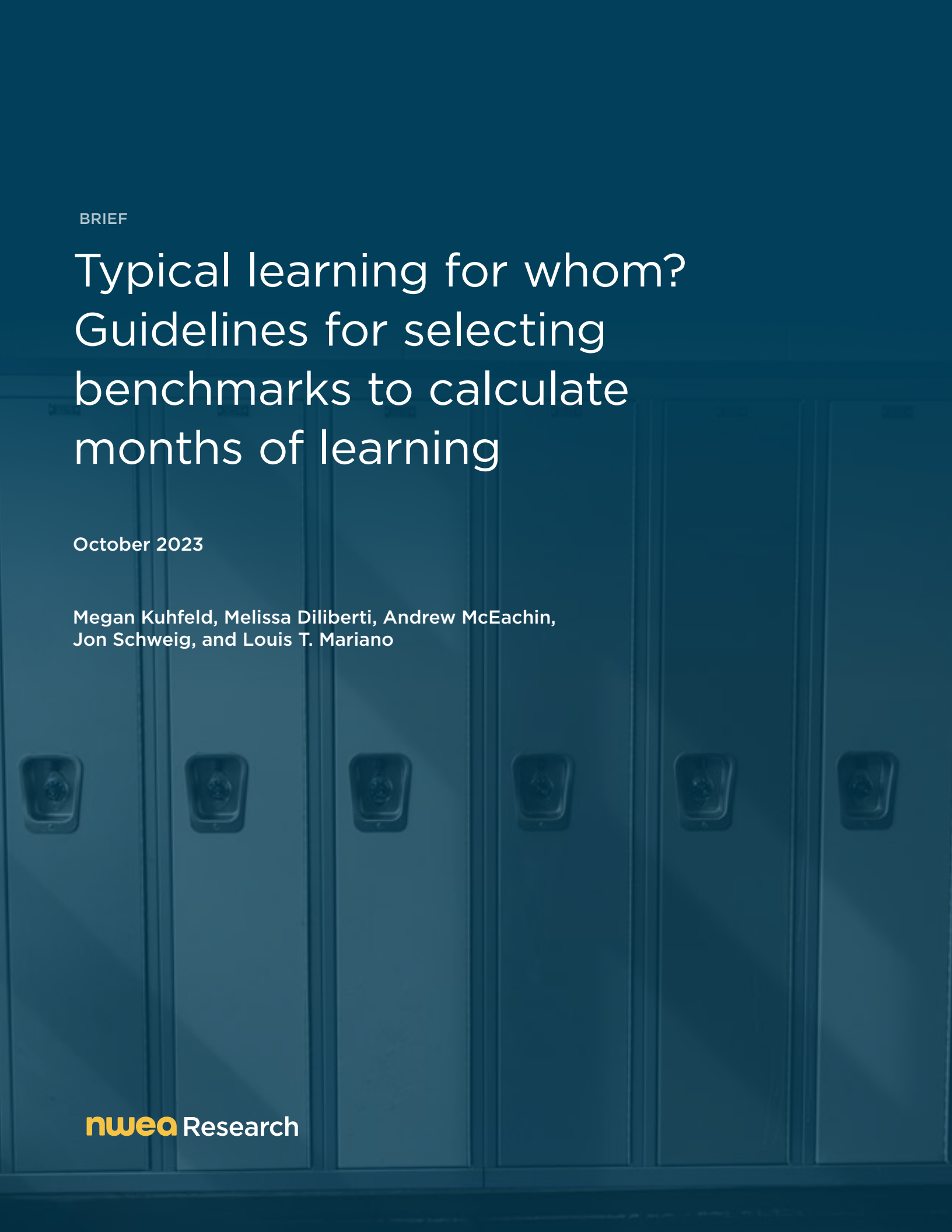


BRIEF

Typical learning for whom? Guidelines for selecting benchmarks to calculate months of learning

October 2023

Megan Kuhfeld, Melissa Diliberti, Andrew McEachin, Jon Schweig, and Louis T. Mariano

nwea Research

Numerous headlines have proclaimed just how far behind students have fallen since the COVID-19 pandemic first shuttered schools in March 2020 (e.g., "Children lost about 35% of a normal school year's worth of learning," "U.S. students ended the pandemic school year 4 to 5 months behind," "Nation's report card: two decades of growth wiped out by two years of pandemic") and how much time is needed for students to return to pre pandemic score levels (e.g., "Full academic rebound likely 5 or more years away"). Many headlines report test score changes in "months (or years) of learning" from recent studies. Increasingly, researchers have translated test-scores differences into months of learning because this metric resonates with audiences—including policymakers and practitioners—who may be less comfortable or less familiar with alternate metrics such as standard deviations or percentile ranks.

Despite the accessibility (and therefore popularity) of the "months of learning" metric, it has some major downsides. For one, the metric can easily be manipulated. Researchers have a great deal of leeway in choosing the typical growth benchmark to use and can select benchmarks that make a relatively small effect size in SD units appear to be more substantial. There are potential technical problems too. For example, when typical learning rates are close to zero—which commonly occurs when advanced middle and high school students take tests that mostly cover K–8 curriculum—it's possible (even probable) to get nonsensical estimates like a personalized learning intervention resulting in 276 years of learning lost. In sum, the validity of the months of learning translation depends on the accuracy of the original effect size *and* the choices made in selecting the typical learning benchmark. We focus on the latter here because while researchers often spend a lot of time describing the methods used to calculate an effect size, the choice of benchmark is often unclear and/or under justified.

Because of these problems, we (and others) have publicly criticized the months of learning metric despite occasionally using it when the perceived benefits (interpretability) outweigh the costs. Nonetheless, there are more and less defensible ways of calculating months of learning.

In this brief, we walk through four important questions that should be considered when selecting a typical learning benchmark with the aim of helping researchers make more informed decisions about how to calculate months of learning:

1. **Which assessment (or set of assessments) should I use to determine "typical" learning rates?**
2. **Is the population of students used to calculate the typical growth estimate comparable to my sample population?**
3. **Do I need a benchmark for a specific grade/subject or am I trying to generalize across multiple grades/subjects?**
4. **Are my months of learning estimates plausible? Would I get a substantially different answer if I made different choices for questions 1 through 3?**

To produce these estimates, researchers take differences in test scores and a benchmark and divide these differences by some rate of "typical" learning. For example, consider the article reporting that students lost learning equivalent to 35% of a school year due to the COVID-19 pandemic. To arrive at this estimate, the authors first compared students' pre-COVID and COVID test scores and found that students' test scores in the COVID era were 0.14 standard deviations (SD) lower than the pre pandemic era. Researchers then assumed that "students generally improve their performance by around 0.40 standard deviations per school year"—an assumption they obtained from a 2008 study documenting typical learning rates by grade/subject. To translate this into months of learning, researchers divided these two numbers (0.14/0.40) to arrive at 35 percent of the school year.

$$\textit{months of learning} = \frac{\text{effect size (COVID- preCOVID)}}{\text{typical learning benchmark}}$$

While we discuss these questions in the context of understanding important dynamics of the impact of COVID-19 on student learning, the underlying principles discussed apply more broadly to all education research using assessments to benchmark months of learning.

1. Which assessment (or set of assessments) should I use to determine "typical" learning rates?

What are the choices?

Researchers have used two main approaches to contextualize COVID-19 learning losses as months of learning: (a) use pre-COVID data from the same assessment to get a benchmark for typical growth or (b) use a standardized benchmark estimate from prior research based on a different set of assessments. As shown in Table 1, there are tradeoffs with each choice. An advantage of choice (a) is the typical learning benchmark will be on the same scale as the outcome. But because the typical learning benchmark relies on a consistent scale and sample from before and after the pandemic, this information may not be available for many assessments (e.g., many states have changed their testing programs and/or state standards in the last ten years). Choice (b) may be more generalizable because it pools together typical learning growth rates from multiple assessments, but it may not be as well matched to the design of the assessment in question.

Table 1. Comparison of benchmark choices to obtain a pre-COVID benchmark

	Choice A: Use pre-COVID data from the same assessment to calculate benchmark	Choice B: Use standardized benchmark estimate from prior research
Pros	• Benchmark is on the same scale as the outcome • Benchmark can be estimated with a customized sample	• Benchmarks are publicly available and citable • Consistent benchmarks can be used across studies
Cons	• Requires a consistent assessment and sample of students prior to and post-COVID • Requires multiple timepoints/grade levels (e.g., a test like NAEP that is not administered in consecutive grades should not be used to calculate "typical learning" rates) • Sensitive to sample/year inclusion rules in estimating the benchmark so may not be consistent across different analyses • Requires either access to raw test-score data or normative documentation, which may not be publicly available	• Widely used empirical benchmarks are based on assessments that are not widely administered now • Typical growth on one assessment may not generalize to the assessment administered to your sample

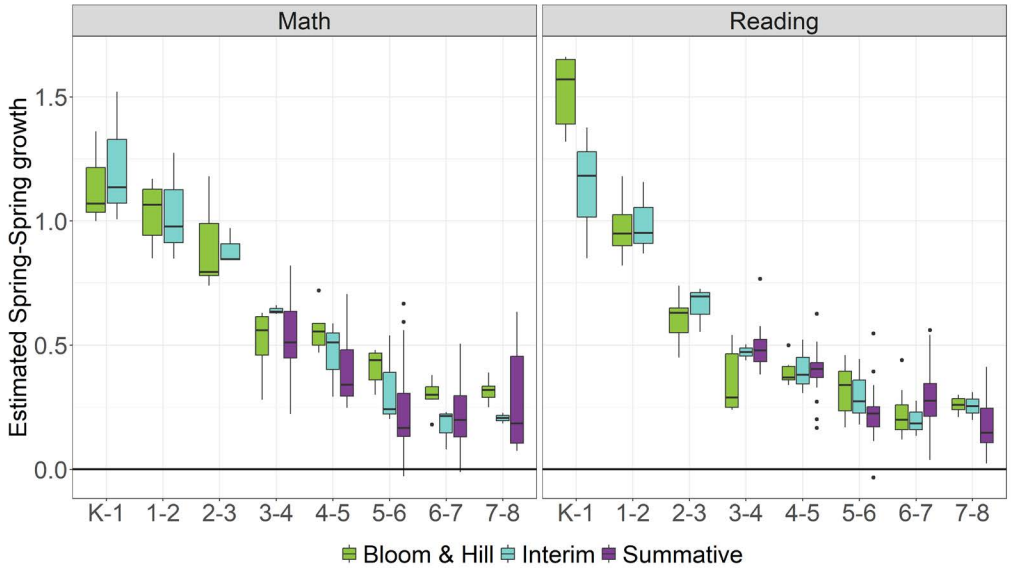
When and why does the choice matter?

The degree to which the choice of (a) or (b) matters depends on how consistent growth benchmarks are across assessments. While it has been widely demonstrated that year-to-year gains in achievement vary by subject and grade level, less is known about how much year-to-year gains vary across different types of assessments. Figure 1 illustrates a set of empirical benchmarks for spring-to-spring standardized gains calculated based on three sources: (1) the seven tests used in the empirical benchmarks discussed by Howard Bloom and Carolyn Hill (which come from norming samples collected in the 1990s to early 2000s); (2) the norms reports from three widely used computer adaptive interim assessments (MAP® Growth™, i-Ready, and Renaissance tests); and (3) the technical manuals of 20 vertically scaled summative assessments collected by Sanford Student.

There are a few important takeaways from Figure 1. First, as previously demonstrated, typical growth varies greatly between subjects and across grade levels. According to the Bloom and Hill empirical benchmarks, the average growth in test scores for a student moving from K to 1st grade is 1.52 SDs in ELA and 1.14 SDs in math, while the average growth between 7th and 8th grade is only 0.26 SD in ELA and 0.32 SDs in math. Second, there is variability in typical growth across three sets of assessment data. The Bloom and Hill benchmarks imply the highest typical growth estimates in most middle school grades, followed by interim assessments and then the summative tests. For example, the average growth in math test scores for a student moving from 5th to 6th grade is 0.41 SDs using the Bloom and Hill benchmarks but only .33 SDs using benchmarks from interim assessments and only .24 SDs using benchmarks from summative assessments. Third, there is heterogeneity in typical growth estimates within a category of assessment, even for a given subject and grade. In particular, the summative assessment results from the middle school grades show extremely large variation in expected growth estimates across states. For example, expected growth estimates for students moving between 6th and 7th grade in both subjects range from essentially zero growth on one summative test to over 0.50 SDs on a different summative test.

In sum, the observed heterogeneity in typical learning growth benchmarks across grades, subjects, and assessments shown in Figure 1 means that months of learning calculations are going to be sensitive to which benchmark researchers choose as their denominator.

Figure 1. Range of standardized spring-to-spring growth estimates by assessment type



Note. Reported estimates are calculated by subtracting a grade-level mean from the prior grade's mean and dividing by the pooled standard deviation across the two grades. Each box shows the variability in estimates within a category (with the horizontal line in the middle representing the median estimate). Achievement gains in 2022–23 lagged prepandemic trends across groups

What should you do?

Given the variability across typical learning rates, we concur with prior researchers who have recommended using typical learning growth benchmarks from the same assessment that is used to calculate effects whenever possible. Before doing so, you should confirm that your assessment and associated data satisfy three conditions: (1) your assessment is vertically scaled to allow for the calculation of growth estimates across grade levels, (2) there a minimum of two pre-COVID test administrations of the assessment, and (3) there have not been any major blueprint or scale changes between the pre-COVID and COVID period that would affect score interpretation. If your assessment does not satisfy these conditions, standardized benchmark estimates from prior research are your best option.

2. Is the population of students used to calculate the typical growth estimate comparable to my sample population?

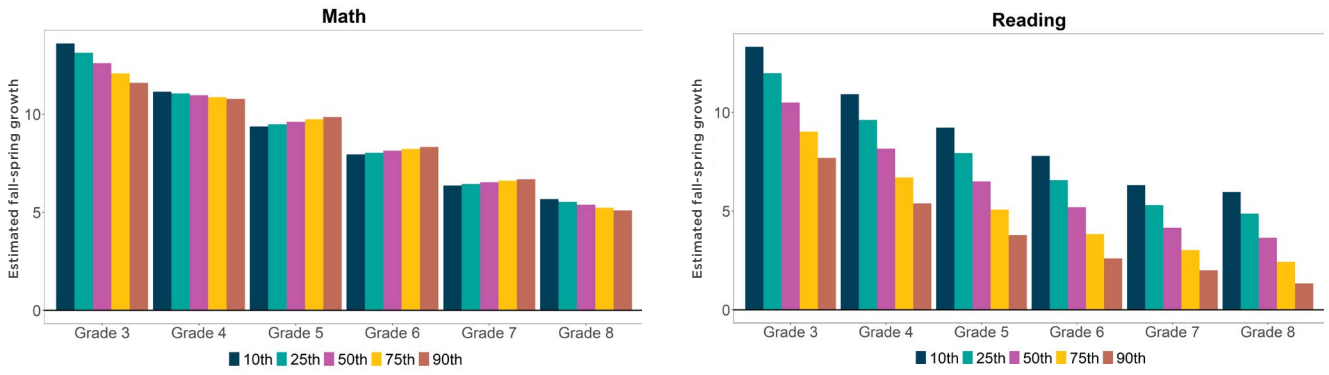
What are the choices?

It is also important to consider if the sample composition of your COVID sample mirrors the sample used to produce the "typical learning" benchmarks. Growth rates can be associated with student and school characteristics in ways that can affect the representativeness of an average benchmark for your particular sample. While most researchers use average growth benchmarks from a national or state sample, there are times when it may be more appropriate to calculate a benchmark from your own sample or a sample that is more similar in composition.

When and why does the choice matter?

While it is not possible to review all the various situations in which a sample may differ from the state/national population, we will focus on potential complications arising from benchmarking growth in districts that differ in their achievement levels. As seen in Figure 2, expected growth on MAP Growth in reading is associated with students' starting achievement, which means that students who are further behind are expected to grow more than students at the average. Suppose we have two school districts that each have a 4-point RIT gap between COVID and pre-COVID test scores, but in one district, the median 3rd grade student is at the 10th percentile and in the other, the median student is at the 90th percentile in reading. As shown in Figure 2, while the average (pre-COVID) 3rd grade student was expected to gain 10.5 RIT points across the school year in reading, students at the 10th and 90th percentiles would be expected to gain 13.3 versus 7.7 RIT points, respectively. If we were to benchmark the gap in both districts based on average growth rate (10.5 RIT points), we'd conclude districts have the same additional learning needs (4 RIT points / 10.5 RIT points, or 38% of a school year). However, this would overstate the need in the low-achieving district (4 RIT points / 13.3 RIT points = 30% of a school year) and understate the need in the high-achieving district (4 RIT points / 7.7 RIT points = 52% of a school year).

Figure 2. Expected growth rates by starting percentile from the 2020 MAP Growth norms



Note. Estimates are estimated within year (fall to spring) gains on MAP Growth in RIT points conditional on starting RIT. Calculations are reported in the NWEA 2020 MAP Growth norms.

What Should You Do?

The most important thing to do is to compare the characteristics of your sample (e.g., student achievement levels, race/ethnicity, gender) to characteristics of the normed sample being used. If the characteristics of your sample align well with those of the national sample, it should be fine (and potentially preferable from a generalizability standpoint) to use benchmarks calculated from the national sample. But if your sample is not well-aligned, we recommend using (a) more localized estimates (ideally from your own sample, if possible) or (b) stratifying the normed sample to obtain estimates that are closer to the sample of students being used in your analysis.

3. Benchmarking an effect size based on a single grade vs. multiple grades/subjects

What are the choices?

In some situations, researchers have separate effect size estimates per subject/grade. In this situation, the appropriate choice is to benchmark your estimates based on typical learning rates from the corresponding subject/grade. For example, if you have an effect size for students in grade 7 in math, it is most appropriate to use a corresponding grade 7 math benchmark.

Sometimes, however, researchers may not have separate effect-sizes by subject and/or grade and may instead only have a single pooled effect size. For example, in meta-analyses, data may have already been pooled across grades (or even across grades and subjects) to a single effect size. In this situation, it is less clear which benchmark estimate to use to translate the effect size, and you must make a choice. One option is to aggregate existing benchmarks across grade levels. For example, pretend you have average spring–spring growth estimates for grades 3–4, grades 4–5, grades 5–6, grades 6–7, and grades 7–8. You might average these five spring–spring growth estimates and use the average as your benchmark. A second option is to pick multiple benchmarks representing the ranges of grades included in the sample. In the second option, you might state something like "a 4-point gap between the COVID and pre-COVID period would represent 40% of a year of learning for 3rd graders and slightly over a year for 8th graders in reading."

When and why does the choice matter?

As we have seen in Figures 1 and 2, there can be considerable variability across grades and subjects in typical learning rates. As a result, the same effect size could imply very different months of learning by grade level. The choice of whether to use a single aggregate benchmark versus multiple benchmarks to demonstrate the range of the effect is going to come down to whether it is useful to highlight the variability across grade levels.

What should you do?

If you are benchmarking an effect size that is pooled across grades/subjects, we have two primary recommendations. First, it is important to be transparent about the grades/subjects used to create the composite benchmark (e.g., say "averaging across grades 3–8, the typical math learning rate is expected to be 0.40 SDs" rather than "the average student is expected to gain 0.40 SDs" with no acknowledgment of *who* is included). Second, it can be helpful to highlight the potential variability across grades in months of learning from the same effect size (e.g., a gap of 0.10 SDs may be more substantial for middle schoolers, given their lower growth rates on average).

4. Confidence check: Examining the sensitivity of your translated estimate to the benchmark choice

Finally, it is worth considering how the choices made in the previous three steps may have impacted your months of learning estimate. First, if there are several plausible choices for the benchmark, it is recommended that you calculate months of learning with each benchmark and report a range of months of learning estimates as a sensitivity check. Second, it is important to do a gut check on the plausibility of the reported months of learning. If the range of estimates across benchmarks is broad (e.g., one month to two years) or if estimates are implausibly large (e.g., students are 25 years behind) it is probably better not to report the months of learning translation.

Final recommendations

If you go down the route of calculating months of learning, we believe the questions outlined in the brief will help researchers think carefully about the benchmark used. In addition to careful consideration of the benchmark options, it is important to be transparent about the choice of benchmark in the reports.

1. **Details about (and justification for) the benchmark selected should be reported transparently in the paper.**

2. **Typical learning estimates for each grade/subject should be available in technical appendices or publicly available norms reports.**

3. **It is worth conducting sensitivity checks and potentially report a range of values. Additionally, language cautioning about the imprecision of this conversion approach is often warranted.**

Additionally, we recommend all consumers of research be more suspicious of these calculations and challenge them instead of just accepting them as is.

ABOUT THE AUTHORS

Dr. Megan Kuhfeld is a research manager for the Collaborative for Student Growth at NWEA. Her research seeks to understand students' trajectories of academic and social-emotional learning (SEL) and the school and neighborhood influences that promote optimal growth. Dr. Kuhfeld completed a doctorate in quantitative methods in education and a master's degree in statistics from the University of California, Los Angeles (UCLA).

Melissa Diliberti is an assistant policy researcher at the RAND Corporation and a doctoral fellow at the Pardee RAND Graduate School. Her research primarily focuses on how state and local policy conditions influence teachers' instructional practices. Diliberti has a master's degree in public policy from George Washington University.

Andrew McEachin is the VP of Research and Policy Partnerships at NWEA. His team's work is devoted to broadening understandings of educational equity and transforming education research and practice through research conducted in partnership with district and school communities. Andrew completed a doctorate in education policy and a master's degree in economics at the University of Southern California.

Dr. Jonathan Schweig is a senior behavioral and social scientist at the RAND Corporation. His research focuses on teaching and the school climate conditions that support the education and wellbeing of young persons. Dr. Schweig completed a doctorate in quantitative methods in education and a master's degree in statistics from the University of California, Los Angeles (UCLA), and a master's degree in curriculum and teacher education from Stanford University.

Dr. Louis T. Mariano is a senior statistician at RAND. His research interests in education have focused on evaluation of the efficacy of education programs, policies, and reforms; experimental and quasi-experimental design methodology; and statistical applications to mental measurement, including student assessment. Dr. Mariano received his PhD in statistics from Carnegie Mellon University.

About NWEA

For more than 40 years, NWEA has been a pioneer in educational research and assessment methodology with a focus on improving learning outcomes for every student. NWEA continues this discovery through dedicated research that explores foundational issues in education, practical challenges in today's schools, and the evolving role of technology in the lives of students. As a mission-based educational research organization, NWEA's research agenda reflects our commitment to attacking big challenges in education and measurement and empowering education stakeholders with actionable insights.

The research reported here was sponsored by the Institute of Education Sciences, U.S. Department of Education, through grant [R305U200006]. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

nwea

NWEA, a division of HMH, supports students and educators worldwide by providing assessment solutions, insightful reports, professional learning offerings, and research services. Visit NWEA.org to find out how NWEA can partner with you to help all kids learn.

NOV23 | WELTSK6813