Standards-Based Grading: History Adjusted True Score

Jay Hooper and Ryan Cowell
Champaign Unit 4 School District

There has been much research and discussion on the principles of standards-based grading, and there is a growing consensus of best practice. Even so, the actual process of implementing standards-based grading at a school or district level can be a significant challenge. There are very practical questions that remain unclear, such as how the grades should be calculated. Common methods include a simple average, averaging only the more recent scores, mathematical models of growth over time, and basic teacher judgment. It is difficult to choose a single method that is justifiable in all circumstances. This article proposes a new method, the history-adjusted true score, that can be applied in a wide variety of situations. A rigorous method of evaluating each grading method is also described.

For more than two decades, discussions of standards in education have taken on an increasingly prominent role. Guskey (2009b) described the power of standards to offer direction to education reform when he stated,

In education, “standards” represent the goals of teaching and learning. They describe what we want students to know and be able to do as a result of their experiences in school. Well-defined standards identify the specific knowledge, skills, abilities, and disposition that we hope students will acquire through interactions with teachers and fellow students in school learning environments. (p. 1)

Guskey explained that the inherent promise of well-defined standards led many professional organizations to develop discipline specific standards. Individual states also worked to identify learning standards for their students. Presently, there is significant work and discussion around the creation and implementation of national standards, the Common Core State Standards. A 2010 report from The National Governors Association Center for Best Practices and The Council of Chief State Schools Officers states that 48 states, the District of Columbia, Puerto Rico, and the Virgin Islands have voluntarily agreed to share Common Core Standards in
English language arts and mathematics. The report indicates that “the ultimate goal is for all American children to graduate from high school ready for college, career pathways, and success in a global economy” (p. 1).

In addition to the creation and implementation of standards, the standards movement in education has caused many to think deeply about grading and reporting practices that reflect student proficiency of identified standards. Consequently, there is increasing discussion around the implementation of grading practices that differ significantly from traditional grading practices. Traditional grading practices are defined, for the purpose of this article, as students earning points on various types of assignments and assessments throughout a grading period and a teacher averaging those points on a 100-point scale to determine a student’s overall grade.

AN ARGUMENT FOR STANDARDS-BASED GRADING

Purpose for Grading

One of the inherent challenges with the traditional grading system is that the purpose for this grading system is not clearly defined. Within this system, a grade can be used to serve myriad purposes: communicating student learning, communicating student effort, sorting and selecting students, motivating students, and punishing students (Guskey, 2009a; Marzano & Heflebower, 2011; O’Connor, 2009). In addition to grades in a traditional grading system serving multiple purposes, Guskey (2009a) pointed out that teachers explicitly and implicitly assess many different types of evidence to determine a student’s grade when agreed-upon standards and indicators are not the sole measure of student learning (p. 17).

In essence, when a specific group of standards and learning targets are not used to measure student learning, then almost anything can be assessed and become part of a student’s grade. Some typical sources for determining students’ grades in a traditional grading system Guskey identifies are as follows:

- Major exams or compositions
- Homework completion
- Class quizzes
- Homework quality
- Reports or projects
- Class participation
- Exhibits of students’ work
- Effort
- Laboratory projects
- Attendance
- Students’ notebooks or journals
- Punctuality of assignments
- Classroom observations
- Class behavior or attitude
- Oral presentation
- Progress made
Because a traditional grading system typically requires a teacher to report a single grade for each student, the result is a grade that is derived from many different types of evidence that is meant to represent multiple aspects of performance or behavior and consequently rarely represents student proficiency with course content (Conley, 2000; Guskey, 2002; Guskey, 2009a). In this case, a lack of a singular, defined purpose for student grades leads to grades that do not clearly communicate anything (O’Connor & Wormeli, 2011). O’Connor (2011) stated that the primary purpose for grades is “communication about achievement, with achievement being defined as performance measured against accepted published standards and learning outcomes” (p. 7). This purpose is the presumed purpose for grading henceforward in this article.

Shortcomings of Traditional Grading Practices

Beyond those just described, this purpose for grading, communicating achievement on accepted standards and learning outcomes or targets, is at odds with the traditional grading system in several fundamental ways. One significant shortcoming of the traditional grading system is the inability to communicate students’ proficiency on established standards when assessment and grading is organized around assessment number or type rather than around standards. Another shortcoming is the inherent ambiguity in communicating student learning when averaging all student scores to determine a single final grade. These shortcomings are clearly evident when examining the following example of two students’ scores on 10 consecutive assessments.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>80</th>
<th>80</th>
<th>80</th>
<th>80</th>
<th>80</th>
<th>80</th>
<th>80</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student 1</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>Average</td>
</tr>
</tbody>
</table>

The first fundamental shortcoming with the traditional grading system made evident by the preceding example is that the student scores fail to communicate what each student really knows. In this example, it is easy to see how each student performed on each of the 10 assessments. For example, Student 1 received an 80 on Assessment 10, but it is impossible to determine the extent to which Student 1 has mastered course content as defined by standards and learning targets. Stiggins, Arter, Chappuis, and Chappuis (2006) argued that grade books should be organized “according to achievement target, not according to type or source of information such as tests, quizzes, labs, or homework” (p. 311). This allows the grade book to clearly communicate student performance on each identified standard or learning target as opposed to student performance on specific assessments or types of assessments.

The second fundamental shortcoming with the traditional grading system made evident by the preceding example is the inherent ambiguities in communicating student learning when averaging all student scores to determine a single final grade. The averaging of all or most assignments on a 100-point scale insufficiently communicates what a student knows and is able to do (Brookhart & Nitko, 2008; Guskey & Bailey, 2001; O’Connor, 2009; Reeves, 2011). In the preceding example, both students receive the same final grade, but it is difficult to argue that these data represent students with identical levels of understanding of course content. For example, Student 2, a student who never scored higher than 64 on any assessment, has the same final grade as Student 1, a student who scored 80 on the final eight assessments.
There are two flaws at play here. First, when averaging on a 100-point scale, the use of zeros is particularly damaging to a student’s grade. It is impossible to determine the reason for the zeros received in the preceding example, but typically these scores are reserved for students who have not turned in an assignment. O’Connor (2011) noted that “the mathematical problem with zeros is that they represent very extreme scores and their effect on the grade is always exaggerated” (p. 96). O’Connor (2011) went on to explain that this is because zeros are typically used on a grading scale similar to the following:

- A = 90–100%
- B = 80–89%
- C = 70–79%
- D = 60–69%
- F = below 60%

This means that there are 10 points for an A, B, C, and D, and 60 points for an F. Consider the implications on Student 1’s grade from the preceding example. If, for example, Student 1 received zeros because he or she did not complete Assessments 1 and 2, then the zeros communicate nothing more than a lack of completion and do nothing to communicate proficiency on standards and learning targets. In this scenario it is difficult to justify the use of zeros if the goal of grading is to communicate students’ proficiency on standards and learning targets. If, on the other hand, Student 1’s zeros represent the quality of his or her work on Assessments 1 and 2, then there are still mathematical concerns with the use of the zero relative to the established grading scale (O’Connor, 2009, 2011).

The second flaw at play here is that the grading practice of averaging each student’s scores does not allow the grade to communicate a student’s present level of performance because it does not reflect learning at the current time. Stiggins et al. (2006) claimed that grades should be based on

the most current evidence of the student’s level of achievement on intended learning outcomes.

The practice of averaging over an entire grading period does not yield a summary of current level of achievement for learning targets that reflect continuous growth—reasoning, skills, and products, such as writing, oral presentation, experimental skills in science, or speaking a foreign language, for example. (p. 314)

If it is presumed that students learn over time, then Student 1 and Student 2’s scores represent clear differences in understanding between these two students, with Student 1 performing at a significantly higher level than Student 2 on the final eight assessments of the grading period. If the presumed purpose for grading is to communicate student learning on identified standards, then the traditional approach to grading, as just described, is inherently flawed.

Subjectivity of Grading

Despite obvious flaws, the traditional grading system persists, in large part, due to habits of practice. This traditional system is so entrenched that many believe that, if nothing else, it is
objective. This belief is simply not true when one considers the human element of curriculum, instruction, and assessment present with all grading practices. O’Connor (2009) wrote,

> Grades are as much a matter of values as they are of science; all along the assessment trail, the teacher has made value judgments about what type of assessment to use, what to include in each assessment, how the assessment is scored, the actual scoring of the assessment, and why the scores are to be combined in a particular way to arrive at a final grade. Most of these value judgments are professional ones; these are the professional decisions that teachers are trained (and paid) to make. It should be acknowledged that grades are, for the most part, subjective, not objective judgments. (p. 19)

O’Connor’s observations are important because grading in any form, as part of the larger human activity of teaching and learning, is inherently subjective; consequently, there is no perfectly objective grading system. The goal, then, is not to undertake the impossible task of developing an objective system but rather to adopt a grading system that best communicates student learning.

Brookhart (2011) wrote, “Standards-based grading is based on the principle that grades should convey how well students have achieved standards. In other words, grades are not about what students *earn*; they are about what students *learn*” (p. 12). This statement echoes the work of many others (Bailey & Guskey, 2001; Cooper, 2001; Marzano, 2006; O’Connor, 2009; Tomlinson & McTighe, 2006) and strongly suggests the use of standards-based grading, described by O’Connor (2009) as a system where grades are directly related to standards and learning targets for each grading period, as the most appropriate method (Brookhart & Nitko, 2008; Marzano, 2010).

Logistical Concern with Standards-Based Grading

Although there is a continually growing consensus around the validity of standards-based grading as best educational practice, several important logistical concerns remain. One such concern is the way a student’s final grade, or true score, on each standard or learning target is determined. Marzano (2006) described the concept of a “true score.” He argued that a student’s performance on a test item (or any assessment) is an indicator of a student’s level of understanding and skill, but one observation is not in itself definitive. A student could be affected by environmental issues or provide seemingly correct responses with little true understanding (Marzano, 2006, p. 38). Because of these uncertainties, a true score is inherently unknown. Therefore, all measurements of a true score for an individual student are considered to be an estimate. When a “final grade” is discussed henceforward in this article, it is interpreted to be an *estimated* true score for a specific standard.

There are, of course, many different approaches for determining a student’s final grade within a system of standards-based grading, making it a daunting task for a school or district to decide upon the best approach. Some approaches work well in specific circumstances. Others work well in most circumstances but are particularly subject to bias. None are perfect. This article will provide an analysis of several commonly used methods for determining a student’s final grade, propose a new alternative, and establish a method of comparing the performance of each.
METHODS OF CALCULATING A TRUE SCORE

Assumptions and Conditions

There are many different methods of implementing standards-based grading. It will be helpful to define specific grading conditions for the remainder of the article. Grades, as discussed in this article, are recorded for a single standard. There would be multiple standards within a course, but this article is focused on the calculation of grades within a single standard with multiple assessments of the same standard. For a standard or learning target to be assessed multiple times, it might be something broad like a standard cluster from the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) such as, “Write expressions in equivalent forms to solve problems” (p. 64) or “Analyze functions using different representations” (p. 69). We will assume that all grades are on a scale from 0 to 4. We will allow half-points (a student could score a 3.5) on individual assessments but no finer detail than half-points (a student could not score a 3.6 on a single assessment linked to a single standard). Even though individual grades are scored to the nearest half-point, we will not constrain the final grade calculation to the same rounding (a final grade summary for a single standard could be reported to be 3.62). The analysis and results of this article are specific to this choice of grading scale, but the methods could be replicated in other grading environments.

A statistician is always interested in the assumptions being made before any statistical procedure is conducted. For standards-based grading, this article makes the following assumptions about individual scores on an assessment linked to a standard.

- Observations of individual scores are independent.
- Individual scores (observations) reported for each standard are similar in their ability to estimate a true score.

In addition, a set of criteria is needed to evaluate each method of calculating a final grade (predicted true score). We (the authors) suggest the following criteria:

- The predicted true score should be within the range of observed scores.
- The most recent scores should be considered as more accurate estimators of the current true score.
- The history of scores should be considered (no significant range of scores is completely ignored).
- The method should be robust enough to apply to all reasonable scenarios, taking into account the need for professional judgment regarding the validity of individual scores.

A Simple Average

One commonly used method for determining a student’s final grade is to use a simple average (arithmetic mean). As previously discussed, a simple average may not accurately communicate what a student knows and is able to do. In terms of the four criteria just listed, the simple average ignores the second criterion. All scores are considered equal, and no attempt is made to
calculate the *current* true score. High levels of understanding at the present time are “averaged” with earlier observations.

### Average of the Most Current “n” Scores

An improvement on the simple average is to average the most current “n” scores. For example, the two most current scores could be averaged and reported as a summary of each standard or learning target. This is a better approach to calculating grades. It will more accurately reflect a student’s true score at the current time.

There is an important flaw in this type of calculation, however. By ignoring the previous scores in the calculation, it does not consider the history of scores (the third criterion). For example, a student could score (over time) a 4, 4, 4, and a 2. The final score of a 2 could have been due to any number of factors outside of what the student knows and is able to do. Averaging the last two scores would assign a final score of 3. The strong history of 4s is ignored. This could be overcome by averaging more than the last two scores, but as more and more scores are averaged, the same problems arise as with the average of all scores. With this method one is left with the dilemma that it might make sense for one student to average more scores, but it might not make sense for others. In addition, if a student knows that only the final “n” scores will be factored into the grade, the student may not be encouraged to put forth a sincere effort near the start of a grading term. This could be another consequence of ignoring the third criterion.

### Mathematical Models of Growth Over Time

In response to rejecting a simple average, mathematical models of growth over time could be used to estimate a current true score. These models attempt to describe the observed growth over time with a mathematical function such as a linear, exponential, logarithmic, or power function. The mathematical function is then used to make a prediction of the student’s true score at the current time. However, we would suggest that these models (the power law being the forerunner) should not be applied as a hard rule for all students. Using the power law as a model, the following assumptions are made:

- The logarithms of a student’s scores, when plotted against the logarithms of the assessment numbers (time), form a linear pattern.
- A student’s growth, although not necessarily linear, can be predicted by a single mathematical function.

Although these assumptions may account for a wide variety of student growth, these assumptions cannot be applied with integrity to many situations.

Consider Figure 1. Fitting a power-law \((y = a \times x^b)\) curve to a pattern of growth over time can produce a predicted true score that is higher than any individual score that the student has achieved. Similarly, the model can also predict a true score lower than any actual score achieved by the student. When using repeated scores for a given standard or learning target, a student’s predicted true score should axiomatically be within the range of observed scores (the first criterion established previously).
FIGURE 1 Each integer point represents an actual assessment score. Note. The curve is generated by modeling the scores with a function of the form $y = a x^b$. The predicted true score at the current time is labeled (color figure available online).

Furthermore, the first score recorded for a student can have a significant impact on the final predicted true score (Figure 2). In the figure, observe how the first score changes with the other scores remaining constant. The predicted true score decreases each time the first score increases. With virtually any single mathematical model (linear, exponential, power, etc.), a student would be encouraged to fail the first assessment. This would violate the notion that individual scores should be independent.

No single mathematical function, linear or nonlinear, will be able to systematically model the complexity of actual individual student growth for all students. The process of learning is too complex with too many variables. For example, a student may suddenly spend a significant amount of time studying for a particular standard or learning target several weeks into a grading period. This sudden, unpredictable change would imply a sharp turn (cusp) in the growth model that no single regression technique will be able to model and predict. The best option in this scenario would be to start using a second model to predict growth at the instant the unpredictable studying event occurred.

The Method of Mounting Evidence

Marzano also documented the method of mounting evidence as a means of predicting a true score. Essentially a teacher would look at an individual student’s score history and decide that there is “mounting evidence” for a certain score (Marzano, 2006, pp. 100–101). By the four evaluation criteria, this method is the strongest of the commonly available methods. It meets all four of the criteria. Although this method may currently be the strongest available method, it is prone to teacher bias, intentional or unintentional. For example, a teacher may remember the poor behavior of a student in class and choose to assign a lower grade. A teacher may also know that a student has been recently trying very hard and choose to assign a higher grade, even though the score history does not justify the grade. Any bias of this kind leads to a grade that does not communicate student proficiency on a standard or learning target. This method could be available as an override based on professional judgment, but it would be ill advised to apply this method universally.
FIGURE 2  The first score varies from 1 to 4 with the three remaining scores remaining constant (color figure available online).

PROPOSED METHOD—FEWER ASSUMPTIONS

Nonparametric Method

The concept of a simple average is rejected because it does not account for student growth over time. The methods of mathematically modeling growth over time make too many assumptions about the complex process of learning. The method of mounting evidence is prone to bias. What is needed is a method of evaluating the evidence that makes fewer assumptions.

In statistics, there are many procedures that are based on statistical distributions such as the normal distribution. These procedures make assumptions about the data that need to be verified before a valid inference can be made. If these assumptions cannot be made given a specific set of data, the researcher cannot, with integrity, carry out the procedure. This same rule of integrity should be applied to the way a student’s final grade is determined.

A reader knowledgeable in statistical inference would likely turn attention to nonparametric methods of inference at this point. A simplified description of this branch of statistical inference is that the procedures are not directly connected to any specific distribution and therefore make fewer assumptions about the data. We need to create a similar method for standards-based grading that does not use a particular distribution (or mathematical family of functions). One of the most basic nonparametric procedures is the sign test.
In a sign test for the median, each observation is given a sign (positive or negative) depending on whether it is above or below the value being tested as the hypothesized median. We propose a similar approach to estimating a current true score. No assumptions need to be made about the form of the data (such as linearity of logarithm transformations). Because no specific distribution or mathematical model is used, the assumption that a single model fits the data is no longer needed. These are significant improvements.

Detailed Description of the Proposed Model

Under the assumption that individual scores are time-based estimators of the true score, if the model is to consider growth over time and estimate the current true score, the most recent score should draw our attention. However, because this is simply an estimator of the true score, the single, most recent score is not sufficient. It is therefore proposed that the two most recent scores should be used in the calculation. Using more than two scores invites similar problems discussed previously with averaging.

If the two most recent scores are the same, then a strong argument is made for the value of the current true score. If they are different, it is reasonable to assume that the true score is within the range of the two most recent observations. The question remains as to where, within this range, the true score should lie. A simple formula for the proposed process cannot be provided because the process is an algorithm (a sequential list of well-defined instructions). The algorithm can be described, and a computer can easily be programmed to follow the algorithm.

The mechanics of the algorithm are as follows. The smaller of the two most recent scores is taken as the baseline score (call this score $x_b$). The median of the two most recent scores (call this value $x_m$) is calculated. Finally, the absolute value of the difference in the two scores ($\Delta x$) is calculated. If the true score is to be “pulled” toward the higher score, there should be evidence in the history of scores to determine how far it should be “pulled.” This is where the methods of the nonparametric sign test are applied.

To apply the process of the sign test, the history of scores is compared to the median value $x_m$. Scores above this value $x_m$ would be labeled with a positive sign. Scores below this value would be labeled with a negative sign. Scores equal to this median value $x_m$ are ignored. The number of signs is then tallied, and the number of positive signs is recorded. Using the results of the sign analysis, in this context, we are able to determine how far the true score should be “pulled” above the baseline. The proportion of positive signs out of the total number of signs determines the proportion of the difference ($\Delta x$) that is added to the baseline.

Consider the following example: If a student scores a 3, 4, 1, 2, 4 on a given standard or learning target, the two most recent scores are a 2 and a 4. The baseline score is $x_b = 2$, the median of the two is $x_m = 3$, and the difference is $\Delta x = 2$. Ignoring scores at the median of 3 in the history of scores, we have two scores above and two scores below the median (four scores in total being considered after ignoring the 3). Two out of the four signs are positive. We will therefore report a score of $2 + \frac{2}{4} (2) = 3$ as the predicted true score for this standard or learning target. For comparison, this same set of scores would produce a simple average of 2.8, and the power law would predict a score of 2.28.

Because the history of scores is used to determine the location of the predicted true score, this method is appropriately named the history-adjusted true score (HAT-score). A traditional symbol used for the sample mean is $\bar{x}$. The proposed symbol for a HAT-score is $x_H$. 
Criticisms of the Proposed Method

One possible critique of the HAT-score method is that if growth is continually happening over time and the observed scores are continually increasing, the HAT-score method will systematically underpredict the true score. The “history-adjusted” portion of the calculation will pull the predicted score closer to the second-most recent observation rather than the most recent observation. This criticism, although valid, is quite easy to overcome and is discussed in the next section. Regardless, the HAT-score method will produce a reasonable summary of the current true score, taking into account the score history.

One other critique of the HAT-score model is the unspoken assumption that a student starts at a baseline of the lower of the two most recent scores. It can be easily mathematically demonstrated that the method is exactly equivalent to starting at the higher of the two scores and using the proportion of negative signs to determine the predicted true score. The method is not biased toward lower or higher predictions by using the median of the two most recent scores in the sign analysis.

USING SIMULATION TO EVALUATE METHODS OF CALCULATING GRADES

Description of the Simulation Method

A rigorous analysis of any model is still possible even though a student’s true score is inherently unknown. A theoretical growth curve can be created, representing a student’s true score as it changes over time as learning occurs. Random times along the curve are selected to represent the student taking an assessment. The student’s score is then determined by the simulated growth curve at that time with an added (random) measurement error. The measurement error serves the purpose of modeling a real assessment. The power behind this method is the ability to simulate an actual assessment environment while knowing the exact true score at each moment in time. This provides the ability to compare the multiple grading methods available. Each theoretical growth curve represents a different type of student growth and conducting many iterations allows long-run patterns to emerge.

Procedure for Conducting the Simulation

We employ this simulation method within the grading context previously established. On an individual assessment, a student can only receive scores in increments of 0.5 with a scale of 0 to 4. Because individual assessments are discrete, we will model the theoretical growth curve with discrete levels at each moment in time. Figure 3 represents a possible theoretical growth curve for which a student’s true score steadily grows over time.

Next, we randomly choose times at which assessments occur. We will simulate five assessments, with the fifth assessment taking place at the very end of the growth curve. Consider the time axis to be scaled to begin at zero and end at one, representing the entire passage of time for this learning standard. We assume that no recorded assessments will take place during the
first fourth of the time line. We will randomly choose a time between each of the following intervals: 0.25, 0.40, 0.55, 0.70, 0.85. This will simulate the timing of four of the assessments. The fifth assessment takes place at 1 on the time line, indicating a final assessment.

Once the times have been randomly chosen, we use the theoretical growth curve to determine the actual true score at that time. A simulated measurement error is then added to each observation. For this simulation, the error was chosen to be normally distributed with a mean of zero and a standard deviation of 0.334. This would allow 99.7% of the observations to be within 1 (remember our grading scale is from 0 to 4) of the actual true score at that time. Finally, the scores with the randomly assigned errors are rounded to the nearest 0.5. These five scores now represent one simulated set of possible scores generated by the theoretical growth curve. An example is shown in Figure 4 where the plotted points represent simulated assessments.

FIGURE 3 A model of a student’s true score as it changes over time (color figure available online).

FIGURE 4 The points represent randomly chosen assessment scores with simulated measurement error at randomly chosen times. Note. This process will be repeated many times on the same growth curve (color figure available online).
Results of the Simulation

We can use this process to repeatedly generate possible scores for any theoretical growth curve. A computer simulation now allows all of the grading methods (with the exception of the method of mounting evidence) to be compared. The final true score (the actual true score for a student on a given standard at the end of a grading term) is what each grading method will attempt to predict. For each of the following results, this process was simulated 1,000 times. Each grading method is applied to the each of the 1,000 simulated sets of scores, and the predicted true score (based on the simulated data) minus the theoretical true score (as determined by the end of the theoretical growth curve) is recorded. The results for this first theoretical growth curve are displayed in Figure 5.

The vertical scale, as described, is the difference between the calculated (predicted) true score and the actual true score at the end of the growth curve (the student’s actual score at the end of the grading term). An ideal method would have a center of zero (the median line on the boxplot should be centered at zero) and have as little variability as possible (a shorter, more compact plot is more ideal). Dashed lines at 0.5 and −0.5 have been added to indicate the region that is within 0.5 of the final true score. In Figure 5, no method seems to stand out as superior, although it can be observed that all but the basic average have at least 75% of the observations (three of the four regions on the box-and-whisker plot) within 0.5 of the true score. The reader might also note that some regions of the box-and-whisker plots land on discrete increments. This is due to the grading environment where individual scores have discrete levels, such as when a teacher scores an assessment using a rubric. A similar analysis can be done in a continuous environment to remove this feature, but it was the intent of the authors to conduct this analysis in an environment with discrete levels to model the use of assessments scored by rubrics.

When the student shows steady growth, all of the discussed methods seem to underpredict the true score. The power-law has a median that is the closest to zero, which would make
FIGURE 6  The student remains at a final level of 3.5 for a slightly extended amount of time. Note. HAT = history-adjusted true (color figure available online).

sense because the growth pattern can actually be modeled by a linear function. The HAT-score method accurately predicts a true score if the last two measurements are taken close together in time or if the student is at a similar level for both measurements. The student likely cannot control the date of the assessments, but the student is more able to control the mastery of learning over time. Notice the effect in Figure 6 of lengthening the duration of the final true score, indicating a stable level of understanding for the student. The final score of 3.5 lasts a little longer on the time scale, creating a slight “plateau” at the end. Notice now that several of the methods now have a median closer to zero. This “plateau” at the end of a growth curve is a likely phenomenon. This represents a scenario where a student reaches a stable level of mastery before the final assessments.

There are many other possible scenarios to consider. For example, a student could have steady growth and then forget the material by the end of the grading term. Or a student could reach a peak, decline, and then grow again. These two scenarios are included in Figures 7 and 8. Each scenario could represent a different type of student. Although these three patterns do not represent all students, they would account for many different types of student growth. It is important to focus on the possible patterns of growth (such as a gradual increase over time) when testing the various grading methods. Actual student scores, although theoretically readily available, do not reveal the true score as it changes over time. The measurement error prevents the true score from being exactly known. By focusing on patterns of growth by using theoretical growth curves, many different sets of actual student data are represented. However, a more robust collection of qualitative data (such as student and teacher interviews) and quantitative grade data would provide another useful lens through which to compare the various grading methods.

Discussion of Results

In all of these scenarios, it can be observed that the HAT-score is within 0.5 of the true score for at least 75% of the observations. As previously demonstrated, a simple extension of the final
“plateau” on the theoretical growth curve causes the HAT-score to show a significant increase in its ability to serve as an unbiased estimator of the final true score (moving its center to zero). The method of averaging the two most recent scores behaves similarly.

Some scenarios just presented demonstrate that a certain grading method would be a poor choice for that type of growth pattern. For example, Figure 8 demonstrates that a basic average would be a poor choice in this context because the entire plot is outside of the ±0.5 band. Even the method of averaging the last three observations stands out as a poor choice in Figure 8. The HAT-score stands out as a reasonable choice in each scenario. This is a significant observation.

As noted, the grading method that performs the most similar to the HAT-score is the method of averaging the last two scores. This occurs because the HAT-score algorithm begins by...
averaging the two most recent scores. Figures 7 and 8 demonstrate significantly less variability for the HAT-score when compared to the method of averaging the last two scores. When comparing the HAT-score method with averaging the two most recent observations, the HAT-score method has a stronger foundation by design by taking into account the history of all scores (one of the previously established criteria). It also displays a different and desirable behavior in terms of long-run patterns, as seen in the decreased variability.

The presence of outliers also needs to be addressed for the figures. Some of the outliers fall within the ±0.5 band, but there are still some that fall outside this band. In terms of this simulation, the outliers occur as a result of the measurement error and the structure of the specific theoretical growth curve. With 1,000 sets of discrete scores generated using each theoretical growth curve, the features of the growth curve allow possible but unlikely combinations of scores as a result of the random timing and the measurement error. For example, an allowable time window within the simulation may just barely allow scores to be sampled before making a jump to the next discrete level of the true score. An unlikely large random measurement error could be added in addition to create an extremely unlikely random event. Enough repetitions bring about observations of these unlikely events. More attention should be paid to the median and interquartile range of each box-and-whisker plot, as the simulation is intended to uncover the overall central tendency of the various grading methods for comparison.

Comparing Grading Methods Using Student Data

The simulation method described in this paper allows patterns to emerge with each grading method. The HAT-score method is shown to be able to be applied in many different scenarios. Even so, there is some information that is not communicated in the box-and-whisker plots. Specifically, the reader may want to see actual student scores along with the grade calculations for each method. The following table of data (Table 1) allows the reader to see the application of each grading method to specific sets of scores. The scores are to be interpreted as sequential in time throughout a grading term. When viewing each individual set of scores, the reader could determine what a reasonable score should be (the method of mounting evidence), keeping in mind that the goal is to report the true score at the end of the grading term. Each calculation is based on four scores. The first two scores in each set of four have been limited to scores of 2 and 3 in order to provide a more compact table, rather than including all possible sets of four scores.

Upon visual inspection, the most recent scores should draw attention, but the remaining history of scores should not be ignored. The HAT-score method supplies a reasonable result in a large majority of situations. When it differs from what the reader may assign by visual inspection, it is likely due to an uncharacteristically low score near the end. Even in these cases, the score predicted is still reasonable, given the assumption that each assessment is similar in its ability to measure a true score. For example, the reader may disagree with the score of 1.75 assigned to the set of scores {2, 2, 4, 1}. If the final score is indeed a valid score of 1, then it could be argued that the score is deserved. If, for example, the student stayed up all night working on a project for another class and was simply exhausted, the teacher may have to use professional judgment if the validity of an individual assessment is questioned. An uncharacteristically low score may need to be excluded from the grade book or reassessed, no matter what grading method is used. It is important to note that the professional judgment
being exercised is in regard to whether an individual assessment is a valid measurement of the student’s true score and not whether the HAT-score algorithm produced an unreasonable result. The human element of teaching and learning is impossible to remove from grading.

CONCLUSIONS

Standards-based grading, as a philosophy, offers an improvement in the accuracy and relevance of grade reporting. Grades are neither inflated nor deflated by mistakes on homework,
completion grades, attendance, or behavior. Grades reflect mastery of standards. The intent of this article is to establish a method of evaluating the various options for determining a final grade (a predicted current true score) for repeated assessments of a single standard as well as to define a new method, the HAT-score, that reliably reports the current true score in a wide variety of scenarios.

The HAT-score can be described as a simple algorithm. It operates under the assumption that the two most recent scores serve as the best evidence of the current true score. The history of scores is then analyzed to determine which of the two most recent scores the predicted true score should lie closer to. If the validity of each assessment as an indicator of the true score is not under question, such as an uncharacteristically low score for a student due to nonacademic reasons, the HAT-score algorithm will produce a reasonable summary of the current true score in all scenarios.

The impetus for this article is the need for a satisfying method of calculating grades within a single standard or learning target. Of the current, widely used methods, only the method of mounting evidence (visually inspecting a score history and assigning a grade) can be applied in all standards-based grading settings. Existing formulas and algorithms fall short when programmed into grading software and applied to all students. After rejecting a simple average, an average of the more recent scores, and a mathematical model such as the power law, there is currently not another calculation method that could be implemented to calculate a current true score for a given standard. This problem can serve as a hindrance to schools and districts as they move toward standards-based grading because the method of mounting evidence allows the possibility for unchecked teacher bias. A system based on the method of mounting evidence would allow students with the exact same score history to be assigned different scores by different teachers (or even the same teacher) based on any number of unknown factors.

Much research has been done on the philosophy of standards-based grading. The HAT-score algorithm aligns with the philosophy of standards-based grading and makes fewer assumptions about the history of scores than alternatives such as the power-law. This article provides a simulation environment to examine the most widely used calculation methods in order to compare them to the HAT-score. However, more research is needed in the actual implementation of the HAT-score at a school, district, or state level to more fully understand the dynamic nature of student growth patterns in the environment of the classroom. Raw student grade data alone would not provide the context needed to understand this dynamic range because it would not take into account the human nature of teaching and learning. Because of this, further research should include both qualitative and quantitative teacher and student data.

REFERENCES


