

Reclaiming “months of learning”: An accessible metric for aggregate educational progress
Andrew Ho, *Harvard Graduate School of Education*

Presented to the Reporting and Dissemination Committee of the National Assessment Governing Board,
November 16, 2023

Background: NAGB and NCES report educational progress using average scale scores and percentages of proficient students.

Example: Average Grade 4 NAEP Reading scores [declined 3 scale score points](#) from 2019 to 2022. The percentage of 4th graders proficient in Reading [declined 2 percentage points](#).

Problem: General audiences cannot understand whether these trends are large or small. Absent an engaging and intuitive reference point, trends in scale scores and proficiency percentages seem, to most audiences, to be less substantial than they are. This raises questions about the usefulness of NAEP.

Recommendation: NAGB and NCES can work together to gather evidence for the NAEP “cross-grade (vertical) scale.” With sufficient validity evidence, NAGB and NCES can decide whether and how to deploy “grade level” and “months of learning” metrics to report educational progress.

For Example: “The Nation’s Report Card uses a provisional metric known as ‘NAEP months of learning’ to describe aggregate trends for low-stakes purposes. In NAEP Reading, fourth graders in 2022 lagged 2 NAEP months of learning behind fourth graders from 2019.”

Caveats: Requires validity evidence for cross-grade comparability from Grade 4 to 8. Some interpretations require linear assumptions of progress. The metric is better suited to indicate progress, not achievement gaps. If adopted, the metric should complement not supplant existing trend metrics.

In social measurement, there are three metrics that the general public tends to accept as intuitive: Dollars, Durations, and Counts. Compared to seemingly arbitrary test scores scales that range, for example, from 200 to 500, Dollars, Durations, and Count metrics raise fewer follow-up questions like, “what does that really mean?” Consider the claim from Doty et al. ([2022](#)), “if allowed to become permanent... recent [NAEP] losses would represent a 1.6% decline in present value of lifetime earnings for the average K-12 student (or \$19,400), totaling \$900 billion for the 48 million students enrolled in public schools...” (p. 2). And consider the aforementioned decline of 2 percentage points in NAEP Reading proficiency. Because there are roughly [3.7 million 4th graders](#), this means that there are 100,000 fewer U.S. public school 4th graders who are proficient in Reading in 2022 than there were in 2019. Although critical readers may ask, “how can we link lifetime earnings to NAEP?” Or “what does proficient mean?” Most find Dollars and Counts metrics themselves intuitive.

I recommend that we revitalize and harness a long-dormant empirical linkage in the NAEP scale that will add the Durations metric to our toolkit of NAEP reporting strategies. I will briefly explain how my colleagues and I use this historical empirical linkage to support “grade level” interpretations for low-

stakes aggregate interpretations. I then describe how NAGB and NCES could conduct a special study that could re-establish the basis for this linkage. Finally, I suggest approaches to using the results of such studies for reporting for special purposes, in particular for aggregate monitoring of progress over time for low-stakes interpretations.

How we (the Educational Opportunity Project) use NAEP scales to support grade level interpretations

The [Educational Opportunity Project](#) is my longstanding collaboration with my colleagues Sean Reardon, Erin Fahle, and Benjamin Shear. Our [widely used](#) Stanford Education Data Archive (SEDA) uses NAEP data to compare school and district average scores across states and over time ([Reardon et al., 2021](#)). We offer results in both “grade levels” familiar to general audiences and “standard deviation units” familiar to researchers. How do we convert NAEP scale score units to grade levels?

Using NAEP scales to measure progress in “grade levels” relies on one key assumption: There must be a defensible basis for comparing (linking) Grade 4 and Grade 8 scores.

To meet this essential assumption, we rely on the initial design of modern NAEP dating back to the dawn of NAGB. Early NAEP provided a basis for comparing Grade 4 and Grade 8 scores by administering common questions to both 4th and 8th graders, in 1990 in mathematics and in 1992 in reading ([Thissen, 2012](#)). These common questions act as a kind of bridge spanning grades 4 and 8. The common questions give us a basis for saying that an average 8th grader in 2009 (a score of 280, see below) would also have scored a 280 had they taken the 4th grade test. The common questions also give us a basis for saying that the average 4th grader in 2009 (a score of 238) would have scored a 238 had they taken the 8th grade test. We can therefore conclude that the 4th grader who scores a 280 (at the 8th grade average score) has the same score as an average 8th grader in mathematics. And we can claim that a state that has an average score of 280 is therefore 4 grade levels higher than a state with an average score of 238.

Table 1. Average scale scores (means) and standard deviations (SDs) of NAEP Mathematics scores for U.S. public school students, 2009-2019.

	Grade	2009	2011	2013	2015	2017	2019
Means	8	280	281	283	280	281	280
	4	238	239	240	239	238	239
SDs	8	38	37	37	38	40	41
	4	30	30	30	31	32	33

These “grade level” interpretations are inherently relative to a “reference group.” This is not controversial, as long as the reference group is clear. All the claims from the previous paragraph are referenced to 2009. The claims would be slightly different had we selected a different year or cohort. For example, in 2013, as you can see in Table 1 above, a score of 280 would be slightly below the 8th grade average score. In our work, we average over multiple cohorts of data. For example, Table 1 shows that the average distance between grade 4 and grade 8 scores is roughly 42 NAEP scale score points. We can define a grade level therefore as roughly $42/4 = 10.5$ NAEP points. In our EOP/SEDA work, we report trends in terms of NAEP points converted to grade levels to improve accessibility and interpretability. By this metric, the pandemic-era Grade 4 math decline of 5 scale score points was half a grade of learning,

and 8 scale score points was 3/4 of a grade of learning. By establishing some assumption about the number of months in a school year (e.g., 9), a “NAEP months of learning” metric could be established.

Recommendations for a special study to gather validity evidence for the cross-grade scale

Thissen ([2012](#)) notes that there have been few publicly available checks on the validity of cross-grade comparisons over time. One exception is a presentation by McClellan et al. (2005), who tested whether subsequent vertical linking would have made a difference on the reading assessment. They concluded that “the current cross-grade scale design used in NAEP seems stable to the alternate design studied” (p. 37). Thissen concluded his report by arguing, “validity evidence can and should be assembled to support, and make more precise, interpretive statements... [like] ‘how many NAEP scale points is one year’s growth?’ I agree wholeheartedly.

Thissen (2012) recommends administering items to adjacent grade populations in grades 3, 5, and 7. I would support this enthusiastically if budgets allow. I would also consider it valuable to simply administer cross-grade questions in grades 4 and 8 to re-establish the cross-grade scale. This is important because I am proposing a metric for “progress,” not a metric for performance at the 5th grade level, 6th grade level, and so on. To support an interpretation like, “Mississippi 4th graders are performing at the 5th grade level,” I need to know what 5th grade scores are nationally, or I need to assume linear growth between grades. However, I believe it is a lower bar to say, “Mississippi 4th graders gained 3 NAEP months of learning in mathematics from 2022 to 2024,” in the sense that maintaining this level of progress would eventually take them to grade 8 levels, without reference to performance at intermediate grades where there may be no basis for absolute comparability.

How should results of this special study be analyzed and actionable? If the cross-grade scale holds, that is, if Grade 4 questions embedded in Grade 8 tests function as models predict, and vice versa, this would be evidence supporting the existing cross-grade scale. It is also possible and perhaps even likely that items do not function as predicted, because grade 4 content is not covered in middle schools, and because grade 8 content is not covered in elementary schools. This would weaken the argument for cross-grade comparability, and NCES and NAGB may judge on the basis of the results that grade-level interpretations of educational progress do not have sufficient evidentiary support. This may motivate adjacent-grade studies like those Thissen (2012) recommended, longitudinal analyses using NAEP, and other research linking NAEP to other tests with stronger adjacent grade comparability.

Appropriate use and predictable misuse of “grade level” interpretations of “NAEP months of learning”

Grade-level reporting has a long history in educational measurement (e.g., [Hoover, 1984](#)). Grade levels are used rarely for student reporting due to widespread misinterpretation of results as indicating that students should be promoted or retained in grades on the basis of results. This is not relevant to my proposed use of NAEP grade-level reporting, because 1) NAEP reports scores only at aggregate levels, and 2) because I am not proposing interpretations of scores at absolute grade levels but instead interpretations of trends and progress in terms of grade levels and months of learning. Because of lessons learned from decades of grade-level reporting at the student level, however, I would generally 3) recommend against interpretations of achievement gaps between sociodemographic groups in terms of

grade levels. These may make disparities seem more immutable and reduce prioritization of equality of educational opportunity (e.g., [Quinn & Desruisseaux, 2022](#)).

If validity evidence supports cross-grade comparability, I would recommend cautious adoption that complements rather than supplants existing scale-score metrics of progress. This could motivate a research study exploring whether different metrics are more or less engaging to various audiences and whether interpretations are more or less accurate. Desired inferences could include progress, inequality (i.e., achievement gaps), changes in inequality, and cross-state comparisons. Some of these inferences have known mismatches with metrics; for example, achievement gap trends should never be measured by proficiency percentages ([Ho, 2008](#); [Polikoff, 2016](#)). Alternative approaches to communicate magnitudes of progress include Dollars and Counts as I mentioned previously, as well as market-basket reporting ([National Research Council, 2000](#)) and criterion-referenced benchmarking (e.g., [Briggs, 2019](#)).

Another metric for educational progress is known as the effect size (e.g., [Kraft, 2020](#)), which expresses a change in means in terms of standard deviation units. This metric is widely understood by researchers but offers limited interpretive value to general audiences. It is also underappreciated that effect sizes, like “months of learning metrics,” are scale-dependent and require a choice of reference to a particular standard deviation unit ([Ho, 2009](#)). Although some researchers have criticized months of learning metrics for being arbitrary and imprecise (e.g., [Baird & Pane, 2019](#)), these assume poorly specified reference groups and substantial grade-to-grade overlap which proposed NAEP validation studies would anticipate and avoid. In our SEDA research, we show that district rankings using standard deviation units and district rankings using grade-level units correlate above 0.99 ([Fahle et al., 2019](#)). This is a reassurance that grade levels are a simple linear transformation that improve interpretation with negligible risk to quantitative results.

Conclusion

I propose in this memo an approach and validation agenda for “grade level” and “months of learning” interpretations of NAEP trends. The proposal goes against my psychometric instincts to discourage accessible interpretations of scores. We psychometricians are generally trained to caution against overinterpretation, avoiding at all costs what I call the three fallacies of test score interpretation, that scores are more 1) meaningful, 2) precise, and 3) permanent than they are. What I learned in my time on NAGB is the importance of balancing this caution with our mission of not simply measuring but engaging the public around the assessment of educational progress. When a pandemic erases decades of educational progress, and people ask me, “is 3 points a lot,” I feel like we have missed an opportunity to report and disseminate as is this committee’s charge and this board’s commitment. I hope you will support evaluating and, if warranted, support adopting an accessible metric for educational progress to the NAGB and NCES reporting toolkit: NAEP months of learning.