***Special Issue**: Exploring Current & Future Grading Practices in K-12 Schools*
Research Article

JSARD
open access | est. 2016
Volume 11, Issue S1 (2026), pp. 105-120
Journal of School Administration Research and Development
ISSN: 2470-8496 Print/2470-850X Online
ojed.org/JSARD

A Matter of Standards: Addressing Inconsistencies in Grading Practices in High School English, Mathematics, Science, and Social Studies Courses

Allie Kollman
TrueMarks Education, USA
Erin Sinner
Hannah Nurmi
Cailen O'Shea
North Dakota State University, USA

ABSTRACT

This study examines the alignment between high school students' summative assessment scores and their final course grade in the point-based grading system to identify patterns of grade inflation or deflation. Using a non-experimental quantitative design, we analyzed gradebook data from 2,494 students in four subject areas at three high schools in the same Midwestern school district. Paired *t* tests revealed statistically significant discrepancies, with one-third of students receiving grades at least five percentage points different from their cumulative summative score averages, and 36% experiencing at least one letter grade shift. Grade inflation was notably more common than deflation. These findings raise concerns about the fairness and accuracy of point-based grading systems and support the need for more consistent, standards-based approaches that prioritize defined learning outcomes, transparent assessment criteria, and the separation of academic achievement from behavioral compliance to ensure grades accurately reflect student proficiency.

***Keywords*:** evidence-based grading, grade inflation and deflation, grading reform

Historically, grades have shaped K-12 education's approach to measuring student achievement, influencing everything from parental expectations to college admissions and scholarship decisions (Brookhart, 2004; Guskey, 2009). Grades carry significant weight in shaping a student's academic trajectory, including honor roll, course pathways, access to advanced courses, and eligibility for extracurricular programs (Griffin & Townsley, 2021).

Research shows that high school grades are a better predictor of four-year college graduation than standardized test scores (Allensworth & Clark, 2020; Galla et al., 2019). Given the high-stakes role of grades, it is critical that they accurately reflect students' learning. A course grade should communicate the extent to which students have achieved the course objectives (Allen, 2005), providing a fair and valid summary of what a student knows and can do (Brookhart et al., 2016; Linn & Miller, 2005; Marzano & Heflebower, 2011).

In practice, however, ensuring accuracy and consistency in grading remains a challenge. Teacher-assigned grades are often a composite of academic performance and non-cognitive behaviors (Brookhart et al., 2016; Guskey & Link, 2018), such as homework completion, class participation, effort, and behavior (Griffin & Townsley, 2021). Teachers commonly incorporate not only summative assessment scores but also homework completion, class participation, effort, and behavior when determining final letter grades (Guskey & Link, 2018; McMillan, 2018). As a result, grades may reflect a mix of achievement and compliance, creating a "hodgepodge" of disparate elements that do not consistently indicate mastery of standards (Brookhart, 1991).

This study investigates the alignment between students' summative assessment scores and their final course grades to determine the extent of grade inflation or deflation in current grading practices. While previous research on grade inflation

has often compared grades or GPAs to standardized test results or college outcomes, there is limited analysis of grade accuracy at the classroom level (Brookhart et al., 2016; Griffin & Townsley, 2021; Pattison et al., 2013). This study addresses that gap by examining how final grades, expressed both as numerical percentages and as letter grades, differ from students' summative assessment averages. Because summative assessments are designed to evaluate how well students have mastered specific learning standards (Brookhart et al., 2016; Marzano & Heflebower, 2011), they serve as a meaningful baseline for judging the validity of final grades.

By comparing final reported grades to the evidence of summative proficiency, we can identify whether grades are being systematically inflated or deflated relative to actual achievement. Such findings have direct implications for grading reform efforts, as they can reveal patterns of inconsistency that compromise fairness and transparency in student evaluation.

The purpose of this study is to determine the degree to which final course letter grades diverge from students' summative assessment scores, thereby uncovering potential patterns of grade inflation or deflation. Our research question is as follows: *To what extent do final grades, measured both as percentages and as letter grades, differ from summative assessment scores?*

By investigating this alignment, the study seeks to assess the accuracy of current grading practices and provide actionable insights for school leaders and educators committed to advancing standards-based and equitable grading systems and norms.

REVIEW OF THE LITERATURE

In this literature review, we examine grading practices, inconsistencies, grade inflation and deflation, and concerns of equity. We begin with a review of overall grading practices and transition into the full review.

Starch and Elliott (1912) revealed wide inconsistencies in grading, with teachers assigning vastly different grades to identical student work. More than a century later, little has changed. Grading practices still vary significantly from one teacher to another, even within the same school or district (Brookhart, 2004; Guskey & Link, 2018; McMillan, 2001). This variability stems from long-standing traditions, limited formal teacher training in effective grading methods (Guskey, 2009), and inconsistent or absent district policies outlining grading expectations. As a result, individual teachers often create their own systems, some relying on total point accumulation models, others using weighted categories such as tests, homework, or participation, leading to inconsistent criteria for calculating grades (McMillan, 2001).

Persistent Inconsistencies in Grading Practices

A key inconsistency is the inclusion of behavioral, or non-cognitive tasks, such as homework-related items or in-progress monitored learning, in academic grades. Griffin and Townsley (2021) define "employability scores" as points awarded for classroom behaviors rooted in present-day expectations, such as participation, citizenship, and timeliness, which are identified in this study as non-cognitive behaviors. While these behaviors are often valued in point accumulation grading models, they risk distorting the academic meaning of a student's grade because they can inflate or deflate grades regardless of actual proficiency of standards (Griffin, 2020; Guskey, 2009; McMillan, 2001).

To preserve the integrity of achievement data, researchers recommend reporting non-cognitive factors separately from academic grades (Brookhart, 2004; Feldman, 2018; Griffin & Townsley, 2021; Guskey, 2009). Although behaviors such as work habits or citizenship are important indicators of student development, they should be documented independently to avoid confounding academic achievement (Link & Guskey, 2022).

Mixing non-academic elements with academic performance can yield grades that send misleading signals about what a student actually knows. In one study, homework and employability points in high school math grades resulted in significantly different outcomes compared to assessment scores alone (Griffin & Townsley, 2021). For example, a student with strong content knowledge might receive a low final grade due to missing assignments, while another student with modest understanding could receive a higher grade due to effort or extra credit (Guskey, 2009; Reeves, 2008). Such examples undermine the core purpose of grades as accurate communication of academic proficiency.

Distortions in Grade Meaning: Inflation and Deflation

These inconsistencies have raised growing concern about grade inflation and deflation. Grade inflation is defined as a rise in students' grades without a corresponding increase in achievement (Yeritsyan et al., 2022). Numerous studies have documented rising high school GPAs despite stagnant or declining standardized test scores (Camara et al., 2003; Godfrey, 2011; Woodruff & Ziomek, 2004). For example, one state analysis found that students with relatively low SAT scores were

nonetheless earning higher high school GPAs than in the past, suggesting inflated grades (Godfrey, 2011). Grade deflation can occur in the opposite scenario, when grades underrepresent students' true academic achievement (Brookhart et al., 2016; Guskey, 2009; Pattison et al., 2013). Some researchers note that focusing only on rising average GPAs can mask considerable variability, including cases where certain students receive lower grades than their level of proficiency (Pattison et al., 2013).

Both inflation and deflation highlight a disconnect between report card grades and demonstrated proficiency. For instance, a study in North Carolina found that Algebra I end-of-course exam scores were more predictive of ACT math performance than teacher-assigned course grades, implying that the standardized exam provided a more accurate gauge of content mastery than the teacher-assigned grade (Gershenson, 2018). Similarly, in a large urban high school, 43% of students' semester math grades differed by 5 or more percentage points from their summative averages, and 13% of students experienced inflation or deflation equivalent to an entire letter grade (Griffin & Townsley, 2021). This suggests a substantial number of students receive grades that do not accurately reflect their academic proficiency.

Point Accumulation Grading System and Standard-Based Grading Alternatives

These inconsistencies and misalignments highlight a clear problem: if final grades do not correspond to demonstrated learning, educational stakeholders may be misled about student performance and achievement. In response to these concerns, scholars and practitioners have increasingly turned to standards-based grading (SBG) and Evidence-Based Grading (EBG), both of which aim to improve grading accuracy and consistency by fundamentally restructuring how grades are constructed (Knight & Cooper, 2019; Muñoz & Guskey, 2015). In point accumulation systems, final grades are typically derived by aggregating points or scores across heterogeneous tasks (e.g., tests, homework, extra credit, participation), which combines academic achievement with non-academic factors. In contrast, SBG and EBG organize grades around clearly defined learning standards and rely on multiple sources of academic evidence to determine proficiency levels, with the goal of making grades more accurate reflections of what students know and can do (Reibel et al., 2025; Townsley, 2018). SBG, EBG, and similar reform models advocate for separating academic achievement from non-academic factors in grade reporting. For instance, teachers report work habits and behavior separately from academic performance so that the course grade itself represents only mastery or proficiency of the subject's standards (Brookhart et al., 2020; Link & Guskey, 2022; O'Connor, 2009). By doing so, schools hope to minimize the grade discrepancies caused by homework, extra credit, or behavior points and thereby improve the integrity and equity of grades; the final grade is intended to communicate what students know and can do in relation to specific standards, rather than the extent to which they complied with classroom expectations (Link & Guskey, 2022; O'Connor, 2009). Because summative assessments are designed to evaluate mastery of standards, they function as the primary evidence of proficiency within SBG and EBG frameworks. Comparing final course grades to summative assessment performance provides a lens for evaluating the extent to which points-based grading systems align, or fail to align, with evidence-based conceptions of grading accuracy and equity.

Equity and Bias in Grading Outcomes

Concerns about equity have come to the forefront of grading discussions. Traditional grading practices that include subjective components can inadvertently introduce bias, potentially disadvantageous to students from historically marginalized cultural or socioeconomic backgrounds (Feldman, 2018). When teachers factor in behavior or compliance that may be influenced by cultural norms or teacher expectations, grades can reflect implicit biases rather than purely academic performance. Assessment experts have concluded that integrating non-academic factors into grades mathematically introduces additional subjectivity and potential bias (Guskey, 1994; Ornstein, 1994). Ensuring that grades are an accurate and bias-free representation of learning is thus not only a matter of better measurement, but also a matter of justice, so that students from all backgrounds have equal opportunities to succeed based on their actual learning. The following section addresses the methodology used to address the research question guiding this study.

METHODOLOGY

This study employed a non-experimental quantitative research design to examine grade inflation and deflation across three public high schools. This approach was selected because it enabled the objective analysis of existing numerical data without manipulating variables or introducing interventions. A non-experimental design is appropriate for studies that seek to explore relationships or patterns using existing data without manipulation (Creswell & Creswell, 2018). The study focused

on identifying discrepancies between students' average summative assessment scores and final course grades, both percentages and letter grades, to detect patterns of grade inflation or deflation within real-world classroom grading systems.

The use of digital gradebook data allowed for a direct and systematic comparison between students' final course grades and underlying measures of summative performance, non-cognitive behaviors, and formative assignments such as homework. Participants were drawn from three high schools within a single public school district in the Midwestern United States. This district was purposefully selected due to its diverse student population, its ongoing transition to evidence-based grading practices at the secondary level, and its expressed commitment to improving grading consistency. Additionally, the district's centralized access to electronic gradebook systems and its active engagement in broader grading reform efforts made it a strategically relevant site for studying patterns of grade inflation and deflation. District and building-level approval was obtained in accordance with the district's research policy. To protect privacy, strict confidentiality protocols were followed throughout the study: no individual student or teacher could be identified, and all findings are reported in aggregate. The research protocol was reviewed and approved by the appropriate authorities, and all procedures adhered to ethical guidelines for the use of student data, including compliance with FERPA standards for de-identification and data security.

The final sample included 48 high school teachers (16 per school) and their students. At each school, four teachers were randomly selected from four core subject areas: English, mathematics, science, and social studies. This approach ensured disciplinary representation and reduced selection bias. These teachers' classes yielded a combined student sample of 2,558, though 64 students were excluded due to incomplete summative data, resulting in a final student sample of $N = 2{,}494$. Students with incomplete summative data were excluded from the analysis. Students with incomplete summative data ($n = 64$) were excluded because partial summative records (e.g., only one early summative score with later assessments blank) could produce misleading summative averages and biased paired comparisons with final course grades. Many of these 64 students transferred into the district late in the marking period. This sample size is consistent with comparable peer-reviewed studies on grading practices and is considered sufficient for generating statistically significant findings (Griffin, 2020; Griffin & Townsley, 2021; Pollio & Hochbein, 2015).

Data Collection

Gradebook data were collected for each selected teacher following institutional and district-level approvals. After receiving permission from the school district, the research team obtained digital copies of the teachers' gradebooks from the district's electronic grading system. To protect privacy, all student and teacher identifiers were anonymized prior to data access. This included the removal of names, student ID numbers, and teacher identifiers, which were replaced by randomized numerical codes.

Each gradebook contained detailed records of student performance, including individual assignment scores, grading category weights, and final course grades. From these records, key data points were extracted and organized for analysis: each student's final letter grade, final overall percentage, average summative assessment scores, and non-cognitive behaviors (e.g., homework and non-academic factors). A summative average was calculated for each student by averaging the scores of all summative assignment scores in that course.

Similar to the work of McMillan and Hellsten (2010), summative assessments in this study are defined as major evaluations of learning because they are intended to assess learning against course standards and learning objectives. In each teacher's gradebook, these assessments were identified as unit tests, projects, or other assessments designated as "summative." To identify these summative assessments consistently across gradebooks, a standardized extraction protocol was used based on the teacher's gradebook category labels (e.g., "Summative," "Tests," "Projects," "Exams") and assignment labels within the district's electronic grading system. Extracted data were reviewed and cross-checked for coding consistency and calculation accuracy prior to analysis. Accordingly, the analysis reflects summative designations as recorded in teachers' operational gradebook practices. See Table 1 for the example of summative assessment averages.

Once the data were cleaned and organized, the next step involved accounting for variations in grading systems across participating teachers. Some teachers used a total points system (unweighted), while others used weighted categories, such as a 70/30 or 60/40 split between summative assessments and non-cognitive behaviors. To ensure comparability across systems, summative scores were isolated from non-academic components in every case. For teachers using weighted categories, summative assessment percentages were taken directly from the gradebook's summative category calculation. For teachers using total points, summative assignments were identified by their labels or gradebook categorization (e.g., "Exam" or "Test"), and the student's summative percentage was computed as points earned out of points possible on those summative assessments.

Each student's final course percentage and final letter grade were recorded alongside their summative average percentage and its corresponding letter grade equivalent. All data were cross-checked for accuracy (e.g., ensuring letter grades correspond to the reported percentage range). The complete dataset was then entered into SPSS (Version 28) for statistical analysis.

Table 1

Example of summative assessment averages from the point accumulation model

Assessment	Student 1	Student 2
Homework (%)	100	65
Participation (%)	95	60
Homework (%)	100	70
Formative Quiz (%)	92	75
Summative #1 (%)	78	92
Homework (%)	100	80
Summative #2 (%)	72	86
Participation (%)	90	50
Summative #3 (%)	64	94
Summative Avg (%)	71.33	90.66
Final Course Grade	B	C

Note. Student 1's higher final course grade (B), despite a lower summative average, reflects the influence of non-cognitive factors (e.g., formative assessments, homework, participation) in a point-accumulation grading model.

Statistical Analysis

To address the research questions, we conducted a series of quantitative analyses comparing students' summative performance to their final grades. For each student, the summative assessment average (expressed as a percentage) was directly compared to the final course percentage. This enabled the calculation of the difference, which effectively measured grade inflation or deflation. A positive difference (i.e., a final grade higher than the summative average) was interpreted as grade inflation, whereas a negative difference was interpreted as grade deflation.

Next, to examine potential discrepancies at the categorical level, we converted both the summative assessment average and the final letter grade into a standardized 4.0 scale (as a categorical value, A = 4, B = 3, C = 2, D = 1, F = 0). This transformation allowed for direct comparison of letter grades in numeric terms (A= 4, B = 3, C = 2, D = 1, F = 0), enabling us to detect shifts equivalent to full or partial letter grade changes. The district used a common letter-grade cut-score scheme across courses (e.g., 90–100 = A, 80–89 = B, 70–79 = C, 60–69 = D, below 60 = F), so formal syllabus review was not required to establish the cut-score framework used for the analytic letter-grade conversion.

We conducted two primary statistical tests using IBM SPSS Statistics (Version 28), as these tests can detect significant differences between two related sets of data: students' summative assessment average and their final course grades. The first test compared the numeric differences between final letter grade percentage and the summative assessment percentage to determine whether grades were inflated or deflated. The second test compared the summative and final letter grades after converting them to a standard numerical scale (A = 4, B = 3, C = 2, D = 1, F = 0) to examine whether similar patterns emerged when grades were treated categorically. Both these analyses allowed for patterns to be identified in grading consistency and accuracy between assessed performance and reported grades.

Paired-Samples t-Test on Percentage Grades

The study performed a paired-samples *t*-test comparing the mean of students' final percentage grades to the mean of all the students' summative assessment percentage scores (Griffin & Townsley, 2021; Pollio & Hochbein, 2015). Each student's summative average was calculated by averaging all summative assessments recorded in the course gradebook. Each student contributed a pair of values: final letter grade percentage and summative average percentage. This test determines whether, on average, final letter grades differed significantly from what purely summative scores would predict.

The paired *t*-test was used because the research question focuses on detecting inflation and deflation between two related measures for the same individuals. The analysis focused on within-student differences between two paired measures (summative average and final course grade); therefore, multilevel modeling was not used. This approach has been used in prior grading research to evaluate alignment between assessment-based performance and final grades (Griffin & Townsley, 2021). The analysis was two-tailed and used $\alpha = .05$, a conventional threshold for significance that allows for the detection of both grade inflation and deflation, rather than assuming a directional hypothesis.

Paired-Samples *t*-Test on Letter Grades

To address the research questions, we conducted a paired *t*-test on categorical letter grades to examine discrepancies in final letter grades across grade categories. Each student's final letter grade and the letter grade corresponding to their summative average were converted into numeric values on a standard 4.0 scale (A = 4, B = 3, C = 2, D = 1, F = 0). For example, if a student's summative average was 78%, it was treated as a "C" (2.0) for comparison to the final letter grade ("B" = 3.0). This approach follows Griffin and Townsley (2021) and Pollio and Hochbein (2015), who similarly used numeric conversions of letter grades for analysis. Using these numeric equivalents, a paired *t*-test compared the mean final letter-grade points with the mean summative-assessment points for the sample. This test provides insight into whether students tended to receive higher or lower letter grades than their level of proficiency would recommend.

While letter grades are ordinal, this numerical translation allowed us to apply parametric analysis to gauge average differences in letter grade (with 1.0 indicating a one-letter grade difference or with 2.0 indicating a two-letter grade difference). Both tests were run on the full sample of 2,494 students, yielding corresponding *p*-values and effect size estimates. This is significant because it quantifies whether students are receiving grades that accurately reflect their demonstrated academic proficiency based on summative assessments. Understanding these differences is essential for evaluating the integrity of grading practices and identifying systemic inflation or deflation patterns across schools or teachers.

To quantify the extent of grade inflation and deflation in practical terms, we also calculated the frequency of discrepancies beyond certain thresholds. Following the approach of Griffin and Townsley (2021), we identified each student who had a notable gap between their summative performance and final grade. Two percentage-based thresholds were used: 5% and 10%. If a student's final percentage exceeded their summative average by 5% or more, we classified that as a case of grade inflation; if it fell 5% or more, it was classified as grade deflation. A difference of five percent corresponds roughly to half a letter grade in typical grading scales (Guskey, 2013; Schinske & Tanner, 2017; Schneider & Hutt, 2014). More stringently, a 10% or greater difference was recorded as a case of major inflation or deflation, approximately equivalent to a full letter grade discrepancy. For example, a student with an 80% summative assessment average who received a 90% final grade would be counted in the "≥10% inflation" group, while one with 80% summative but a 70% final would be "≥10% deflation."

We tallied the number and percentage of students falling into each category of inflation or deflation. Similarly, we examined categorical grade shifts. We noted how many students' final letter grades were at least one or two full letters higher or lower than the letter grade indicated by their summative score. All metrics (5%/10% thresholds and 1-letter/2-letter discrepancies) were computed to supplement the *t*-test results, providing a clear picture of the prevalence of grade inflation/deflation in the sample.

Building on the foundational work of Cohen (1988) on effect size interpretation, effect sizes were calculated to assess the magnitude of any observed differences. Cohen's *d* was computed for each paired comparison (percentage and letter grade) by dividing the mean difference by the standard deviation of the difference scores (Borenstein et al., 2009; Cohen, 1988; Lakens, 2013). This provides a standardized measure of how large the grade inflation/deflation effect was. We interpreted *d* using conventional benchmarks: ~0.20 was considered a small effect, ~0.50 a medium effect, and ~0.80 or above a large effect (Cohen, 1988; Cohen, 1992; Lakens, 2013; Sullivan & Feinn, 2012). All analyses were conducted at the 95% confidence level, and results are presented with appropriate statistical notations (e.g., mean differences, *t* values, degrees of freedom, *p* values, and confidence intervals) (American Psychological Association, 2020; Wasserstein & Lazar, 2016; Wilkinson et al., 1999). This analytical strategy directly aligns with the research questions, determining whether final grades differ significantly from summative assessment averages and quantifying how many students are affected by such differences and to what extent.

RESULTS

To address the research question, *"To what extent do final grades, measured both as percentages and as letter grades, differ from summative assessment scores, indicating grade inflation or deflation?*" two complementary analyses were conducted, as shown in Table 2. First, a paired-samples *t*-test was used to compare students' final course percentages to their average summative assessment percentages. Second, a paired-samples *t*-test was performed on converted letter grades, using a 4.0 GPA scale (A = 4, B = 3, etc.), to examine discrepancies between categorical summative grades and final reported grades. These analyses allowed for both numerical and categorical evaluation of grading discrepancies, offering an in-depth understanding of grade inflation or deflation at both the system and student levels. This dual approach provided a more comprehensive understanding of the extent to which grading misalignment was present, or absent, within the district's current practices.

Final Grade Percentage vs. Summative Assessment Percentage Grades

The percentage-based analysis compared students' final overall grade percentages to their summative assessment grade percentages. The analysis revealed significant differences ($t_{2494} = 14.71$, $p < .001$) when comparing the means between students' final grade percentages ($M = 83.15$, $SD = 13.16$) and their summative assessment percentage grades ($M = 81.40$, $SD = 14.34$), suggesting an average inflation of 1.75 percentage points. The 95% confidence interval for the mean difference [1.52, 1.99] indicated consistency with grade inflation. This finding is important because it demonstrates that final course grades, which are the most visible on transcripts and used for GPA calculations, are, on average, modestly inflated compared to students' summative assessment performance. This suggests that non-cognitive behaviors may be elevating reported achievement beyond what assessments alone indicate, prompting deeper reflection on whether current grading practices genuinely represent student proficiency or merely reward compliance and conduct, thereby distorting our understanding of academic achievement.

To further examine the alignment between final grades and summative assessment performance, a Pearson correlation coefficient was calculated. This tool was chosen because it measures the strength and direction of the linear relationship between two continuous variables, providing a clear indication of how closely students' overall course outcomes align with their performance on key assessments. The analysis revealed a very strong positive correlation between final grade percentages and summative assessment percentages ($r = .91$, $p < .001$) (Goodwin & Leech, 2006). This indicates that students who performed well on summative assessments generally received correspondingly high final grades, although those final grades were, on average, slightly inflated. The effect size for the mean difference was also calculated using Cohen's *d*. Cohen's *d* was selected because it provides a standardized measure of the magnitude of the difference between two groups, allowing for easier interpretation and comparison across studies. The resulting value was 0.295, indicating a small to moderate effect.

Final Letter Grades vs. Summative Assessment Letter Grades

In addition to understanding the inflation and deflation percentages, the researchers also analyzed letter grades. A similar pattern of inflation emerged, with differences that are more pronounced in practical terms, depending on how letter grades are configured and the thresholds that are predetermined to generate an A, B, C, D, or F. Furthermore, transcripts report final letter grades, which may or may not include final percentage grades. Each student's final letter grade (A = 4, B = 3, C = 2, D = 1, F = 0) was compared to the letter grade equivalent of their summative assessment average. There were statistically significant ($t_{2494} = 17.35$, $p < .001$) differences when comparing the means between students' final letter grades ($M = 2.96$, approximately a B) and their summative assessment letter grades ($M = 2.74$, approximately a C), suggesting an average inflation of 0.22 letter grades. The average difference of +0.22 on a letter grade scale indicates that, on average, final grades were about a quarter of a letter grade higher than what summative assessments alone would justify. This is significant because it highlights how even a fraction of a letter grade in categorical grade inflation can have meaningful implications for student transcripts, college admissions, and scholarship eligibility. A discrepancy of 0.22 on a 4.0 GPA scale, while seemingly minor, may alter a student's reported academic standing and influence external perceptions of their achievement.

To measure the effect size, Cohen's *d* was utilized again. The effect size for this difference was $d = 0.348$, indicating a moderate effect that was larger than the effect observed in the percentage-based analysis ($d = 0.295$) (Cohen, 1988). There was also a strong correlation between final letter grades and summative letter grades ($r = .855$, $p < .001$), suggesting that students who performed well on summative assessments generally received correspondingly higher letter grades (Goodwin & Leech, 2006). This correlation was somewhat weaker than the correlation observed with percentage grades, which was

expected since converting percentages into letter grades introduces categorical thresholds (A = 4, B = 3, C = 2, D = 1, F = 0). This means that even small percentage-based discrepancies can lead to meaningful shifts in categorical outcomes, reinforcing the impact of grade inflation. See Table 2 for the summary of descriptive and inferential statistics.

Prevalence of Grade Inflation and Deflation

While the average discrepancy between final grade and summative assessments was modest, a significant proportion of students experienced substantial individual-level differences, highlighting instances of grading inconsistency. For example, approximately one-third (30.83%), or 769 students, demonstrated statistical discrepancies, with their final grades differing by at least 5 percentage points from the summative assessment averages. Specifically, grade inflation was more prevalent, affecting 21.97% of students, or 548 students, compared to grade deflation, which affected 8.86% of students, or 221 students. The reasoning for a higher discrepancy with grade inflation could be attributed to non-cognitive factors often used in point accumulation models that become perceived reasons for student success (Wormeli, 2017). These include points awarded for peripheral components like compliance, participation, effort, extra credit, etc. These factors can artificially elevate overall grades independent of true academic proficiency. This suggests that a portion of students had final grades that were significantly higher or lower than the summative assessment averages, beyond typical measurement variability.

Table 2

Summary of Descriptive and Inferential Statistics for Paired-Samples Comparisons of Final Grades and Summative Assessment Scores

Comparison	Final Grade Mean	Summative Assessment Mean	Final Grade *SD*	Summative Assessment *SD*	Mean Difference	*t*	*p*	Cohen's *d*	95% Confidence Level
Final Grade Percentage vs. summative percentage grades	83.15	81.40	13.17	14.34	1.76	14.73	<.001	0.295	[1.52, 1.99]
Final letter grades vs. summative letter grades	2.96	2.74	1.14	1.23	.22	17.36	<.001	0.348	[0.20, 0.25]

Further analysis revealed that 10.79%, or 269 students, had even larger discrepancies, with 7.66%, or 191 students, experiencing an inflation rate of 10% or more, and 3.13%, or 78 students, experiencing a deflation rate of the same magnitude. This results in approximately one in nine students receiving a final grade that differs from their summative assessments by at least 10 percentage points.

These discrepancies equate to half- or full-letter grade shifts relative to students' summative assessment percentages. For instance, a student achieving a summative assessment average of 77% could potentially receive a final grade percentage as high as 87% or as low as 67% when additional non-cognitive behaviors are considered. This highlights the significant practical implications for accurately reporting and interpreting student academic grades. See Table 3 for the frequency and percentages of grade inflation and deflation. Then see Figure 1 for the grade discrepancies between final grade percentages and summative assessment percentages, and Figure 2 for the distribution of discrepancy scores across the full sample.

Letter Grade Discrepancies

When letter grade discrepancies were analyzed categorically, the prevalence of grading inconsistencies was further highlighted. A total of 36.25%, or 904 students, received a final letter grade, differing by at least one full letter from their summative assessment letter grades. Grade inflation remained significantly more common, affecting 28.31% of students (706 students), compared to grade deflation, which impacted 7.94% of students (98 students). This indicates that nearly one-third of students ended the semester with higher letter grades than their summative performance suggested.

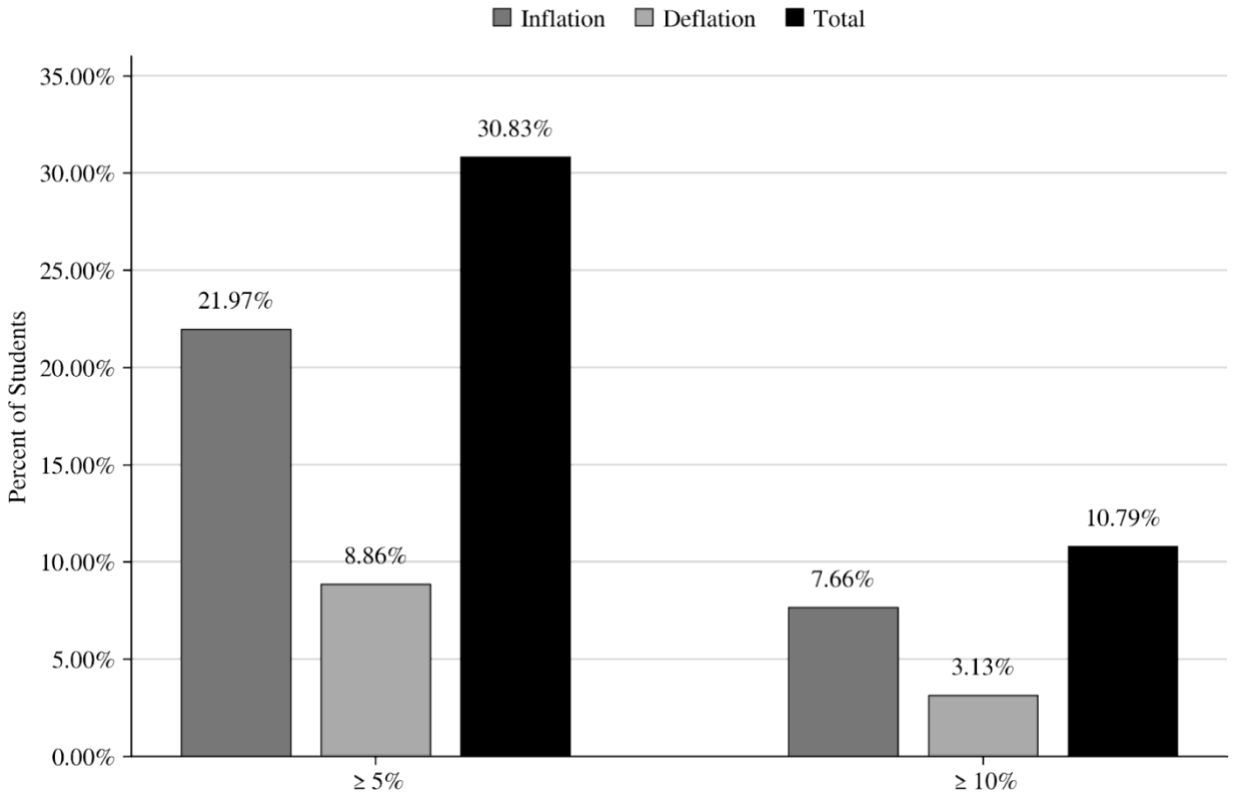
Table 3

Frequency and Percentage of Grade Inflation and Deflation

Metric	Inflation	Deflation	Total
≥ 5% Percentage Difference	21.97% (*n* = 548)	8.86% (*n* = 221)	30.83% (*n* = 769)
≥ 10% Percentage Difference	7.66% (*n* = 191)	3.13% (*n* = 78)	10.79% (*n* = 269)
≥ 1 Letter Grade Difference	28.31% (*n* = 706)	7.94% (*n* = 198)	36.25% (*n* = 904)
≥ 2 Letter Grade Difference	2.25% (*n* = 56)	0.56% (*n* = 14)	2.81% (*n* = 70)

Figure 1

Grade Discrepancies Between Final Grade Percentages and Summative Assessment Percentages

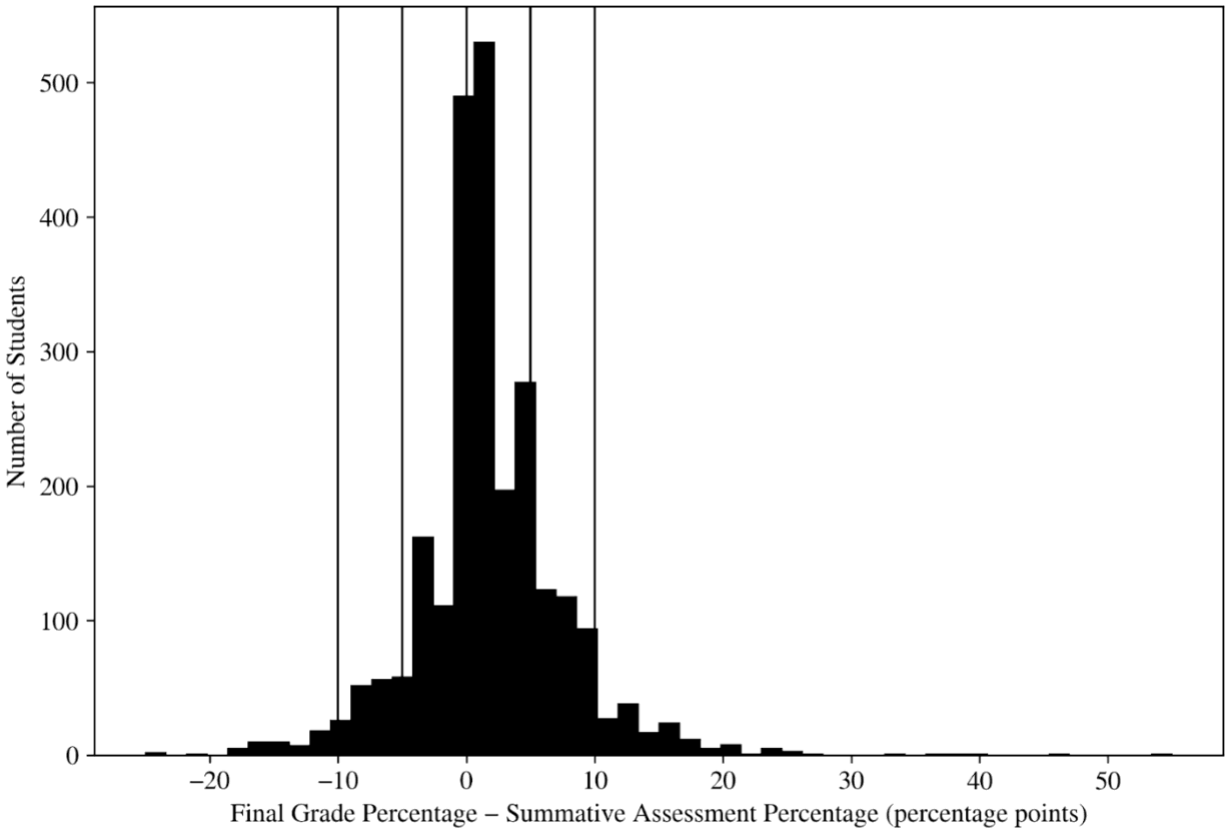


Note. Bars display percentages. *N* = 2,494. ≥ 5% inflation *n* = 548, ≥ 5% deflation *n* = 221, total *n* = 769; ≥ 10% inflation *n* = 191, ≥ 10% deflation *n* = 78, total *n* = 269.

Most discrepancies involved a one-letter grade difference; however, a smaller subset of students experienced even larger variations. Specifically, 2.81%, or 70 students, had a discrepancy of two or more letter grades. Within this group, 2.25%, or 56 students, had their final grades inflated by two letter grades. This means that a student's grade could be inflated from a C to an A. Comparatively, only 0.56%, or 14 students, experienced deflation by two letter grades, such as from a B to a D. These categorical discrepancies carry meaningful academic consequences. A one or two-letter increase could enhance a student's academic standing in transcripts and GPA, while even minor deflation may adversely affect confidence, motivation, scholarship, and other funding sources and perceptions of equity (Brookhart, 2011; Feldman, 2018; Guskey, 2020; Reeves, 2008). The disproportionate ratio of grade inflation to deflation (approximately 3.6 to 1) highlights systemic patterns and reinforces the need for greater consistency and transparency in grading practices. See Figure 3 for the grade discrepancies between final letter grades and summative assessment letter grades.

Figure 2

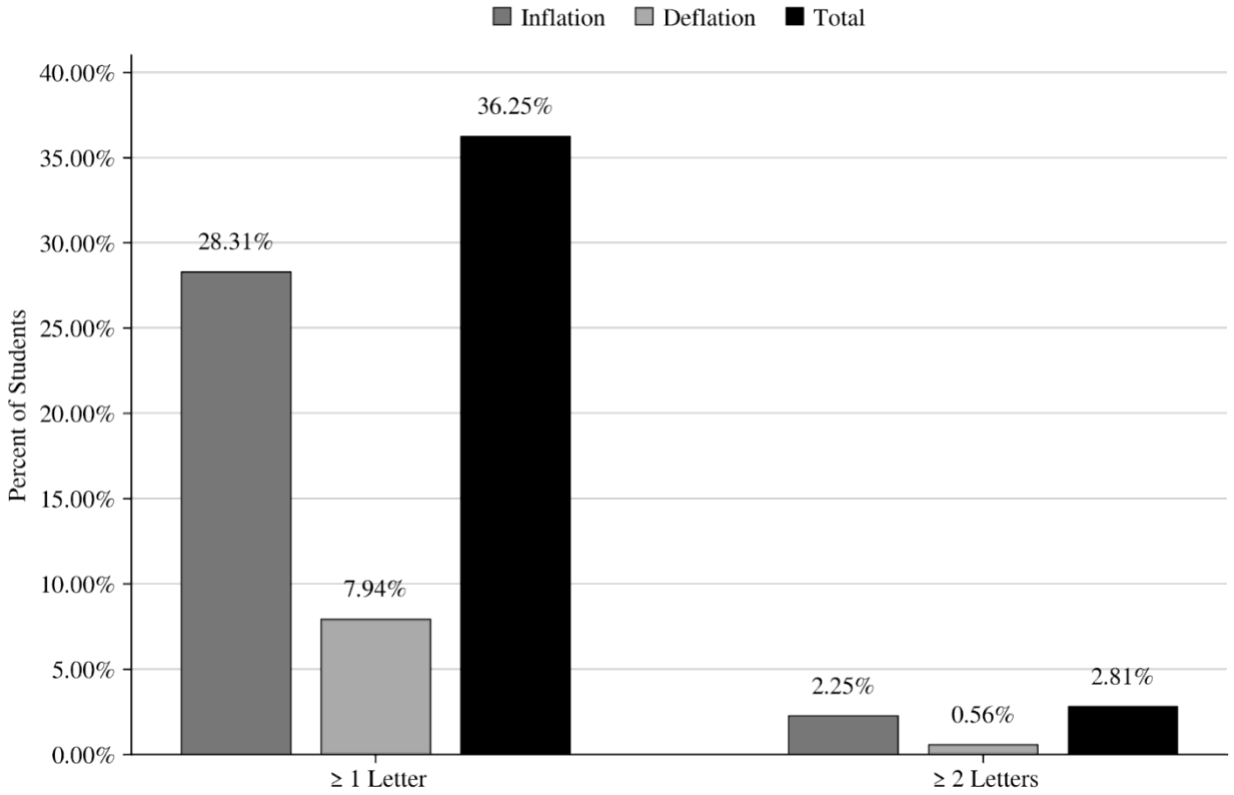
Distribution of Discrepancy Scores Across the Full Sample



Note. Discrepancy scores were calculated as final grade percentage minus summative assessment percentage (in percentage points). Positive values indicate that final grade percentages were higher than summative assessment percentages, and negative values indicate that final grade percentages were lower than summative assessment percentages. Vertical reference lines indicate −10, −5, 0, +5, and +10 percentage-point discrepancies.

Figure 3

Grade Discrepancies Between Final Letter Grades and Summative Assessment Letter Grades



Note. Bars display percentages. N = 2,494. ≥ 1 letter inflation n = 706, ≥ 1 letter deflation n = 198, total n = 904; ≥ 2 letters inflation n = 56, ≥ 2 letters deflation n = 14, total n = 70.

DISCUSSION AND IMPLICATIONS

To what extent do final grades, measured both as percentages and as letter grades, differ from summative assessment scores, indicating grade inflation or deflation? Using the analytic sample of N = 2,494, we observed an overall association between final grades and summative assessment grades but frequent and asymmetric discrepancies. At the ≥ 5% threshold, inflation occurred for 21.97% of final grades (n = 548; 95% CI [20.35, 23.60]) and deflation for 8.86% of final grades (n = 221; 95% CI [7.75, 9.98]); the combined inflation and deflation rate was 30.83% (n = 769; 95% CI [29.02, 32.65]). At ≥ 10%, inflation was 7.66% (n = 191; 95% CI [6.61, 8.70]) and deflation 3.13% (n = 78; 95% CI [2.44, 3.81]); the total was 10.79% (n = 269; 95% CI [9.57, 12.00]). For letter thresholds, at ≥ 1 of a letter grade, inflation was 28.31% (n = 706; 95% CI [26.54, 30.08]) and deflation 7.94% (n = 198; 95% CI [6.88, 9.00]); the total was 36.25% (n = 904; 95% CI [34.36, 38.13]). At ≥ 2 letter grades, inflation was 2.25% (n = 56; 95% CI [1.66, 2.83]) and deflation 0.56% (n = 14; 95% CI [0.27, 0.85]); the total was 2.81% (n = 70; 95% CI [2.16, 3.45]). This study revealed a statistically significant and practically meaningful pattern of grade inflation and deflation, confirming concerns raised in prior research about the accuracy and consistency of traditional grading practices (Brookhart et al., 2016; Griffin & Townsley, 2021; Guskey & Link, 2018).

Grade inflation was more common than deflation, consistent across both percentage and letter grade comparisons. These findings directly address the research question: final course grades often do not accurately reflect summative test results when non-academic factors are taken into account, with a substantial minority of students experiencing discrepancies in favor of higher grades. These discrepancies have essential implications for the meaning and fairness of grades. Grades are a primary currency in education, influencing high-stakes decisions such as GPA, honors, scholarships, graduation, and college admissions (Brookhart et al., 2016; Guskey & Link, 2018; Nowruzi, 2021). If final grades systematically overestimate achievement, they can provide stakeholders with an inaccurate understanding of students' true proficiency and mastery of learning. Our findings reinforce longstanding concerns in the literature that traditional grading can be inconsistent

and incorporates non-cognitive behaviors beyond academic mastery (Brookhart et al., 2016; Duncan & Noonan, 2007; Guskey, 2011; Guskey & Link, 2018; McMillan, 2001; Nowruzi, 2021; Nowruzi & Amerian, 2020).

In a century-spanning review, Brookhart et al. (2016) concluded that teacher-assigned grades are often subjective and not reliably aligned with standardized measures of learning. This study's evidence of grade inflation and deflation aligns with that conclusion, suggesting that many teachers award course grades based on a mix of achievement and non-cognitive behaviors, resulting in grades that do not strictly reflect proficiency of summative assessment performance that assesses state or national standards (Brookhart, 2011; Guskey, 2002).

Guskey and Link (2018) similarly argued that large inconsistencies in grading practices, such as what counts towards a grade and how different teachers weigh components, are a fundamental obstacle to fairness in grading. Our results support this, as the variation of grade versus summative assessment alignment indicates a lack of a uniform standard for what a course grade represents.

The prevalence of grade inflation and deflation observed here is consistent with recent empirical studies. For example, in a single-district study of high school math, Griffin and Townsley (2021) found that when homework and "employability" points were included in grades, 61.5% of students ended up with higher final grades than their test scores alone would justify, whereas about 38% saw grade deflation. They reported that 43% of students had grade differences of at least five percentage points, and 12.6% differed by a full letter grade when compared to a summative-only grade calculation. This study, which examined a broader set of courses that included all core area courses (English, math, science, and social studies), found roughly 31% with ≥ 5% discrepancies and ~36% with ≥ 1-letter discrepancies, with inflation more prevalent, a pattern similar to Griffin and Townsley's (2021) findings. This convergence suggests that the tendency for course grades to be higher than summative assessment performance is widespread and not confined to a single subject area or school. Grade inflation can mean that some students receive passing or even better course grades despite not being proficient in key content and state standards.

These outcomes exemplify why our findings matter: They highlight how current grading practices may inflate student records, potentially giving a distorted sense of student achievement or masking situations where a student is struggling with the content. To illustrate the stakes under a typical grading scale (e.g., 90/80/70), a student averaging 79% on summative assessments but receiving an 89% final (≥ 10% inflation) records a B+ rather than evidence consistent with a C, directly affecting GPA thresholds, honors lists, and scholarship eligibility. Our study strengthens the case made by prior researchers that the point-accumulation model often produces a "hodgepodge" of academic and nonacademic factors, distorting communication of what students have learned and what they are proficient in (Brookhart et al., 2016; Duncan & Noonan, 2007; Guskey, 2011; Guskey & Link, 2018; McMillan, 2001; Nowruzi, 2021; Nowruzi & Amerian, 2020). Because many gradebooks include achievement alongside non-academic factors, frequent total discrepancies of ≥ 5% and ≥ 1 letter appear even when correlations with summative evidence are high.

Taken together, these findings warrant caution in interpreting final course grades as precise measures of proficiency. Therefore, we outline key limitations and future directions.

LIMITATIONS AND FUTURE RESEARCH

While the findings of this study offer meaningful insights, it is important to acknowledge several limitations that may affect the interpretation and generalizability of the results. First, this research was conducted using gradebook data from a single school district. Because the school district's population, curricular standards, and grading culture may differ from those in other contexts, this may limit the generalizability of the findings. Beyond this, final course grades were collected from teacher-managed electronic gradebooks, introducing potential fidelity issues. We assumed that gradebook records were accurate and consistently coded; however, there may be inconsistencies in teacher practices when using gradebook software, such as data entry errors.

Although the research team used a standardized extraction protocol and cross-checking process to identify summative assessments from teacher-assigned gradebook category and assignment labels, the study examined classifications as they were operationalized in authentic classroom gradebook practice, rather than applying an external validation or reclassification process. While we acknowledge the value of examining gradebook classifications as practiced authentically in classrooms, this approach risks conflating teacher labeling inconsistencies with true assessment characteristics, potentially inflating interrater variability beyond what standardized protocols would reveal (Brimi, 2011). To strengthen validity, future studies could incorporate a brief expert review of a subsample using established reliability benchmarks. We also did not control for differences in teachers' grading scales or weighting policies beyond the summative assessments; thus, some portion of the observed discrepancies could stem from technical or recording inconsistencies rather than

intentional grading policies. Although course syllabi were not formally reviewed, the gradebook records analyzed captured teacher-reported final grade outcomes, including the effects of course-level grading practices visible in the gradebook (e.g., added points or extra credit).

It is also important to note that we used students' summative assessment averages as a benchmark for academic proficiency. This is a proxy, and it may not perfectly capture all facets of student learning or state/national standards. If a summative assessment was of uneven quality, narrow in scope, or misaligned with standards, a discrepancy between the course grade and the summative assessment may not truly indicate grade inflation. In other words, some "inflation" might reflect limitations of the assessments themselves (e.g., a student demonstrated a standard in the homework that the summative assessment did not measure). The study's design cannot disentangle true grade inflation from misalignment caused by misaligned summative assessments.

Similarly, we were unable to account for all factors that might influence grade discrepancies in a point-based grading system. Variation in grading policy among teachers could vary and not be accounted for in our data. In addition, student-level factors such as motivation or access to support might systematically affect both a student's class participation and their summative assessment performance. Without data on these variables, our analysis cannot determine which specific factors contribute most to the inconsistencies. This is a methodological limitation that leaves some ambiguity about the causes of variation in grade inflation and deflation detected in the findings.

These conclusions highlight several areas where additional inquiry is needed, prompting the following recommendations for future research. First, it remains unclear whether grading inconsistencies affect all students and subjects uniformly. Because the data spans multiple content areas analyzed in aggregate (English, math, science, and social studies), subject-specific patterns were not fully examined. Future research could investigate how grade discrepancies differ across subject areas or impact specific student groups. Additionally, employing a qualitative lens to this topic may provide a more nuanced and complex picture of the cause of grade inflation or deflation. For example, interview or observational studies might explore teacher philosophies and rationales that justify grades that diverge from summative assessment scores (e.g., What rationales do teachers give for including non-academic factors into final course grades?)

This study confirms that final course grades in point-accumulation models frequently diverge from students' demonstrated summative assessment performance, with grade inflation occurring more regularly than grade deflation. These discrepancies raise important concerns about the accuracy, fairness, and equity of point-based grading practices in K–12 education. For educational leaders, these findings suggest several actionable steps to support students, educators, and families in this work.

Educational leaders can begin by clarifying the purpose of grades and establishing shared expectations that course grades should primarily communicate academic proficiency. Leaders may also support greater grading consistency by adopting common definitions for summative and formative assessments, providing professional learning focused on separating academic achievement from behavioral compliance, and auditing gradebook practices to examine alignment between summative evidence and final grades. Leaders can promote equity by encouraging transparent grading criteria, limiting the influence of non-academic factors on grades, and using summative assessment data as a reference point when reviewing grading policies. These steps not only improve the technical accuracy of grades but also help ensure that grading practices do not unintentionally advantage or disadvantage particular groups of students. It is clear that the work behind grades needs attention (Griffin & Townsley, 2021). The findings of this study highlight the need for intentional, system-wide shifts towards evidence-based grading practices so grades can serve as valid, transparent, and equitable indicators of student learning for all stakeholders. This study can serve as a grounding framework to begin these conversations. Educators of all levels, students, and communities need to reimagine the purpose of grades, and that starts with hard conversations about the inequities already occurring.

REFERENCES

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 78*(5), 218–223. https://doi.org/10.3200/TCHS.78.5.218-223

Allensworth, E., & Clark, K. (2020). High school GPA: Predicting success in college. *Educational Researcher, 49*(6), 393–397. https://doi.org/10.3102/0013189X20902110

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

Brimi, H. M. (2011). Reliability of grading high school work in English. *Practical Assessment, Research & Evaluation, 16*(17). https://doi.org/10.7275/4z0x-7z

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice, 10*(1), 35–36. https://doi.org/10.1111/j.1745-3992.1991.tb00182.x

Brookhart, S. M. (2004). *Grading*. Pearson/Merrill/Prentice Hall.

Brookhart, S. M. (2011). Starting the conversation about grading. *Educational Leadership, 69*(3), 10–14.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Brookhart, S. M., McTighe, J., Wiliam, D., & Guskey, T. R. (2020, September 1). Eight essential principles for improving grading. *Educational Leadership, 78*(1). https://www.ascd.org/el/articles/eight-essential-principles-for-improving-grading

Camara, W. J., Kimmel, E. W., Scheuneman, J. D., & Sawtell, E. A. (2003). *Whither the SAT? Trends in college admissions and student performance*. College Board.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.

Duncan, R., & Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *The Alberta Journal of Educational Research*, *53*(1), 1–21. https://doi.org/10.11575/ajer.v53i1.55195

Feldman, J. (2018). *Grading for equity: What it is, why it matters, and how it can transform schools and classrooms*. Corwin Press.

Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., D'Mello, S. K., Finn, A. S., & Duckworth, A. L. (2019). Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *American Educational Research Journal, 56*(6), 2077–2115. https://doi.org/10.3102/0002831219843292

Gershenson, S. (2018). *Grade inflation in high schools (2005–2016)*. Thomas B. Fordham Institute.

Godfrey, K. E. (2011). *Investigating grade inflation and non-equivalence* (College Board Research Report No. 2011-2). College Board. https://eric.ed.gov/?id=ED521179

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education, 74*(3), 249-266. https://doi.org/10.3200/JEXE.74.3.249-266

Griffin, R. T. (2020). *Grading and equity: Inflation/deflation based on race, gender, socio-economic and disability statuses when homework and employability scores are included* (Doctoral dissertation, University of Northern Iowa). UNI ScholarWorks. https://scholarworks.uni.edu/etd/1045

Griffin, R., & Townsley, M. (2021). Points, points, and more points: High school grade inflation and deflation when homework and employability scores are included. *Journal of School Administration Research and Development, 6*(1), 1-11. https://doi.org/10.32674/jsard.v6i1.3460

Guskey, T.R. (1994). Making the grade: What benefits students? *Educational Leadership, 52*(2), 14-20.

Guskey, T. R. (2002). *Developing grading and reporting systems for student learning*. Corwin Press.

Guskey, T. R. (2009). *Practical solutions for serious problems in standards-based grading*. Corwin Press.

Guskey, T. R. (2011). *Five obstacles to grading reform*. *Educational Leadership, 69*(3), 16–21. https://www.ascd.org/el/articles/five-obstacles-to-grading-reform

Guskey, T. R. (2013). The case against percentage grades. *Educational Leadership, 71*(1), 68–72.

Guskey, T. R., & Link, L. J. (2018). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice, 26*(3), 303–320. https://doi.org/10.1080/0969594X.2018.1555515

Guskey, T. R. (2020). Breaking up the grade. *Educational Leadership, 78*(1), 40–46. https://www.ascd.org/el/articles/breaking-up-the-grade

Knight, M., & Cooper, B. (2019). Taking on a new grading system: The interconnected effects of standards-based grading on teaching, learning, assessment and student behavior. *NASSP Bulletin, 103*(1), 3–21. https://doi.org/10.1177/0192636519826709

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Link, L. J., & Guskey, T. R. (2022). Is standards-based grading effective? *Theory into Practice, 61(*4), 406–417. https://doi.org/10.1080/00405841.2022.2107338

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.). Prentice Hall.

Marzano, R. J., & Heflebower, T. (2011). Grades that show what students know. *Educational Leadership, 69*(3), 34–39.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20–32. https://doi.org/10.1111/j.1745-3992.2001.tb00055.x

McMillan, J. H., & Hellsten, L. (2010). *Classroom assessment: Principles and practice for effective standards-based instruction*. Pearson Education Canada.

McMillan, J. H. (2018). *Classroom assessment: Principles and practice for effective standards-based instruction*. Pearson.

Muñoz, M. A., & Guskey, T. R. (2015). Standards-based grading and reporting will improve education. *Phi Delta Kappan, 96*(7), 64–68. https://doi.org/10.1177/0031721715579043

Nowruzi, M., & Amerian, M. (2020). Exploring the factors Iranian EFL institute teachers consider in grading using personal construct theory. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills), 38*(4), 123–164. https://doi.org/10.22099/jtls.2020.36293.2780

Nowruzi, M. (2021). A study of EFL teachers' classroom grading practices in secondary schools and private institutes: A mixed methods approach. *Language Testing in Asia, 11*, Article 29. https://doi.org/10.1186/s40468-021-00145-2

O'Connor, K. (2009). *How to grade for learning, K–12* (3rd ed.). Corwin.

Ornstein, A. C. (1994). Grading practices and policies: An overview and some suggestions. *NASSP Bulletin, 78*(561), 55–64. https://doi.org/10.1177/019263659407856111

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher, 42*(5), 259–265. https://doi.org/10.3102/0013189X13481382

Pollio, M., & Hochbein, C. (2015). The association between standards-based grading and standardized test scores as an element of a high school reform model. *Teachers College Record, 117*(11), 1–28. https://doi.org/10.1177/016146811511701106

Reeves, D. B. (2008). Leading to change: Effective grading practices. *Educational Leadership, 65*(5), 85–87.

Reibel, A. R., Twadell, E., Onuscheck, M., & Gobble, T. (2025). *Pathways to proficiency: Implementing evidence-based grading* (2nd ed.). Solution Tree.

Schinske, J., & Tanner, K. (2017). Teaching more by grading less (or differently). *CBE—Life Sciences Education, 13*(2), 159–166. https://doi.org/10.1187/cbe.cbe-14-03-0054

Schneider, J., & Hutt, E. (2014). Making the grade: A history of the A–F marking scheme. *Journal of Curriculum Studies, 46*(2), 201–224. https://doi.org/10.1080/00220272.2013.790480

Starch, D., & Elliott, E. C. (1912). Reliability of grading high school work in English. *School Review, 20*(7), 442–457.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Townsley, M. (2018). Mastery-minded grading in secondary schools. *School Administrator, 75*(1), 16–19.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Woodruff, D. J., & Ziomek, R. L. (2004, March). *High school grade inflation from 1991 to 2003* (ACT Research Report Series 2004–04). ACT, Inc. https://files.eric.ed.gov/fulltext/ED484784.pdf

Wormeli, R. (2017). *Fair Isn't Always Equal: Assessment & Grading in the Differentiated Classroom* (2nd ed.). Routledge. https://doi.org/10.4324/9781032681122

Yeritsyan, A., Mjelde, J. W., & Litzenberg, K. K. (2022). Grade inflation or grade increase. *Journal of Agricultural and Applied Economics, 54*(2), 375–393. https://doi.org/10.1017/aae.2022.15

ALLIE KOLLMAN, MEd, is an educational consultant, supporting educators across grade levels in the development and implementation of instructional practices and district-wide initiatives. Her research interests include evidence-based grading, grading reform, assessment practices, and science education. She brings a background in teaching middle and high school biological and physical sciences, along with experience as a laboratory assistant and researcher in the biological sciences. Email: alliekollman@gmail.com

ERIN SINNER, MEd, is a high school assistant principal in North Dakota and a doctoral candidate in educational leadership at North Dakota State University. Her research interests center on grading reform, instructional equity, and the implementation of evidence-based grading practices in secondary schools. Before her current role, she served as a high school instructional coach and English teacher. Erin is passionate about making grading practices more accurate, fair, and student-centered. Email: sinnere@fargo.k12.nd.us

HANNAH NURMI, is an instructional coach in North Dakota and a doctoral candidate in educational leadership at North Dakota State University. Her work centers on advancing systemic change to ensure all students, regardless of background, have access to rigorous academic opportunities. As an instructional coach, she believes that reflective dialogue and intentional support empower teachers to challenge inequitable practices and cultivate more just, student-centered classrooms. Prior to her coaching role, Hannah taught high school English. Email: nurmih1@fargo.k12.nd.us

CAILEN O'SHEA, PhD, is an assistant professor of educational and organizational leadership at North Dakota State University. His research interests focus on school transformation, specifically enhancing instruction to provide more equitable settings for all students. Previously, Dr. O'Shea served as a behavior interventionist, 5th-grade teacher, and instructional technology coach in Title I schools in Lincoln, Nebraska. Email: cailen.oshea@ndsu.edu

Manuscript submitted: ***September 7, 2025***
Manuscript revised: ***March 1, 2026***
Accepted for publication: ***March 15, 2026***