Verifying Proficiency
Performance Indicators

> "We know that students will rarely perform at high levels on challenging learning tasks at their first attempt. Deep understanding or high levels of proficiency are achieved only as a result of trial, practice, adjustments based on feedback, and more practice."
>
> —Jay McTighe, "What Happens Between Assessments?," *Educational Leadership*

In a proficiency-based system, assessment, grading, and reporting practices are designed to (1) accurately measure and describe the knowledge and skills students have acquired, and (2) emphasize and encourage learning growth over time. For these reasons, the achievement of specific learning standards is tracked and reported by teachers, which requires that grade books be reformatted to report assessment results by standard, rather than report results by test or assignment. In Proficiency-Based Learning Simplified, we call the standards for a course or learning experience performance indicators to distinguish them from graduation standards—the standards students must demonstrate to be eligible for grade promotion or a diploma.

Verifying achievement of performance indicators is derived from a student's performance on assessments over time. The achievement of performance indicators requires students to demonstrate that they have acquired the knowledge and skills—i.e., the learning objectives or learning targets—addressed in units and lessons, which are reported in course-based assessment scores.

When designing assessments, teachers begin with the performance indicators that a specific assessment is intended to address (a process generally known as "backward design"). If an assessment is intended to measure four performance indicators, for example, teachers create four entries in their grade books—one entry for each performance indicator—and scores on the assessment are reported for each indicator.

To determine the extent to which students have demonstrated achievement of performance indicators, the Great Schools Partnership recommends that scores be calculated in a way that assigns the greatest weight to *the most recently assessed student work*. In this way, students are not penalized for poor performance earlier in a term when more recent assessments indicate they have met or exceeded expectations.

The three most widely used grading options that assign greater weight to more recent assessment results are *Power Law*, *Decaying Average*, and *Most Recent Score*.

> ***NOTE:** While the Great Schools Partnership recommends the use of the following three methods, both power law and decaying average may require districts and schools to heavily modify existing grading systems or invest in specialized online systems—both of which could have financial implications. While the Great Schools Partnership does not endorse any specific grading platform or product, we have created a guide to selecting online grading and reporting systems that will be useful to districts and schools.

Method	Description	Pros	Cons
Power Law	The power-law formula plots different assessment scores over time and attempts to draw a "best-fit" line that effectively answers the question: What score would the student most likely receive on the performance indicator if she were assessed again?	Power law does not penalize students for poor performance at the beginning of a grading period, and it produces scores that more accurately reflect what students know and can do at the end of a semester or year.	Because the formula generates a predictive trend, it's possible that power law could produce, in some cases, a final score that is higher than the highest score earned by a student.
Decaying Average	Decaying-average formulas assign progressively decreasing weight to older assessment scores. In effect, newer assessments "count more" in the final score.	Because skills and knowledge increase over time, giving more weight to more recent assessments can facilitate the learning process and encourage teaching practices that are focused on learning growth.	Decaying averages introduce the possibility that students may not try as hard on some assessments given earlier in a grading period.
Most Recent Score	Teachers use the most recent assessment score (or scores) to determine if students have achieved performance indicators.	Using the most recent assessment score encourages students to improve their performance because new assessment results replace older results, and final grades will more accurately reflect the knowledge and skills they acquired over the course of a term.	Some teachers are uncomfortable using systems that replace older scores because they believe that students may not give every assessment their best effort if they know that some grades won't "count" or that they will be allowed to redo or retake assessments.

How Power Law Works

While the power-law formula is mathematically complex, and requires specialized grading systems, educators only need to know how it works and how to interpret scores (for a detailed explanation of the formula, see *Transforming Classroom Grading* by Robert J. Marzano).

Power law predicts what the student's next score will be based on the scores a student has already earned. In effect, power law answers the question: What score would this student most likely receive on the standard if she was assessed again?

In the table below, for example, the teacher is using a four-point rubric to evaluate proficiency on four distinct assessments. Four students in the class earned the same set of scores (1.00, 2.00, 3.00, and 4.00), but each in a different order. If the scores were averaged, all four students would receive a 2.5, but the power-law formula produces different aggregate scores because it generates a trend that places more weight on more recent assessments.

The following chart* provides a simplified illustration of how power law works in practice:

Assessment	1	2	3	4	Final Score	Proficiency Interpretation
Student 1	1.00	2.00	3.00	4.00	4.00	The scores show continuous improvement, and the student will likely demonstrate mastery on the next assessment.
Student 2	1.00	3.00	2.00	4.00	3.66	The scores show irregular improvement, and the student will likely demonstrate high but not complete mastery on the next assessment.
Student 3	2.00	4.00	1.00	3.00	2.16	The scores show very uneven performance, and the student will likely demonstrate a mid-level of achievement on the next assessment.
Student 4	4.00	3.00	2.00	1.00	1.28	The scores show continuous decline, and the student will likely demonstrate a low level of achievement on the next assessment.

***NOTE:** This section was adapted from useful explanations created by EasyGradePro and JumpRope.

How Decaying Average Works

A decaying-average formula gives more weight to more recent assessment scores. Decaying average is based on the assumption that students—with more instruction, support, and practice—will progressively increase their knowledge, comprehension, and skill, while also decreasing the frequency of errors and incorrect answers. The formula is intended to produce scores that more accurately reflect learning progress on performance indicators—i.e., where students end up, rather than where they started out.

One of the benefits of decaying average is that it can be used with as few as two assessment scores. And unlike power law, which uses a complex mathematical algorithm, decaying average is relatively easy to explain to students and parents. Districts and schools can determine the weight used in the formula, but it needs to be at least a 60-percent weight on the most recent assessments to produce reliable scores.

If a teacher is using decaying average with a .65 weight, for example, and a student takes two assessments and earns scores of 2.00 and 3.00 for a performance indicator [.35(2) + .65(3)], the final score would be a 2.65 (or below proficiency). If the student then takes a third assessment and earns a score of 4.00 [.35(2.65) + .65(4)], the recalculated score would be a 3.53 (or above proficiency). Notice how the formula takes the *last recorded proficiency level* (not the last recorded assessment score) and weights it by .35 to produce the “decaying” average.

> ***NOTE:** There are a variety of ways to calculate decaying average, and online grading systems may offer multiple options. For example, some may offer multiple weight options or allow teachers to assign more weight to certain assessments or types of assessments. For this reason, districts and schools should always review all available options and ask questions to determine whether a specific product or platform will suit a school’s instructional needs and goals.

The following chart* provides a simplified illustration of how decaying average works in practice:

Assessment	1	2		3		4	Final Score	Proficiency Interpretation
Student 1	1.00	2.00	1.65	3.00	3.00	4.00	3.65	The scores show continuous improvement, and the student's proficiency level reflects learning progress made during the grading period. The final score indicates the student's current proficiency level, while also factoring in the student's less successful demonstrations at a diminished weight.
Student 2	1.00	3.00	2.30	2.00	2.10	4.00	3.33	The scores show irregular improvement, which suggests that the student may not have understood an important concept or that outside factors may have adversely affected the student's performance. If a low score is misrepresentative, the student's proficiency level will quickly go up after scores improve on additional assessments.
Student 3	2.00	4.00	3.30	1.00	1.80	3.00	2.58	The scores show very uneven performance. While the student demonstrated proficiency on the last assessment, the current score recognizes that the student has not met the standard with enough consistency to be considered proficient at this time.
Student 4	4.00	3.00	3.35	2.00	2.47	1.00	1.51	The scores show continuous decline. If the student's scores were averaged, the final score of 2.5 would reflect an inflated proficiency level, given the student's most recent assessment results. The decaying average more accurately represents the student's declining assessment results.

***NOTE:** This section was adapted from useful explanations created by EasyGradePro and JumpRope.

How Most Recent Score Works

In some schools, teachers use scores on the most recent assessment (or assessments) to determine proficiency on performance indicators. The method is based on the assumption that a student's most recent performance is representative of the knowledge and skills he or she has acquired.

When deciding whether to use the most-recent-score method, school leaders and teachers should consider the structure of the curriculum to ensure that the approach will accurately reflect student learning progress and achievement. For example, the method tends to work best with skill-based standards that require students to refine and improve their abilities over time. With some content-based standards that are demonstrated at a specific point in time and only once, the method may produce less accurate results.

When most recent score is used to determine proficiency, students can quickly recover from poor assessment scores that failed to meet expected standards, while students who met standards initially must also maintain their high performance. That said, the method could produce less representative or accurate proficiency levels when scores are uneven.

The following chart provides a simplified illustration of how most recent score works in practice:

Assessment	1	2	3	4	Final Score	Proficiency Interpretation
Student 1	1.00	2.00	3.00	4.00	4.00	The scores show continuous improvement, and the student's proficiency level reflects that progress.
Student 2	1.00	3.00	2.00	4.00	4.00	The scores show irregular improvement, and the final score may or may not reflect the most accurate proficiency level in some cases.
Student 3	2.00	4.00	1.00	3.00	3.00	The scores show very uneven performance. While the final score meets the standard, the student's proficiency level may not be entirely clear in some cases.
Student 4	4.00	3.00	2.00	1.00	1.00	The scores show irregular improvement, and the final score may or may not reflect the most accurate proficiency level in some cases.

Alternative Methods

Some schools choose to use alternative methods and formulas in their proficiency-based systems, including *Mean*, *Mode*, and *Highest Score*.

> ***NOTE:** These three options are described here for informational purposes only—the Great Schools Partnership does not recommend the use of these methods.

Method	Description	Pros	Cons
Mean	All assessment scores are averaged together to determine proficiency.	This method will be familiar to teachers, students, and parents because it has historically been the most common grading method used in schools.	Averaging can distort and misrepresent proficiency, particularly when students make significant progress over the course of a grading term.
Mode	The most common score is used to determine proficiency.	Mode is relatively easy to explain to parents and students.	If the grading scale used by schools has a lot of graduations, the mode is much more difficult to calculate and may not accurately reflect a student's proficiency level.
Highest Score	The highest score achieved by a student is used to determine proficiency.	This method could encourage students to take risks in their education and explore more challenging learning opportunities after they have demonstrated proficiency.	The highest score may not accurately reflect a student's level of knowledge and skill, especially when performance is inconsistent.

How Mean Works

Most traditional assessment systems are based on the average (or mean) of all grades a student earns—scores are added up and divided by the total number of scores. In some schools, teachers may assign more weight to certain assessments or types of assessments (such as homework scores vs. test scores), or they may decide that a greater percentage of student's final course grade will be based on certain types of assessments (for example, the score on a final project may count for 25 percent of a student's final grade).

While averaging *successful* assessment scores provides a more representative picture of the knowledge and skills students have acquired, averaging *all* scores can distort and misrepresent student proficiency and learning progress. For this reason, some schools choose to delay the numerical grading of assessments—by using placeholders such as "not met" or "insufficient evidence"—when averaging. In these cases, teachers will provide additional opportunities for students to redo assessments or improve the quality of their work.

In general, the Great Schools Partnership does not recommend the use of averaging to determine the achievement of performance indicators for three primary reasons:

1. **Averaging may not accurately reflect academic effort, learning growth, or end-of-term proficiency.** When scores are averaged at the end of a reporting period, the results may penalize students for poor assessment scores at the beginning of a term—even if they worked hard, improved their performance, and ultimately demonstrated proficiency. Even when averages are weighted to distinguish between formative and summative assessments or "major" and "minor" assessments, the results may still provide a distorted representation of achievement and proficiency.
2. **Averaging may introduce a disincentive to improve.** If students fail a few assessments at the beginning of a term, these early failures will impose clear mathematical limits on the final grade they can earn. Consequently, students may be less motivated to work hard or overcome past failures because their final grades won't reflect their effort and learning progress.
3. **Grade averaging advantages students who begin a course prepared and disadvantages those who begin unprepared.** Because effort and learning progress may not be accurately represented in averaged grades, students who begin school with more education, skills, resources, or family support have a strong advantage—in terms of their likelihood of earning a good grade—than students who arrive less prepared. And because academic readiness tends to mirror demographic factors such as socioeconomic and minority status, grade averaging also raises concerns about educational equity.

The following chart provides a simplified illustration of how averaging works in practice:

Assessment	1	2	3	4	Final Score	Proficiency Interpretation
Student 1	1.00	2.00	3.00	4.00	2.50	The scores show continuous improvement, but the final score does not reflect the significant learning progress made by the student—instead, it suggests that the student has failed to meet proficiency.
Student 2	1.00	3.00	2.00	4.00	2.50	The scores show irregular improvement, but the final score does not meet proficiency.
Student 3	2.00	4.00	1.00	3.00	2.50	The scores show very uneven performance. While the average score is somewhat representative the student's proficiency level in this case, the other averaged scores are clearly misrepresentative.
Student 4	4.00	3.00	2.00	1.00	2.50	Even though the scores show continuous decline, the student receives the same final score as Student 1, who made clear and significant improvement.

How Mode Works

The mode is the most common result in a given data set. While the mode is relatively straightforward and easy to explain, many people confuse "mode" with other mathematical terms like *mean* (average) and *median* (middle value).

For most performance indicators, teachers will have more than one assessment result to consider, and the most common score achieved by students may be used to determine proficiency in some schools. Yet when teachers have a limited data set (i.e., fewer scores), when they are using grading scales with more gradations (such as 1–100 scales), or when scores

are widely discrepant, the mode may produce misrepresentative results. For example, a student who scored a 1.00 on the first three assessments, a 3.00 on the next two assessments, and a 4.00 on the final two assessments would receive a final score of 1.00 even though the majority of the assessment results demonstrated proficiency. In this case, the student's learning growth over the grading term would also not be reflected in the final score.

The following chart provides a simplified illustration of how the mode works in practice:

Assessment	1	2	3	4	Final Score	Proficiency Interpretation
Student 1	1.00	2.00	2.00	4.00	2.00	The most common score is 2.0.
Student 2	4.00	1.00	2.00	4.00	4.00	The most common score is 4.0.
Student 3	2.00	4.00	1.00	3.00	IE	Because there is no "most common score" in this data set, more evidence is needed to verify proficiency ("IE" in this case stands for *insufficient evidence*).
Student 4	1.00	3.00	4.00	1.00	1.00	The most common score is 1.0.

How Highest Score Works

The highest-score method is easy explained: the highest assessment score achieved during a grading period is the student's final score for a performance indicator.

While highest score is easy to use, the method will produce misrepresentative results in many cases. For example, if a student scores a 4.00 on one assessment and 1.00 on all other assessments, the highest score (4.00) may not accurately reflect a student's level of proficiency. In addition, the method does not take into account a student's learning growth over the grading term. The advantage of highest score is that it recognizes a student's best work, while the disadvantage is that it may not accurately represent uneven performance.

The following chart provides a simplified illustration of how highest score works in practice, while also revealing the clear disadvantage of the approach:

Assessment	1	2	3	4	Final Score	Proficiency Interpretation
Student 1	1.00	2.00	2.00	4.00	4.00	Because all students achieved a score of 4.00 at *some* point during the grading period, all students earned a 4.00 even though the performance patterns from student to student are clearly dissimilar and representative of different levels of proficiency.
Student 2	4.00	4.00	4.00	4.00	4.00	
Student 3	1.00	4.00	1.00	1.00	4.00	
Student 4	4.00	3.00	2.00	1.00	4.00	