of this article is to elucidate the translation's undesirable properties, most importantly that it is highly sensitive to the method of calculation and can produce implausible results. When a translation suffers from flaws such as these, it can obfuscate rather than clarify the research findings.

We investigate two research questions. First, among four translation options, which possess the best properties and should be pursued, and which possess the worst properties and should be avoided? Second, if an analyst is committed to a particular translation option, how sensitive are the results to how the translation is implemented, and which implementation should be preferred, if any?

For illustration, consider a treatment effect of 0.13 standard deviations. We examine translating this to (1) years or days of learning, e.g., 0.22 additional years of learning (*Years of Learning*); (2) benchmarking results against gaps between demographic groups, e.g., one fifth of the black/white achievement gap, or effect sizes measured in other studies, e.g., approximately the same effect as decreasing class size from 22 to 15 students

[1] RAND Corporation, Pittsburgh, PA

Educational Researcher, Vol. 48 No. 4, pp. 217–228
DOI: 10.3102/0013189X19848729
Article reuse guidelines: sagepub.com/journals-permissions

(*Benchmarking)*; (3) percentile change, e.g., 6.3 percentile point increase (*Percentiles)*; and (4) calculating the likelihood of scoring above a reference value, such as scoring proficient, e.g., a 3 percentage point increase in the probability of scoring proficient (*Thresholds)*.

As a framework for this study, we begin by enumerating six desirable properties for translations. Then, using an empirical data set, we examine the performance of each translation option, as well as its sensitivity to implementation alternatives.

We are not the first to caution against translating to units of time (Dadey & Briggs, 2012; Dorn, 2015; Maul & McLelland, 2013). However, the practice is still frequently requested and implemented, so we feel it is important to summarize the critiques in one place and add empirical insights regarding the limitations and available alternatives. Given the frequency of use, we also discuss preferences for how to implement years-of-learning translation, should it be used.

There are additional issues at play, not fully addressed in this article. We abstract from psychometric concerns regarding test scaling. In particular, years of learning requires an achievement measure reported on a continuous developmental scale; the creation, maintenance, and application of such scales pose psychometric challenges (Briggs, 2013; Dadey & Briggs, 2012; Martineau, 2006; Yen, 1986). There are also concerns about the ability of tests to measure the learning in higher grade levels; when students diversify course-taking it becomes more challenging for tests to accurately measure the full breadth of learning. We discuss one facet of this problem, but otherwise operate on the assumption that tests accurately capture learning. Thus, we start from assumptions commonly made in the wide breadth of program analysis investigating student achievement, namely that we have a measure that enables valid estimation of a program's effect. We focus only on how best to translate findings from the standardized effect scale to more easily interpreted measures.

Desirable Properties

The driving purpose of translation is to mitigate a challenge faced in using research results to guide policy and practice: how to interpret standardized effect sizes. Clearly, one desirable property is interpretability, but if the translation option has other problems, ease of interpretation may actually lead to misinterpretation, with consequent faulty decision-making. We investigate six desirable properties of effect size translations.

1. *Ease of interpretation.* A translation option is useful only if it is easier for practitioners and policymakers to interpret than the original standardized effect.
2. *Transparent and valid assumptions.* Assumptions underlying the translation should be plausible, if not formally validated, and clearly communicated with the translated results.
3. *Added statistical uncertainty is minimized and clearly conveyed.* The original treatment effect estimate has statistical uncertainty. Some of the translation options add additional uncertainty. Smaller uncertainty is preferable, and the uncertainty should be reported along with the translated result. In practice, this advice is often disregarded, with the original estimate's statistical properties implied for the translated result. That is, if the original effect estimate was statistically significant, there is an unverified assumption that the translated effect is too.
4. *Results are bounded within a plausible range of values.* Some translations are ratios, with a benchmark (e.g., typical growth in 1 year) in the denominator. As the denominator approaches zero the translated metric increases without bound, producing either positive or negative results that are not credible.
5. *Results are substantively consistent across calculation options.* Some translations can be performed in a variety of ways depending on data availability or choices made by the analyst. If these variants do not produce consistent results, substantive conclusions may depend on which one is used. This creates a risk that, no matter how rigorous the original study, interpretations could be manipulated by another party selecting the method most supportive of a particular agenda.
6. *Does not discard useful information.* If the treatment effect was calculated for a particular sample, but the translation only uses a small subsample, the translation may not generalize to the whole sample. This fact should be communicated along with the translated result.

Analytic Background and Translation Options

Consider an educational intervention that has potential effects on student learning. This learning is assumed to be captured by differences in achievement levels between students who received the intervention and a comparable group of students that did not. The analyst standardizes the difference by dividing by the standard deviation of the outcome (Hedges, 1981). Our presentation adopts the convention of standardizing the outcome prior to calculating the difference; the analysis and conclusions of this article are independent of that detail. Regardless of when the standardization is carried out, where possible it should be performed against a broad reference population, such as national or state norms or a well-defined subgroup of such populations, in order to enhance comparability across studies with different samples.

As such, the standardized posttest score z_i for student i can be modeled as a function of treatment status T_i, standardized pretest score w_i, a vector of observed baseline factors $\boldsymbol{X}_i$, (mean centered to simplify ensuing discussion) and unobserved factors ε_i:

$$z_i = \alpha + \beta T_i + \lambda w_i + \boldsymbol{X}_i \gamma + \varepsilon_i \qquad (1)$$

Because z_i is already standardized, β is read directly as the standardized treatment effect.

For application of the methods in this article, we use data from an evaluation of personalized learning (Pane, Steiner, Baird, Hamilton, & Pane, 2017). That report provides details about the assessment, data structure, and analytic methods, most of which are omitted here for brevity. However, the following details are relevant.

- The sample includes 100 schools that adopted personalized learning, predominantly located in low-income urban areas of the United States. There are approximately 22,000

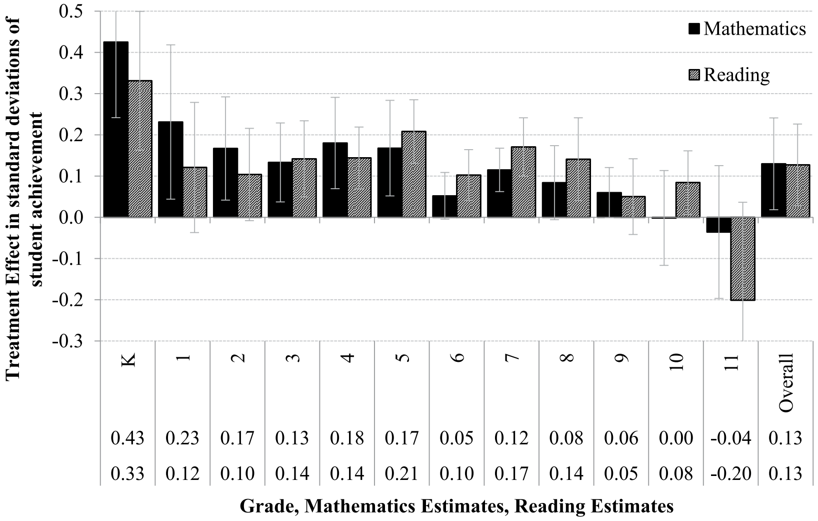


FIGURE 1. *Personalized learning treatment effects, by grade and subject.*

Note. Error bars represent 95% confidence interval. The confidence intervals in some cases extend beyond the viewable frame. Overall represents the sample-weighted average for grades 1–10, the grades for which all of the explored years-of-learning translation options are available. Differences in the sample and methods used in this article cause these results to differ from those reported in Pane et al. (2015, 2017).

treatment group students, and 370,000 matched comparison group students, all of whom were tested in the fall (pretest) and spring (posttest) of the 2014–2015 academic year. As such, this dataset has a relatively large sample and can estimate treatment effects with greater precision than many other studies.

- The pretest and posttest data come from Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP), a computer-adaptive test designed to efficiently determine accurate scores across a wide range of abilities from approximately kindergarten through 11th grade. The scores are reported on a continuous development scale.
- Although Pane et al. (2017) modeled pretest-posttest growth, for the purposes of this article we use the model in equation 1 to estimate the treatment effect, and a sandwich estimator to calculate cluster-adjusted standard errors.

Other studies may employ different research designs, but the resulting interpretation is the same: on average, what difference in student achievement was induced by receiving treatment, in standard deviation units.

Figure 1 presents treatment effect estimates from our data, by grade and subject. These standardized effects may be difficult to interpret, especially for nontechnical audiences. For example, in both math and reading, the overall treatment effects for the full sample are 0.13 standard deviations. Both are statistically significant; are they practically significant?

We next describe the four translation options.

Years of Learning

A popular choice for translating effect sizes is to convert to years (or months, weeks, days) of learning. This is estimated using the ratio of the treatment effect to typical growth on the same scale, i.e., years of learning $= \dfrac{\beta}{\alpha}$, where β is the standardized effect size estimated by the evaluation and α is a measure of typical annual growth in achievement. See the Online Appendix available on the journal website for a derivation of this ratio from the underlying analytic model, and elucidation of the underlying assumptions that both typical achievement and any incremental achievement due to treatment accumulate linearly over time.

There are several potential options for estimating α. One common approach is to use data from Bloom, Hill, Black, and Lipsey (2008), reproduced here as Table 1. They calculate spring-to-spring growth by grade and subject, gathered from a set of standardized tests that enable growth calculations because they are scored on continuous developmental scales (i.e., vertically equated)—a feature not typically present for standardized assessments administered as part of state testing programs (Briggs, 2013).

Table 1 reveals that typical growth varies considerably across grade levels across the tests investigated, immediately calling into question the assumption that achievement grows linearly with time. In particular, given such variation *between* grades, it is highly unlikely that growth is constant *within* grades where seasonality within the academic year may also affect learning rates. The spring-to-spring benchmarks in Table 1 may not be accurate estimates of α for other timespans such as fall-to-spring. Lee, Finn, and Liu (2018) calculate a similar table for fall-to-spring growth. Table 1 also shows standardized growth has a declining trend with increasing age, as has been widely observed (e.g., Dadey & Briggs, 2012). Luyten, Merrell, and Tymms (2017) make the further observation that not all of the measured growth can be attributed to schooling. That is, some of the growth shown in Table 1 is attributable to maturation or other out-of-school factors, further confounding attempts to translate incremental growth into incremental schooling time.

Table 1
Typical Standardized Spring-to-Spring Growth

	Reading Tests		Mathematics Tests	
Grade Transition	Mean	Margin of Error	Mean	Margin of Error
Grade K–1	1.52	±0.21	1.14	±0.49
Grade 1–2	0.97	±0.10	1.03	±0.14
Grade 2–3	0.60	±0.10	0.89	±0.16
Grade 3–4	0.36	±0.12	0.52	±0.14
Grade 4–5	0.40	±0.06	0.56	±0.11
Grade 5–6	0.32	±0.11	0.41	±0.08
Grade 6–7	0.23	±0.11	0.30	±0.06
Grade 7–8	0.26	±0.03	0.32	±0.05
Grade 8–9	0.24	±0.10	0.22	±0.10
Grade 9–10	0.19	±0.08	0.25	±0.07
Grade 10–11	0.19	±0.17	0.14	±0.16
Grade 11–12	0.06	±0.11	0.01	±0.14

Note. Reproduced from Bloom et al., 2008, with permission.

As an example of how this translation works, consider our fall-to-spring estimate of the effect of personalized learning for fourth grade reading of 0.14. Table 1 shows typical growth to be 0.36 for spring of third grade to spring of fourth grade. Setting aside differences in timespan (fall-to-spring vs. spring-to-spring) the translation is $0.14 \div 0.36 = 0.39$ additional years of learning, or 1.39 years of learning for a treated student in one year's time. Assuming 9 months of instruction in a year, this translates to 3.5 additional months of learning. This could further be translated into weeks or days of learning.

Although it is common to use Table 1 for years-of-learning calculations, it was not initially presented for that purpose. Rather, it was one of several types of benchmarks proposed for gauging the practical magnitude of standardized effects (Bloom et al., 2008; Hill, Bloom, Black, & Lipsey, 2008). Those articles concluded that it is useful to apply multiple benchmarks appropriate for the intervention, population, and outcomes under study—each providing a lens for interpreting the effect. This perspective holds the interpretation of standardized effects as a multifaceted and nuanced exercise of informed judgement, a point at risk of being lost in the quest for simple numeric translations of standardized effects.

Notwithstanding the original intent for Table 1, we label it with the lead author of Bloom et al. (2008) to distinguish it from five other estimators of typical annual achievement growth α, presented here in order of increasing specificity to the sample:

- **Hanushek**: The simplest translation, based on Hanushek, Woessmann, and Peterson (2012), estimates a scaling factor of 0.25 standard deviations per year for all grades and subjects.
- **Bloom**: This estimator allows α to vary by grade and subject, however it assumes α is the same for all students and all tests within grade and subject, and that spring-to-spring growth is applicable even for studies covering other timespans.
- **MAP norms**: As previously mentioned, the data in our study come from the NWEA MAP assessments. NWEA estimates national norms of fall-to-spring growth by grade and subject.
- **MAP conditional growth norms (CGN)**: NWEA also uses a flexible model to estimate growth conditional on grade, subject, and starting test score. This allows for students with different achievement levels to have a different expected growth, which we average across students to obtain an estimate of α for our sample.
- **Average control group growth:** This method calculates α as the average standardized score growth of comparison group students within the study, by grade and subject.
- **Regression adjusted:** We may also recover α from the empirical regression. We do so by standardizing the pretest and posttest to the pretest mean and standard deviation, and mean centering all other covariates, so that α represents the change for the average untreated student, controlling for observables.

The last four of these estimators rely on the pretest and posttest coming from the same assessment with scores on a continuous growth scale. Table 2 summarizes assumptions made by each of these variants. Generally, a choice that has fewer assumptions is preferable. Thus, the Bloom translation improves on Hanushek by allowing α to differ by grade and subject, rather than assuming growth is constant across grades. However, both Hanushek and Bloom rely on a set of tests likely to be different than those used in the study at hand and measure typical growth from spring of one grade to spring of the next. By including the time off during summer, a highly studied period when learning rates are low or even negative (e.g., Quinn & Polikoff, 2017), these methods are more likely to violate the assumption of linear growth over time. Lee et al. (2018) estimate fall-to-spring growth but omit information needed to calculate standard errors.

MAP norms improve on these by using the same test and timespan (fall-to-spring, in our application) for α and the estimated treatment effect. MAP CGN additionally controls for variation in starting achievement, which is important if the

Table 2
Assumptions of Growth Scaling Options

Assumes…	Hanushek	Bloom	MAP Norms	MAP CGN	Average Control Group Growth	Regression Adjusted
… growth is constant across grades	✓					
… growth is insensitive to the differences in timespan	✓	✓				
… growth is constant across assessments	✓	✓				
… growth is independent of student's starting level of achievement	✓	✓	✓			
… growth is independent of other observable student characteristics	✓	✓	✓	✓		
… it is unnecessary to further control for student observables	✓	✓	✓	✓	✓	

Note. MAP = Measures of Academic Progress; CGN = conditional growth norms.

sample differs from the national average on baseline achievement because growth may be correlated with baseline achievement. Using average control group growth goes a step further, relaxing the assumption that growth is independent of other observable factors, by using the same comparison sample to estimate both α and β. Finally, the regression adjusted method controls for student characteristics that are associated with typical growth rates.

Matters of inference are typically ignored when presenting years-of-learning translations, even at times incorrectly presuming that statistical significance of the treatment estimator implies statistical significance of the years-of-learning translation. In the Online Appendix, we derive the standard error for the years-of learning estimator, which we employ throughout this article.

Figures 2 and 3 present the translations (of the effects shown in Figure 1) for mathematics and reading respectively, using the six variants of years of learning. To preserve readability, the reading chart displays a range of –0.5 to +1.75 years of additional learning in one year, truncating many of the bars for grades 10 and 11 where the calculation produced extreme values as low as –276 to as high as +37 years. In both subjects, the Bloom estimate of α is unavailable for kindergarten, and MAP norms are unavailable for 11th grade.

Three patterns of sensitivity to grade level are evident in these figures. First, in the early grades the Hanushek method produces much larger years-of-growth estimates than the other methods; this pattern dissipates or even reverses in later grades, an outgrowth of the assumption that α is constant across grades. Second, the various methods appear to produce the most consistent years-of-growth estimates in the middle grades. Finally, even when the various methods are most consistent, there is wide variation in the years-of-learning estimates. For example, in reading, grade 6 has the *smallest* range of effects: 0.15 to 0.41 additional years of learning. Even here, the choice of method can have substantially different implications for the success of the program.

Figure 4 displays averages of these estimators across the whole sample for grades 1–10—the grades for which α is available for all calculations. Averages are calculated by weighting each grade-level estimate by the number of treated students per grade.[1]

To summarize Figure 4, in both subjects, the years of growth estimates based on external norms (Hanushek, Bloom, and MAP norms) decrease when the calculation considers grade level and decrease further when α is derived from the same test and timespan as were used in the study. Once the starting achievement level of the study population is incorporated into the calculation, the three remaining estimators (MAP CGN, Average control group growth, and Regression adjusted) produce nearly identical results in mathematics. However, in reading, these methods produce inconsistent results, influenced by extreme results in 10th grade.

Benchmarking

The next option we explore is benchmarking effect sizes by comparing them against other estimated effects. This calculates a similar ratio as years of learning of $\frac{\beta}{\alpha}$, but here α represents the benchmark rather than typical growth in a year. The standard errors of the ratio follow a similar format as for years of learning, as presented in the Online Appendix.

As discussed in Lipsey et al. (2012), the benchmark can be internal or external to the study. For internal benchmarks, we might compare the effect sizes to other characteristics of the data, such as the baseline achievement gaps between Black and White students or urban and nonurban students. Table 3 displays the results.

In our data, the Black-White gap for reading is 0.59 standard deviations. Given the treatment effect is 0.13, the ratio is 0.22, i.e., the treatment effect is about one fifth of the gap. The urban-rural gap in reading is 0.07, meaning the treatment is 1.83 times this gap. However, the standard error on the urban-rural ratio is quite large and the confidence interval covers a wide band around zero.

For external benchmarks, Lipsey et al. (2012) explain how to use gaps in National Assessment of Educational Progress achievement. We can also compare to the published effects of other inventions. Krueger (1999) found that a decrease of class size from 22 to 15 students for grades K–3 increased achievement by 0.22 standard deviations. As show in Table 3, the average treatment effect of 0.25 in math in K–3 is about the same as this class size reduction effect; for reading, the effect of 0.17 is three-quarters as large. However, the confidence intervals cover a substantially larger range of ratios.

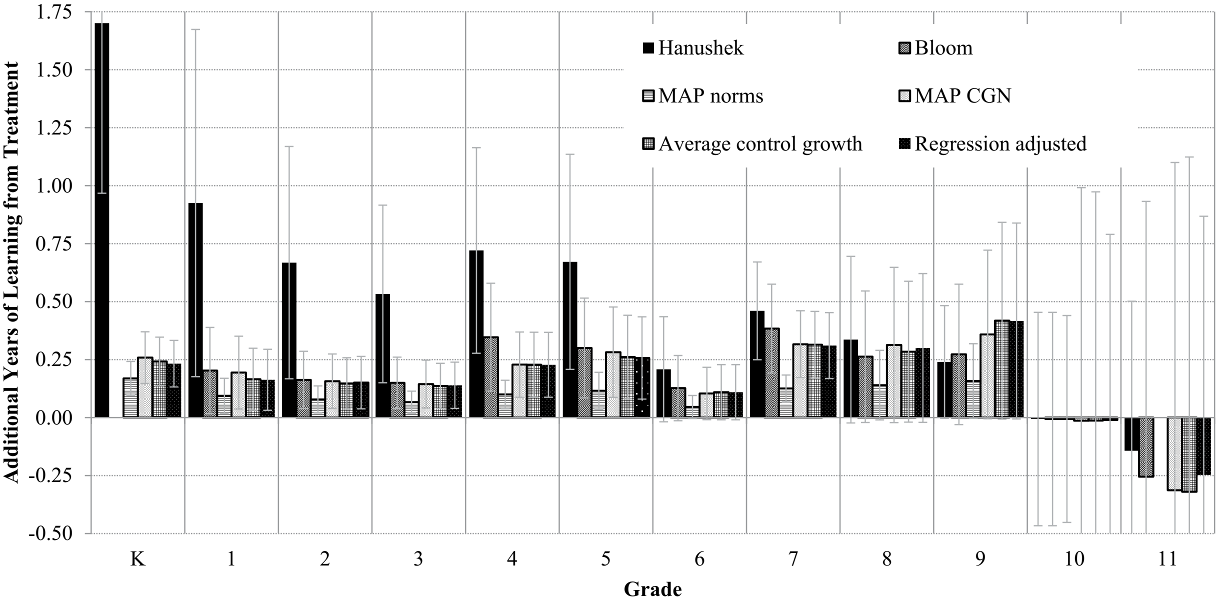


FIGURE 2. *Years-of-learning translations of treatment effects in mathematics.*

Note. Error bars represent 95% confidence interval. Some confidence intervals extend beyond the viewable frame. MAP = Measures of Academic Progress; CGN = conditional growth norms.

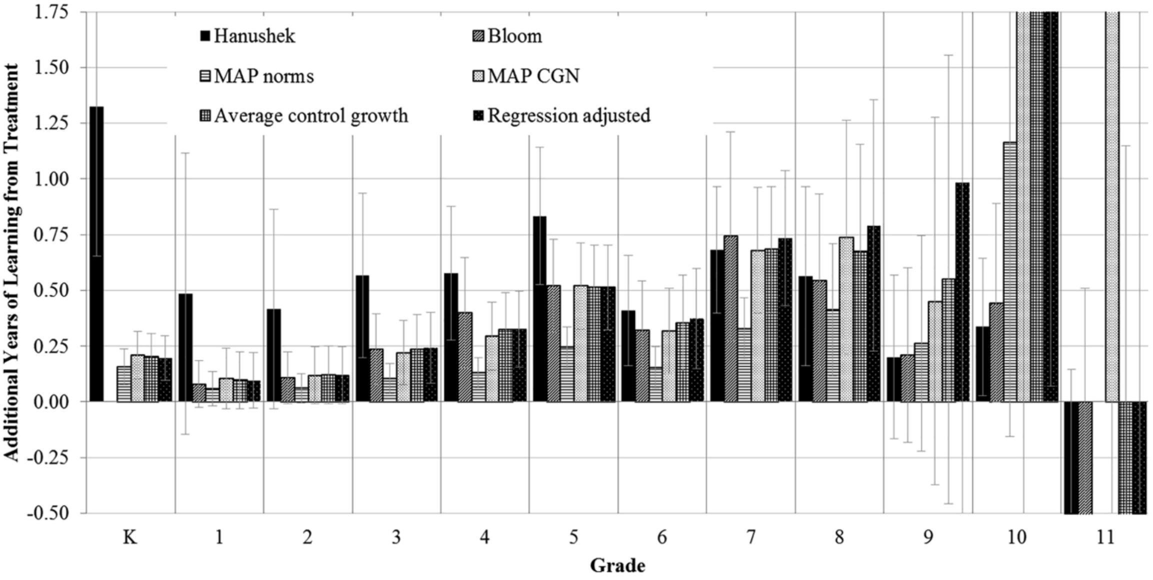


FIGURE 3. *Years-of-learning translations of treatment effects in reading.*

Note. Error bars represent 95% confidence interval. Several bars in grades 10 and 11 as well as several confidence intervals extend beyond the viewable frame. For the bars extending beyond the viewable frame, for grade 10, we find Hanushek = 0.34 ± 0.31, Bloom = 0.44 ± 0.45, MAP National = 1.16 ± 1.32, MAP CGN = 15.89 ± 27.94, average control growth = 5.26 ± 5.19, Regression adjusted = 1.36 ± 2.25. For grade 11, Hanushek = –0.80 ± 0.95, Bloom = –1.06 ± 1.57, MAP CGN = 37.11 ± 103.59, Average control growth = –275.62 ± 5171.28, Regression adjusted = –11.80 ± 47.57. MAP = Measures of Academic Progress; CGN = conditional growth norms.

Percentiles

Another option is to translate to percentile growth. Typically, the translation estimates the change in percentile rank that would have been experienced by the median student in the control group, had they received treatment. Assuming a normal distribution, growth is given by $\Phi(\beta)-0.5$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF). This is the most common implementation of the percentile translation, is among the examples given by Lipsey et al. (2012) and is the "improvement index" used by the What Works Clearinghouse (U.S. Department of Education, 2014). The online appendix generalizes this calculation for any point along the distribution (e.g., enabling calculation of the percentile gain for a first-quartile student), and derives the standard errors.

In this article, we adhere to using the median student as the reference point. We consider two versions of the percentile translation: assuming the standard normal CDF and using

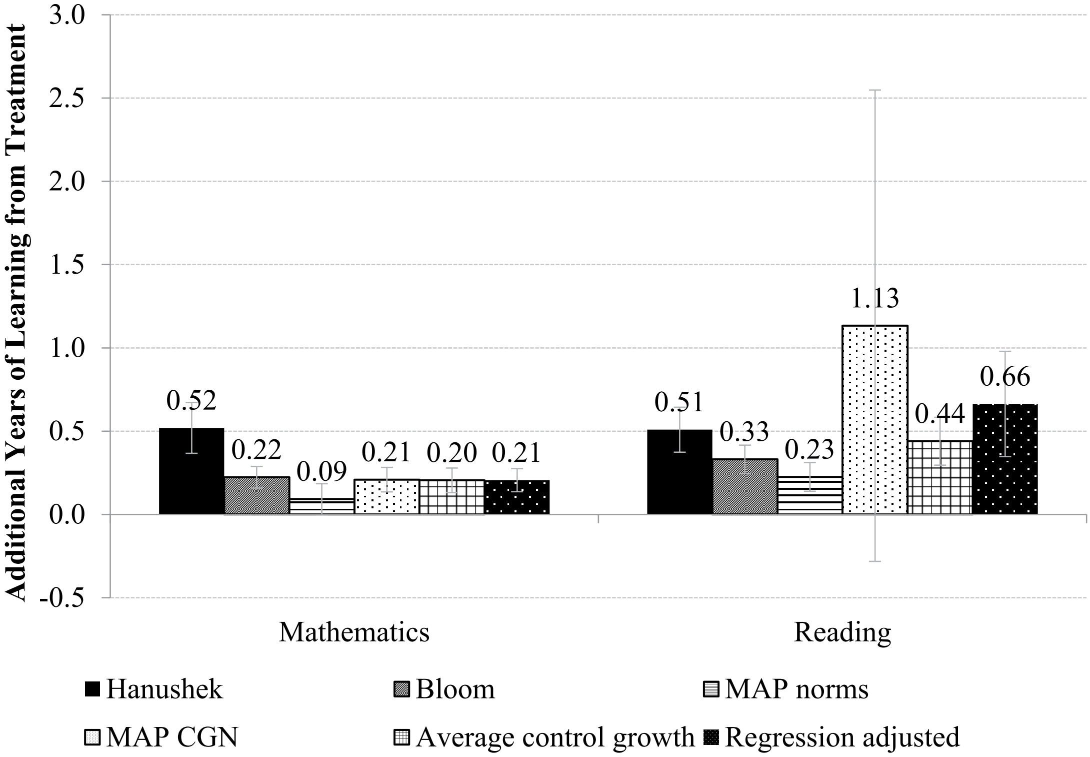


FIGURE 4. *Years-of-learning translations of treatment effects in mathematics and reading, aggregated across grades 1–10.*
Note. Error bars represent 95 percent confidence interval. Only grades 1–10 are included because these are the grades for which all scaling choices are available. MAP = Measures of Academic Progress; CGN = conditional growth norms.

Table 3
Comparisons of Benchmark Translations and Standard Errors

	Black/White Gap		Urban/Rural Gap		Class Size Reduction in Grades K–3	
	Reading	Math	Reading	Math	Reading	Math
Standardized effect estimate	0.13	0.13	0.13	0.13	0.17	0.25
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
Benchmark estimate	0.59	0.66	0.07	0.14	0.22	0.22
	(0.17)	(0.16)	(0.12)	(0.13)	(0.02)	(0.02)
Translated result	0.22	0.20	1.83	0.95	0.79	1.11
	(0.07)	(0.06)	(3.16)	(0.93)	(0.18)	(0.22)

Note. The standardized effect estimates are different for the class size reduction columns because we constrained the sample to grades K–3, the ones evaluated in the benchmark study (Krueger, 1999).

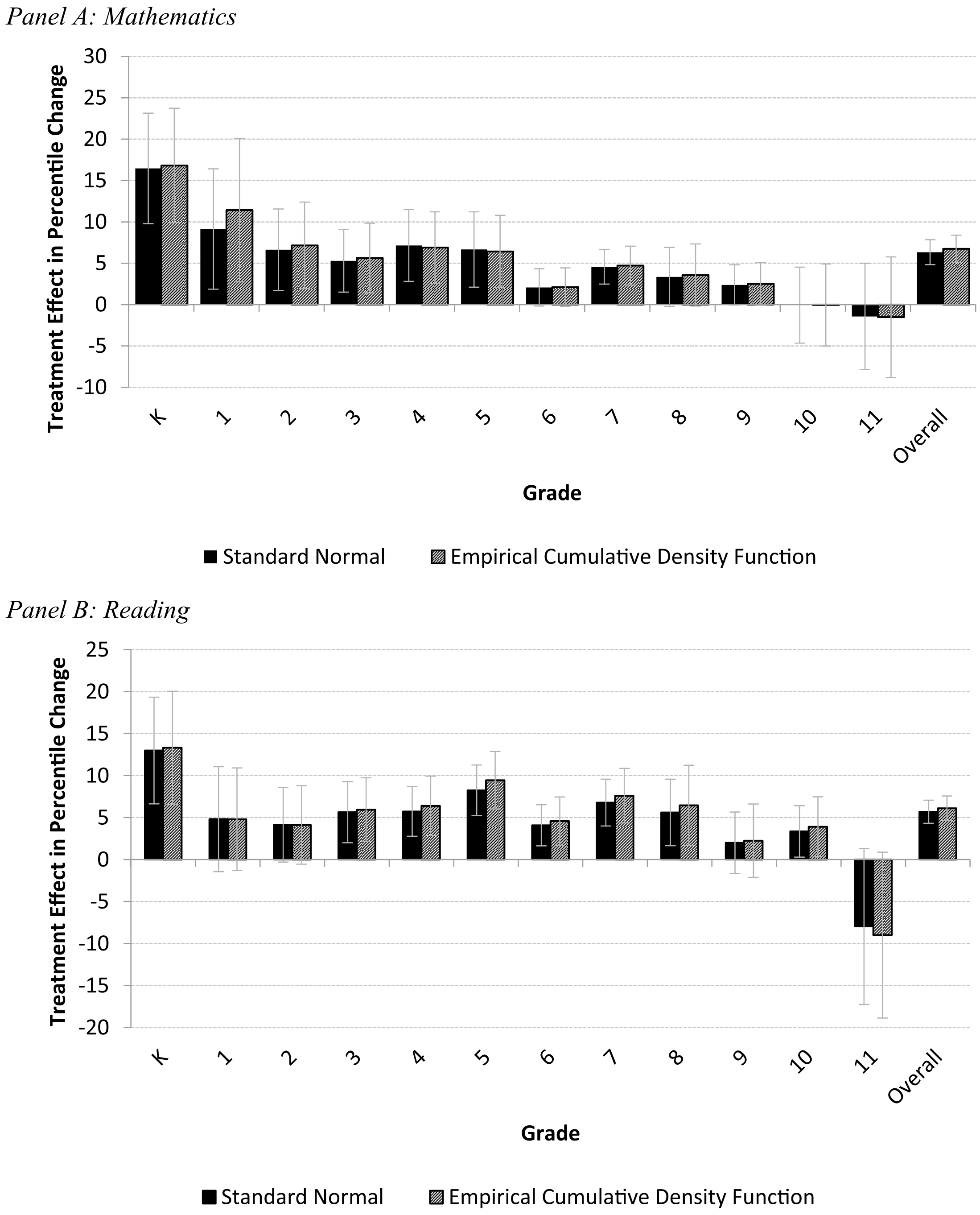
the empirical CDF. The empirical CDF uses a nonparametric estimate of the density traversed in the control group by the increase of score experienced by the treatment group. Intuitively, it just estimates the fraction of control students surpassed in the posttest distribution through the treatment effect. This allows the distribution of scores to take any shape, such as a skewed, bimodal, or highly kurtotic distribution.

Figure 5 presents the percentile results for mathematics and reading. In both subjects, the two methods yield very similar results and have the largest estimates in the earliest grades. In this dataset, the standard normal method generally estimates slightly smaller magnitudes than the empirical CDF method.

Thresholds

The fourth alternative translates to the change in likelihood that a student will attain some level of achievement. Using proficiency as an example, a result of 0.05 indicates treatment induces a five percentage point increase in the likelihood that a student is rated as proficient.

Unlike the previous methods, this requires a reestimation of the treatment effect in a different format. We follow a similar data and regression set-up as for evaluating the treatment effect in standard deviations, except now instead of standardized posttest as the dependent variable we use an indicator variable for the



Panel A: Mathematics

Panel B: Reading

FIGURE 5. *Treatment effects translated to percentile growth.*

Note. Error bars represent 95% confidence interval. Overall mathematics treatment effect, Standard Normal = 6.34 ± 1.51, Empirical CDF = 6.74 ± 1.65. Overall reading treatment effect, Standard Normal = 5.69 ± 1.37, Empirical CDF = 6.11 ± 1.45.

student having a proficient score on the posttest. For our data, we use proficiency as defined by NWEA's alignment with the Smarter Balanced Assessment Consortium standards (NWEA, 2015). In addition to the covariates of the model used above, we include a cubic in baseline scores to account for nonlinearities in those scores' relationship to proficiency. The results are presented in Figure 6.

The same general trends hold, with the largest effects in the elementary school grades. Looking at the weighted averages, we find that the treatment effect for math is estimated as a 3.0 percentage point increase in the likelihood of scoring proficient, from a base of 27%, for about an 11% increase in likelihood. For reading, the magnitude is slightly larger at 3.2 percentage points; working from a base of 34% proficient, this is about a 9% increase in likelihood.

Comparing the Strengths and Weaknesses of Translation Options

We now discuss how the four translations fare on the desirable properties discussed above. Table 4 summarizes the strengths and weaknesses of each translation option across all of the desirable properties.

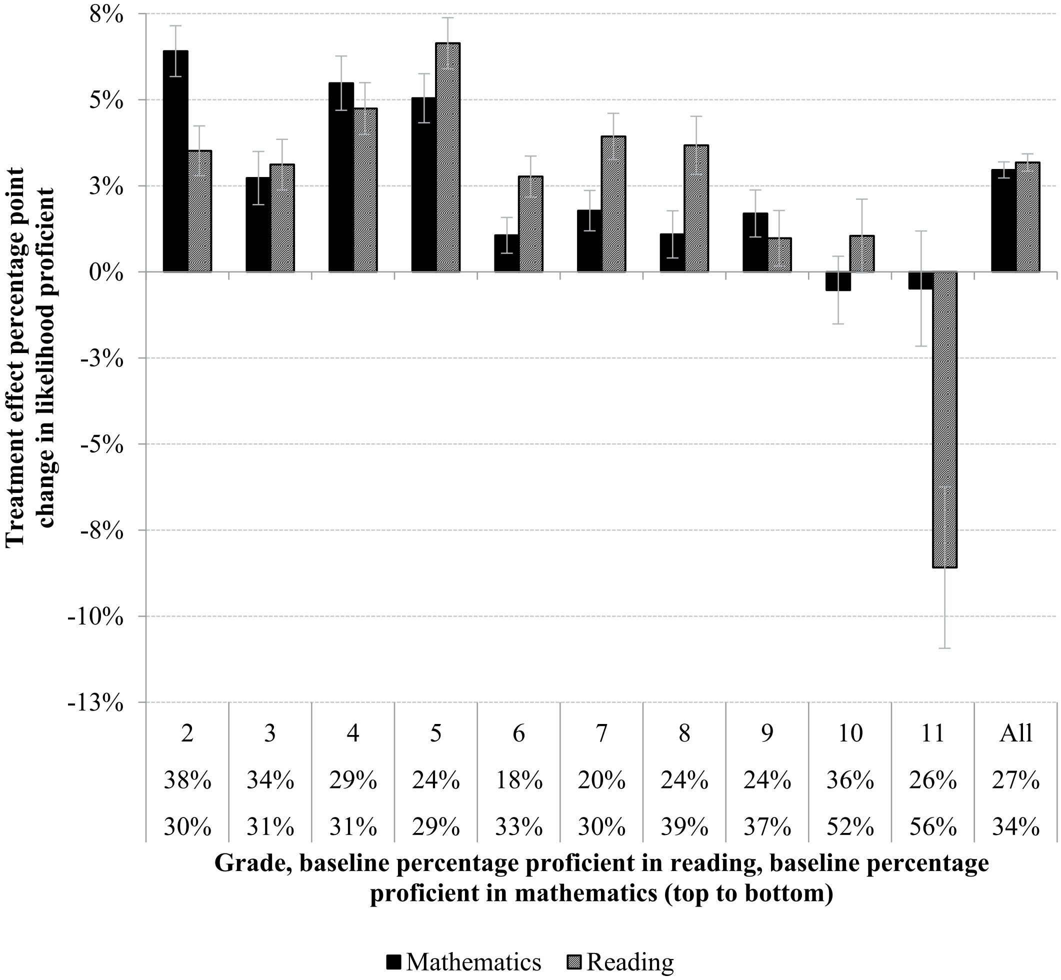


FIGURE 6. *Treatment effect on likelihood of scoring proficient.*

Table 4
Strengths (+) and Weaknesses (–) of Translation Options

	Translation Option			
Desirable Property	**Years of Learning**	**Benchmarking**	**Percentiles**	**Thresholds**
Ease of interpretation	+	–		
Transparent and valid assumptions	–		+	+
Minimizes statistical uncertainty	–		+	+
Bounded to a plausible range	–		+	+
Consistent across calculation options	–	–	+	
Does not discard useful information				–

Strengths and Weaknesses of the Years-of-Learning Translation

The one potential strength of the years-of-learning metric is its ease of interpretation. There is an intuitive appeal to a statement such as: treated students accomplished the equivalent of 1.33 years of learning in the space of one year.

On the other hand, the years-of-learning translation option suffers from many weaknesses across the dimensions. First, the translation is based on a strong assumption that learning rates accumulate linearly over time, an assumption disproven by readily available empirical data showing learning rates are highly dependent on student age and whether school is in session. This basic flaw is rarely discussed when study results are reported in terms of years or other units of time. As a result, research consumers may not be aware of the translation's weak empirical support and may make further extrapolations and erroneous interpretations to address questions of great interest, such as: If control students were taught 0.33 more years would they catch up to the treated students? How does 0.33 years translate to actual days of instruction versus elapsed days? How does this translation relate to different schedules (such as year-round schooling versus 9-month calendars)? Does an effect of 0.33 years imply that 33% of the annual school budget could be saved? Because learning time was not manipulated to estimate the treatment effect in years of learning, the translation does not provide a solid basis for inferences about how additional (or fewer) days of instruction would affect scores.

Further, added statistical uncertainty is typically ignored even though the translation often substantially increases uncertainty over what was present for the original treatment effect estimate. Poor performance is evident in our data, even though in most cases we are only able to estimate a lower-bound standard error. We find that the *smallest* confidence bands have a width of about one quarter of a year, and some exceed ±5,000 years. Indeed, years of learning performs the worst by far when using *t* statistics to objectively rank the translation options. The translations are measuring the significance of the same underlying program effect, so greater precision, i.e., larger *t* statistics, are preferred.

The years-of-learning metric also fails at being bounded within a reasonable set of values, because it is a ratio with a denominator that can in many instances be very small. Highly implausible results are possible, such as many multiples of a year or negative values. In our data for 11th grade reading, variants of the translation produced results between +37 to −276 years of learning. For comparison, percentile translation variants ranged from −8 to −9 for this grade and subject. Implausible results are not confined to our study. For example, Woodworth et al. (2015) reported math effects of about −180 days, or no learning over a 180-day school year, and, in certain states, less than −180 days, implying learning loss over a year of schooling.

Finally, as demonstrated above, the various methods of calculating years of learning can yield substantially different results, an undesirable property. Analysts could select a method they prefer, and it may be infeasible for readers to evaluate the alternatives. In our data, translations of the effect for reading, for grades 1–10 combined, ranged from 0.23 to 1.13 years of learning, a difference of almost a factor of five. The ranges are even wider for individual grade levels. Given available data, the variant that makes the fewest assumptions possible should be preferred (see Table 2). Even within a variant (Bloom), results can depend on which estimates of α are applied. Lee et al. (2018) calculated two distinct estimates of growth per grade and subject, which differ from each other and from Bloom et al. (2008) by factors of 2.5 or more in some instances—resulting years-of-learning translations would vary similarly.

Strengths and Weaknesses of the Benchmarking Translation

This method is susceptible to issues related to the choice of which benchmark to use. First, it may not help clarify whether a *current* study's result is meaningful if readers are left with the question whether *another* study's result is meaningful. A second concern is the risk of misinterpretation. If the treatment is half as large as the Black/White achievement gap, readers may be led to assume the gap could be ameliorated by another year of treatment. Among the flaws of that reasoning is an assumption that students in subgroups experience the same gains as the average treated student. The method is also susceptible to cherry picking of benchmarks by analysts who wish to make results look more or less favorable to suit their own agenda, a fact mitigated somewhat by the possibility that readers could reinterpret results against their own preferred benchmarks. Many of these concerns could be avoided by offering translations relative to more than one benchmark and explaining the limitations of each, enabling research consumers to weigh which is most relevant for their needs. Finally, benchmarks near zero can lead to unbounded comparisons and very large standard errors, a concern that can be mitigated by analysts selecting appropriate benchmarks that are not near zero.

Strengths and Weaknesses of the Percentiles Translation

The percentiles translation option has several desirable properties and no strong weaknesses. First, it does not rely on strong assumptions. The only assumption made is that score distributions are normal, an assumption already made for standardized effect sizes. Where data are available, this assumption can be avoided by using the empirical CDF variant. However, a disadvantage of the empirical CDF is that it is ordinal with respect to scores—results are determined by the ordering of scores traversed irrespective of absolute distance. In our data, the two CDF variants produce very similar results, a result we consider likely to hold in sufficiently large data samples. Second, there is little added statistical uncertainty. Resulting confidence intervals are relatively tight, and percentile conversion is the most precise of the four options when ranked by *t* statistic. Third, results are always bounded within a reasonable range by construction. Values range from positive or negative movements within a 100-point scale. For example, movement from the median is bounded between ±50. In practice, the results rarely, if ever, approach these bounds. Finally, the results across the two versions are very similar, avoiding the risk of analysts selecting the one that favors their agenda.

Strengths and Weaknesses of the Thresholds Translation

Thresholds have several desirable properties. They do not rely on overly strong assumptions: translations based on an alignment study (as used in our example) rely on the assumptions and uncertainties of the alignment study; however, if the threshold is defined on the same test as was used in the original study this issue can be avoided. Thresholds also produce reasonably tight confidence intervals, coming in second behind percentiles in our ranking by *t* statistic. Further, while thresholds could produce implausible results if a linear probability model is used (possibly yielding probabilities outside 0–100%), this can be sidestepped with a nonlinear, bounded model such as a logistic regression.

Practitioners are already familiar with the proficiency categories reported for many assessments. It is relatively easy to interpret an expression of how treatment affects proficiency, for example, raising it by 3% from a baseline of 27%. However, there are potential points of confusion. For example, proficiency may be interpreted as being "at grade level," although this is not typically how cut points are determined. In fact, the cut score is typically somewhat arbitrary, and even the most carefully designed approaches to setting cut scores have a fair amount of subjectivity. Finally, the use of thresholds can distort inferences about changes over time, as discussed in National Academy of Sciences, Engineering, and Medicine (2017, Chapter 6).

One problem with thresholds is that information is discarded by taking the continuous variable of the standardized score and converting it into a discrete variable. All students with scores above (or below) the proficiency threshold are equated, including the top scorer and bottom scorer in the range. Additionally, the estimator requires that there are enough students near the threshold that treatment induces discernable movement across it. The result is a local treatment effect near the threshold that may not generalize for the rest of the sample. Finally, threshold results can vary depending on which threshold is chosen, such as *proficient* or *needs improvement*. These may in some cases lead to substantively different results.

Discussion

Table 4 summarizes the strengths and weaknesses of each of the translation options. Percentiles perform consistently well with no major problems and are our recommended choice for a more interpretable translation of standardized effect sizes. Percentiles are commonly used in education, so most research consumers should be familiar with the metric. As previously mentioned, it is also used by the What Works Clearinghouse (U.S. Department of Education, 2014).

Although converting standardized effect sizes in education to years (or months, weeks, or days) of learning has a potential advantage of easy interpretation, it comes with many serious limitations that can lead to unreasonable results, misinterpretation, or even cherry picking from among implementation variants that can produce substantively inconsistent results. We recommend avoiding this translation in all cases, and that consumers of research results look with skepticism toward research results translated into units of time. If an analyst insists on translating to years of learning, we suggest using the calculation, supported by available data, that makes the fewest assumptions (Table 2). When growth measures are available, typical growth should be estimated within-sample, using the average comparison group growth or regression adjusted methods. When that is not possible, using data from Table 1 or similar is always preferable to using a constant that ignores substantial variation in growth across grades and subjects. Standard errors or confidence intervals of the translated effect should always be reported, and readers should be warned that the translation does not support projections of what would happen if schooling time was increased or decreased. Finally, all analysts should avoid using years-of-learning translations when average growth is small (typically in the higher grades) because these situations often lead to unbounded, implausible results.

NOTES

This work was partially supported by the Bill & Melinda Gates Foundation. This article was completed within the scope of the authors' employment with RAND Corporation. We are grateful for feedback from Laura Hamilton, Mark Showalter, Brad Bernatek, John Engberg, Fatih Unlu, and anonymous referees.

[1] With a goal of calculating an average overall effect, we avoid methods of aggregation that would weight individuals unequally or mask the strong variation in α, such as precision weighting or calculating whole-sample averages of β and α before taking their ratio.

ORCID IDS

Matthew D. Baird https://orcid.org/0000-0003-3016-480X
John F. Pane https://orcid.org/0000-0001-5155-2436

REFERENCES

Bloom, H. S., Hill, C. J., Black, A. B., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328.

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

Childress, S., & Amrofell, M. (2017). *Reimaging learning: A big bet on the future of American education*. Accessed September 7, 2017 from http://www.newschools.org/bigbet/.

Dadey, N., & Briggs, D. C. (2012). A meta-analysis of growth trends from vertically scaled assessments. *Practical Assessment, Research & Evaluation*, *17*(14).

Dorn, S. (2015, September 24). "Weeks/days of learning" is well-intended bad interpretative factoid [Blog post]. *Sherman Dorn: Work to understand how schools have been social institutions*. Retrieved from http://shermandorn.com/wordpress/?p=8079.

Hanushek, E. A., Woessmann, L., & Peterson, P. E. (2012). Is the US catching up? *Education Next*, *12*(4).

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

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

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497–532.

Lee, J., Finn, J., & Liu, X. (2018). Time-indexed effect size for educational research and evaluation: Reinterpreting program effects and achievement gaps in K–12 reading and math. *Journal of Experimental Education*. doi:10.1080/00220973.2017.1409183

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms. (NCSER 2013-3000).* Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Luyten, H., Merrell, C., & Tymms, P. (2017). The contribution of schooling to learning gains of pupils in Years 1 to 6. *School Effectiveness and School Improvement*, *28*(3), 374–405.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35–62.

Maul, A., & McClelland, A. (2013). *Review of National Charter School Study 2013*. Retrieved from National Education Policy Center, University of Colorado Boulder: https://nepc.colorado.edu/think-tank/review-credo-2013

National Academies of Sciences, Engineering, and Medicine. (2017). *Evaluation of the achievement levels for mathematics and reading on* *the national assessment of educational progress*. Washington, DC: The National Academies Press. doi:10.17226/23409

Northwest Evaluation Association. (2015). *Smarter balanced preliminary performance levels: Estimated map scores corresponding to the preliminary performance levels of the Smarter Balanced Assessment Consortium*. Portland, OR. Retrieved from https://www.nwea.org/content/uploads/2015/01/SBAC-Preliminary-Cut-Scores-MAY15.pdf.

Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2015). *Continued progress: Promising evidence on personalized learning*. Santa Monica, CA: RAND Corporation.

Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S., & Pane, J. D. (2017). *Informing progress: insights on personalized learning implementation and effects*. Santa Monica, CA: RAND Corporation.

Quinn, D. M., & Polikoff, M. (2017). *Summer learning loss: What is it, and what can we do about it?* Retrieved from Brookings Institution: http://brook.gs/2wWJKIN.

U.S. Department of Education. (2014). What Works Clearinghouse: Procedures and standards handbook (Version 3.0): Institute of Education Sciences.

Woodworth, J., Raymond, M., Chirbas, K., Gonzalez, M., Negassi, Y., Snow, W., & Von Donge, C. (2015). *Online charter school study 2015*. Center for Research on Educational Outcomes. Accessed September 7, 2017 from https://credo.stanford.edu/pdfs/Online%20Charter%20Study%20Final.pdf.

Yen, W. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299–325.

AUTHORS

MATTHEW D. BAIRD, PhD, is an economist at the RAND Corporation, 4570 Fifth Avenue Suite 600, Pittsburgh, PA 15213; *mbaird@rand.org*. His research focuses on policy at the intersection of education and labor markets to improve outcomes for disadvantaged populations.

JOHN F. PANE, PhD, is a senior scientist at the RAND Corporation, 4570 Fifth Avenue Suite 600, Pittsburgh, PA 15213; *jpane@rand.org*. He studies technology innovations in education using rigorous research methods.

Manuscript received March 8, 2018
Revisions received November 30, 2018;
February 25, 2019
Accepted March 14, 2019