Translating Standardized Effects of Education Programs into More Interpretable Metrics

Matthew D. Baird and John F. Pane

RAND Education

WR-1226-BMGF
March 2018
Funded in part by the Bill and Melinda Gates Foundation

RAND working papers are intended to share researchers’ latest findings and to solicit informal peer review. They have been approved for circulation by RAND Education but have not been formally edited or peer reviewed. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND’s publications do not necessarily reflect the opinions of its research clients and sponsors. RAND® is a registered trademark.
Abstract

Evaluators of education initiatives commonly report results as standardized effect sizes. While this eases comparability across studies, educators and policymakers struggle to interpret the practical importance of results on this scale. This stimulates demand for translating standardized effects to more readily interpretable metrics, such as the number of years of learning necessary to induce the effect. However, the years of learning translation has serious drawbacks and may not be the best choice. We examine four options for translating effect sizes: converting to years of learning, benchmarking against other effect sizes, converting to percentile growth, and estimating the probability of scoring above a proficiency threshold. We describe how each of these perform on a set of desirable properties and use a small survey of researchers to rank them. Years of learning ranks lowest while percentile growth ranks highest and is our recommended translation option.

Keywords: education program effects, years of learning, translation of research results, standardized effect sizes
1. Introduction

Education program evaluations often estimate how a program affects student learning, with changes in test scores used as a proxy for learning. Setting aside statistical complexities, the essence of the analysis is to compare posttest scores of students receiving the program to comparison students serving as the counterfactual. To factor out variations in the posttest scale, researchers typically convert the difference to a standardized effect size by dividing by the standard deviation of the scores. The effects of programs are thus measured in standard deviations of student achievement, allowing for comparison of effects across different programs measured by different tests.

One goal of research into educational effectiveness is to disseminate findings widely and influence policy. This elevates the importance of making results accessible to practitioners and other important decision makers; however, these audiences find standardized effect sizes difficult to interpret. For example, it may not be evident whether an effect of 0.13 standard deviations is meaningful.

Consequently, there have been efforts to translate findings into more readily interpretable metrics. 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.

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. However, others have gone ahead and translated our results (e.g., Childress and
Our resistance is partly because the translation can produce implausible results. 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.

In this paper, we evaluate four options for translation: (1) translation to units of time, such as years or days of learning (Years of Learning); (2) benchmarking results against gaps between demographic groups or effect sizes measured in other studies (Benchmarking); (3) percentile change (Percentiles); and (4) calculating the likelihood of scoring above a reference value, such as scoring proficient (Thresholds).

We discuss a set of desirable properties for translations, and how each of the four translations perform, as rated by a small group of education researchers. Based on those ratings, we recommend translations to be preferred or avoided.

One of our primary conclusions is that years of learning has undesirable properties and should be avoided. We are not the first to caution against using this translation (Dadey & Briggs, 2012; Dorn, 2015). 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.

There are additional issues at play, not fully addressed in this paper. 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 (Yen, 1986; Martineau, 2006; Dadey & Briggs, 2012; Briggs, 2013). There are also concerns about the ability of tests to measure the learning in higher grade levels; when students diversify
coursetaking 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.

2. 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 translator has other issues, apparent ease of interpretation may actually lead to misinterpretation, with consequent faulty decisionmaking. Thus, we enumerate several desirable properties of effect size translations.

1. *Ease of interpretation.* A translation option is useful only if it is easier for practitioners and policy makers 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. It is implied, possibly incorrectly, that if the original effect estimate was statistically significant 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 one 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 a third-party analyst selecting the method most supportive of their own 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.

3. **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. Although there are other options, the analysis and conclusions of this paper are independent of that detail.
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 $X_i$, (mean centered to simplify ensuing discussion) and unobserved factors $\varepsilon_i$:

$$z_i = \alpha + \beta T_i + \lambda w_i + X_i \gamma + \varepsilon_i$$  \hspace{1cm} (1)

In this model, $\beta$ is read directly as the standardized treatment effect.

For application of the methods in this paper, we use data from an evaluation of personalized learning (Pane et al., 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 treatment group students, and 370,000 matched comparison group students, all of whom were tested in the fall (pretest) and spring (posttest) of the 2014-15 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 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 paper 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 non-technical 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 turn to four translation options for making findings more easily interpretable.

### 3.1 Years of Learning

A popular choice for translating effect sizes is to convert to years (or months, weeks, or days) of learning. In this conception, standardized achievement is a function of elapsed time, extending equation 1 with a parameter $D_i$, the fraction of the school year that has elapsed:

$$ z_i = (\alpha + \beta T_i)D_i + \lambda w_i + \gamma X_i + \epsilon_i $$

Setting $D_i$ to one (a full year) recovers equation 1. Although it is highly unlikely that learning rates are constant across the year, this model makes important simplifying assumptions that achievement accumulates at a linear rate after controlling for baseline
characteristics, and that any incremental growth due to a treatment effect also accumulates linearly with time.

Figure 1: Personalized learning treatment effects, by grade and subject

Note: We take the weighted average of treatment effects from grades 1-10 because for the later scaling to years of learning, this is the grade span for which all options are available. Differences in the sample and methods used in this paper cause these results to differ from those reported in Pane et al. (2015, 2017).

The years of learning translation seeks to estimate the additional fraction of a year $\phi$ that must be added to the schooling time of an untreated student to make their achievement equal to a treated student who received one year of schooling. Mathematically, it seeks to equate:
\[ E[z_i|T_i = 0, D_i = 1 + \phi] = E[z_i|T_i = 1, D_i = 1] \]
\[ \Rightarrow \alpha(1 + \phi) = (\alpha + \beta) \]

Solving for \( \phi \),
\[ \phi = \frac{\beta}{\alpha} \]

\( \beta \) is the standardized effect size estimated by the evaluation; \( \alpha \) is a measure of typical annual growth and can be estimated directly from the regression with proper standardization, by other methods such as using the data directly if the pretest and posttest are on a continuous growth scale, or by using an external estimate of typical growth.

One external approach is to use data from Bloom et al. (2008), reproduced here as Table 1. It shows spring-to-spring growth by grade and subject, gathered from standardized tests that enable growth calculations because they are scored on a continuous developmental scale (i.e. vertically equated) – a feature not typically present for standardized assessments administered as part of state testing programs (Briggs 2013).
The table reveals that typical growth varies considerably across grade levels, immediately calling into question the fundamental assumption underlying equation 2, 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 $\alpha$ for other timespans such as fall-to-spring. Lee et al. (2018) calculate a similar table for fall-to-spring growth. Another pattern apparent in Table 1, that standardized growth has a declining trend with increasing age, has been widely observed (e.g., Dadey & Briggs, 2012). Luyten et al. (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

<table>
<thead>
<tr>
<th>Grade transition</th>
<th>Reading tests</th>
<th>Math tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Margin of error</td>
</tr>
<tr>
<td>Grade K-1</td>
<td>1.52</td>
<td>±0.21</td>
</tr>
<tr>
<td>Grade 1-2</td>
<td>0.97</td>
<td>±0.10</td>
</tr>
<tr>
<td>Grade 2-3</td>
<td>0.60</td>
<td>±0.10</td>
</tr>
<tr>
<td>Grade 3-4</td>
<td>0.36</td>
<td>±0.12</td>
</tr>
<tr>
<td>Grade 4-5</td>
<td>0.40</td>
<td>±0.06</td>
</tr>
<tr>
<td>Grade 5-6</td>
<td>0.32</td>
<td>±0.11</td>
</tr>
<tr>
<td>Grade 6-7</td>
<td>0.23</td>
<td>±0.11</td>
</tr>
<tr>
<td>Grade 7-8</td>
<td>0.26</td>
<td>±0.03</td>
</tr>
<tr>
<td>Grade 8-9</td>
<td>0.24</td>
<td>±0.10</td>
</tr>
<tr>
<td>Grade 9-10</td>
<td>0.19</td>
<td>±0.08</td>
</tr>
<tr>
<td>Grade 10-11</td>
<td>0.19</td>
<td>±0.17</td>
</tr>
<tr>
<td>Grade 11-12</td>
<td>0.06</td>
<td>±0.11</td>
</tr>
</tbody>
</table>

Source: Bloom et al., 2008
Temporarily setting aside these concerns, as an example of how this translation works, consider our fall-to-spring estimate for 4th grade reading of 0.14. Table 1 shows typical growth to be 0.36 for spring of 3rd grade to spring of 4th grade, leading to a translation of $\frac{0.14}{0.36} = 0.39$ additional years of learning, or 1.39 years of learning for a treated student in one year’s time. This could further be translated into months, or weeks, or days of learning. Assuming 9 months of instruction in a year, the treatment effect for 4th grade reading translates to 3.5 additional months of learning.

While the use of the Bloom et al. (2008) table is common, it may not be the optimal estimator for $\alpha$. We consider five others, presented here in order of increasing specificity to the sample. These are:

- **Hanushek**: The simplest translation, based on Hanushek et al. (2012), assumes a scaling factor of 0.25 standard deviations per year for all grades and subjects.
- **Bloom**: As discussed above, this allows for $\alpha$ to vary by grade and subject, however it assumes $\alpha$ is the same for all students and all tests within grade and subject, and that spring-to-spring growth is applicable even for studies that cover 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 $\alpha$ for our sample.
• **Average control group growth:** This method uses data from our study, for each grade and subject calculating $\alpha$ as the average standardized score growth of comparison group students.

• **Regression adjusted:** We may also recover $\alpha$ 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 $\alpha$ represents the change for the average untreated student, controlling for observables.

The last four of these 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 $\alpha$ 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 $\alpha$ and the estimated treatment effect. MAP CGN additionally controls for variation in starting achievement, which is important if the sample differs from the national average on baseline achievement, since 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 $\alpha$ and $\beta$. Finally, the regression adjusted method controls for student characteristics that are associated with typical growth rates.

<table>
<thead>
<tr>
<th>Assumes…</th>
<th>Hanushek</th>
<th>Bloom</th>
<th>MAP norms</th>
<th>MAP CGN</th>
<th>Average control group growth</th>
<th>Regression Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>… growth is constant across grades</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>… growth is insensitive to the differences in timespan</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>… growth is constant across assessments</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>… growth is independent of student's starting level of achievement</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>… growth is independent of other observable student characteristics</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>… it is unnecessary to further control for student observables</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We use the typical Taylor expansion approximation to calculate standard errors on an estimated ratio:

$$se\left( \frac{\hat{\beta}}{\hat{\alpha}} \right) = \sqrt{\frac{se(\hat{\beta})^2}{\hat{\alpha}^2} + \frac{\hat{\beta}^2}{\hat{\alpha}^4}se(\hat{\alpha})^2 - 2\frac{\hat{\beta}}{\hat{\alpha}^3}Cov(\hat{\beta}, \hat{\alpha})}$$  \hspace{1cm} (5)

This calculation requires the estimated covariance of $\hat{\beta}$ and $\hat{\alpha}$. If $\hat{\beta}$ and $\hat{\alpha}$ are the same sign (e.g., positive) and estimated using a bivariate regression of score growth on treatment status and a constant, the covariance between the two is given by $-\frac{\hat{\sigma}_e^2}{N(1-T)}$, where $\hat{\sigma}_e^2$ is the estimated variance of the error term, N is the sample size, and T is the fraction treated. Under these assumptions, $-2\frac{\hat{\beta}}{\hat{\alpha}^3}Cov(\hat{\beta}, \hat{\alpha})$ is positive, so the covariance term always increases the standard error of the ratio. Unfortunately, in most cases we do not
have estimates of this covariance. We only have it for the regression-adjusted option, where we recover it from the variance-covariance matrix of the regression coefficients. For the other methods, we assume this term is zero, which produces a lower bound on the total standard errors and confidence intervals of the translated effects.

Figure 2 presents the translations (of the effects shown in Figure 1) for reading, using the six variants of years-of-learning. To preserve readability, the 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. Figure 3 presents results for mathematics on the same scale, without the need to truncate any extreme values. In both subjects, the Bloom estimate of $\alpha$ is unavailable for kindergarten, and MAP norms are unavailable for 11th grade.
Note: Several bars for Grades 10 and 11 truncated on the chart due to very large values. 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.
Three patterns of sensitivity to grade level are evident in these figures. First, in the early grades the Hanushek produces much larger years-of-growth estimates than the other methods; this pattern dissipates or even reverses in later grades, an outgrowth of Hanushek’s assumption that $\alpha$ 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 $\alpha$ is available for all calculations. Averages are calculated by weighting each grade-level estimate by the number of treated students per grade.\footnote{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 $\alpha$, such as precision weighting or calculating whole-sample averages of $\beta$ and $\alpha$ before taking their ratio.}

Figure 4: Years of learning translations of treatment effects in reading and mathematics, aggregated across grades 1-10

![Years of learning translations of treatment effects in reading and mathematics, aggregated across grades 1-10](image)

Note: Only grades 1-10 are included because these are the grades for which all scaling choices are available.

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 $\alpha$ 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 norms, MAP CGN, and Average control growth) all yield similar estimates. In both subjects, the largest estimate is from the Hanushek norm, followed by Bloom, MAP CGN, and Average control growth in decreasing order. Regression-adjusted estimates are generally lower than the unadjusted estimates, indicating that controlling for other factors may reduce the apparent effect of the intervention.
CGN, Average control 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.

### 3.2 Benchmarking

The next option we explore is benchmarking effect sizes by comparing them against other estimated effects. This calculates a similar ratio of \( \frac{\beta}{\alpha} \), but here \( \alpha \) represents the benchmark rather than typical growth in a year. As with years-of-learning, we use the Taylor expansion in equation 5, omitting the covariance term to approximate a lower bound on standard errors.

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 achievement gaps between Black and White students or urban and non-urban students. Table 3 displays the results.

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

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. This helps to illustrate the risks of using a benchmark near zero.

For external benchmarks, Lipsey et al. (2012) explain how to use gaps in NAEP 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.

3.3 Percentiles

Another option is to translate to percentile growth, the change to the distribution of control students had they experienced treatment:

\[
Percentile \text{ Growth} = F(\beta + F^{-1}(C)) - C
\]

(10)

The estimate is conditional on a point (C) in the distribution, such as the median student. This calculation estimates the fraction of control students whose scores are between the score of the Cth-percentile student and that score plus the treatment effect \( \beta \). \( F \) is the cumulative density function (CDF) of the posttest. In this article, we consider two versions of the percentile translation.

- **Standard normal from the median.** Here we assume the standardized posttest scores are distributed standard normal, and focus on the median student. In that case, growth is given by \( \Phi(\beta) - 0.5 \), where \( \Phi(\cdot) \) is the standard normal 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).

- **Empirical CDF.** Instead of assuming normality, the empirical CDF uses a non-parametric 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 (by grade and subject) through the treatment effect. The percentile growth is given by $ECDF(\beta + ECDF^{-1}(C)) - C$, where ECDF is the empirical CDF and $ECDF^{-1}$ is the inverse empirical CDF. This allows the distribution of scores to take any shape, such as a skewed, bimodal, or highly kurtotic distribution.

Standard errors are obtained by using the delta method:

$$se\left(Perc.\left(\hat{\beta}\right)\right) = se(\hat{\beta}) * F'(\hat{\beta} + F^{-1}(C))$$  \hspace{1cm} (11)

Figures 5 and 6 present the percentile results for mathematics and reading. In both subjects, the two methods are highly correlated and yield the largest estimates in the earliest grades. The standard normal method generally estimates slightly smaller magnitudes than the ECDF method. Figure 7 presents the weighted averages for the two methods.
Figure 5: Treatment Effects Translated to Percentile Growth, Math

![Graph showing treatment effects in percentile growth for Math across different grades. The graph compares treatment effects for Standard Normal and ECDF, highlighting notable changes in percentile growth from Kindergarten to Grade 11.]

Figure 6: Treatment Effects Translated to Percentile Growth, Reading

![Graph showing treatment effects in percentile growth for Reading across different grades. The graph compares treatment effects for Standard Normal and ECDF, highlighting notable changes in percentile growth from Kindergarten to Grade 11.]

Translating Standardized Effects of Education Programs into More Interpretable Metrics
3.4 Thresholds

The fourth alternative translates to the change in likelihood that a student will attain some level of achievement such as proficiency. For example, a treatment effect of 0.05 on this scale would mean that the treatment increases the likelihood that a student is rated as proficient by five percentage points.

Unlike the previous methods, this requires a re-estimation 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 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 (Northwest Evaluation Association,
In addition to the covariates of the model used above, we include a cubic in baseline scores to account for nonlinearities those scores’ relationship to proficiency. The results are presented in Figure 8.

**Figure 8: Treatment Effect on Likelihood of Scoring Proficient**

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 percent, for about an 11 percent increase in likelihood. For reading, the magnitude is slightly larger at 3.2 percentage points; working from a base of 34 percent proficient this is about a 9 percent increase in likelihood.
4. Ranking the Translation Options

We now discuss how the four translations fare on the desirable properties discussed above. To help us rank the translation options, we solicited input from attendees of our presentation at the Society for Research on Educational Effectiveness Spring 2018 conference. After describing the four translation options, we presented each of the desirable properties in turn, along with relevant information from this paper, then paused while the attendees ranked the translation options on that property using a paper survey. Ten attendees completed surveys, and we added our own rankings as an eleventh perspective. Table 4 shows the average across respondents for each translation and property. In the following sections we list each translation option in rank order according to the survey, along with some of our own discussion points.

Table 4: Average Rankings of Translation Options

<table>
<thead>
<tr>
<th>Desirable Property</th>
<th>Years of Learning</th>
<th>Benchmarking</th>
<th>Percentiles</th>
<th>Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of interpretation</td>
<td>1.4</td>
<td>3.3</td>
<td>2.9</td>
<td>2.5</td>
</tr>
<tr>
<td>Transparent and valid assumptions</td>
<td>3.7</td>
<td>2.4</td>
<td>1.6</td>
<td>2.3</td>
</tr>
<tr>
<td>Minimizes statistical uncertainty</td>
<td>3.6</td>
<td>2.7</td>
<td>1.5</td>
<td>2.1</td>
</tr>
<tr>
<td>Bounded to a plausible range</td>
<td>3.9</td>
<td>2.9</td>
<td>1.1</td>
<td>2.1</td>
</tr>
<tr>
<td>Consistent across calculation options</td>
<td>3.8</td>
<td>3.0</td>
<td>1.1</td>
<td>2.1</td>
</tr>
<tr>
<td>Does not discard useful information</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Average rank</td>
<td><strong>2.9</strong></td>
<td><strong>2.5</strong></td>
<td><strong>1.5</strong></td>
<td><strong>2.5</strong></td>
</tr>
</tbody>
</table>

Source: Average responses of 11 attendees of authors’ presentation at Society for Research on Educational Effectiveness Spring 2018 conference.

4.1 Ease of Interpretation

1. Years of learning. There is an intuitive appeal to a statement like: treated students accomplished the equivalent of 1.33 years of learning in the space of one year. We personally ranked it lower because of the high risk of misinterpretation caused by
projecting the measured achievement effect into a parameter (time) that was not measured.

2. Thresholds. 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 seem to refer to 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).

3. Percentiles. Percentile rankings are commonly reported with assessment results, so practitioners are familiar with the scale and may easily interpret, say, a 6 percentile increase in the treatment group, versus a comparison group that is at the 54th percentile nationally. However, a limitation of this translation is that some research consumers confuse percentiles with percentages.

4. Benchmarking. Benchmarking 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 on the pretest, 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 the black treated students experience the same gains as the average treated student.
4.2 Transparent and valid assumptions

1. Percentiles. The common application of percentiles makes a plausible assumption that score distributions are normal; where data are available, this can be avoided by using the empirical CDF variant (which made little difference in our data).

2. Thresholds. 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.

3. Benchmarking. There are no hidden assumptions in benchmarking beyond the assumption that the chosen benchmark is relevant and informative.

4. Years of learning. 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 nine-month calendars)? Does an effect of 0.33 years imply that 33 percent of the annual school budget could be saved? Since 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 learning would affect scores.

4.3 Added statistical uncertainty is minimized and clearly conveyed

1. Percentiles. Confidence intervals are relatively tight.

2. Thresholds. Produced the tightest confidence intervals, as discussed below.

3. Benchmarking. Although both benchmarking and years of learning can suffer from very large standard errors when the denominator is small, this problem can be easily avoided for benchmarking by choosing an appropriate (not near zero) benchmark.

4. Years of learning. This translation may not afford a way to avoid small values of typical growth. 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.

For this property, we can also use the empirical results to objectively rank the translations based on t-statistics, the ratio of the translated effects to the standard errors. The translations are measuring the significance of the same underlying program effect, so all else equal, we prefer greater precision, i.e. larger t-statistics. These results are very similar to the survey rankings shown above, except thresholds rank above benchmarking using this method. Rankings are the same whether we average t-statistics across grade levels or calculate them on grade-weighted whole-sample translations. If we substitute t-statistic rankings in place of the survey results, the overall translation rankings do not change.
4.4 Results are bounded within a plausible range of values

1. Percentiles. Results are bounded to 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 because of the non-linear transformation induced by the CDF.

2. Thresholds. Thresholds could run into problems if a linear probability model is used (possibly yielding probabilities outside 0-100 percent), but this can easily be sidestepped with a non-linear, bounded model such as a logistic regression.

3. Benchmarking. Results can be bounded if the analyst selects an appropriate benchmark not near zero. Otherwise, the translated result is unbounded and can lead to insensible results. This concern is mitigated if consumers are aware of the problem; they can select a more reasonable benchmark and re-translate the effect.

4. Years of learning. This translation intrinsically suffers from the risk of unbounded results. Highly implausible results are possible, such as many multiples of a year. 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).

4.5 Results are substantively consistent across calculation options

1. Percentiles. The results across the two versions we present are very similar, and there are no other plausible approaches.
2. Thresholds. Once a particular proficiency threshold is chosen, there is no other choice to be made. The opportunity for manipulation in this method is only in selecting which threshold to use.

3. Benchmarking. A result can be made to look much better or much worse depending on the choice of study or achievement gap to use as the benchmark. However, readers familiar with the metric and existing research can select a different benchmark and re-translate the effect.

4. Years of learning. The various methods of calculating years of learning can yield substantially different results. 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 $\alpha$ are applied. Lee et al. (2018) calculate 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.

4.6 Does not discard useful information

1. (tie) Years of learning, Benchmarking, Percentiles.

4. Thresholds. Information is discarded by taking the continuous variable of the standardized score and converting it into a discrete variable. All students with scores in a given proficiency band are made equal, including the top scorer and bottom
scorer in the band. Additionally, the estimator requires that there are enough students near the threshold so 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.

4.7 Summary of rankings

Table 4 also provides the overall average ranking across properties for each translation option. On average, percentiles performs best, at 1.5, signifying its dominance across the properties. Benchmarking and thresholds rank about the same at 2.5. Years of learning ranks lowest. Although it ranks first on ease of interpretation, it ranks lowest on all other properties except using all information.

5. Recommendations

Although converting standardized effect sizes in education to years (or months, weeks, or days) of learning has an apparent advantage of easy interpretation, it comes with many serious limitations that can lead to unreasonable results or misinterpretation. We recommend avoiding this translation, and that consumers of research results look with skepticism toward research results translated into units of time.

The percentiles translation appears to be the best choice. Percentiles are commonly used in education so most research consumers should be familiar with the metric. It ranked first across all the desirable properties except ease of interpretation. It is also used by the What Works Clearinghouse (U.S. Department of Education, 2014). However, percentiles are confused with percentages with alarming regularity, and researchers should take care to emphasize the distinction. When data are available we recommend
using the empirical distribution of scores in the control group as a precaution against violation of a normality assumption (in practice, this may not matter).

If an analyst insists on translating to years of learning, we suggest using the most suitable control group supported by available data. When growth measures are available, typical growth should be estimated within-sample, using the average control growth or regression adjusted methods. When that is not possible, Table 1 or similar is always preferable over using a constant that ignores substantial variation in growth across grades and subjects. Standard errors or confidence intervals 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.

Acknowledgements

This work was partially supported by the Bill & Melinda Gates Foundation. We are grateful for feedback from Laura Hamilton, Mark Showalter, Brad Bernatek, John Engberg, Fatih Unlu, and attendees of presentations at the Western Economic Association Annual Meetings, the RAND Center for Causal Inference, and especially the SREE Spring 2018 conference (who also provided input we used to rank the translation options).

References


