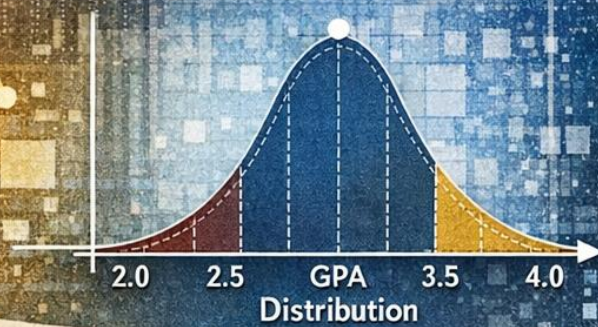


The Grading Architecture

Measurement, Systems, and Reform
in Higher Education



- Assessment Design
- Scale Cutoffs
- AI & Evaluation
- Policy Analysis

Mark Ryan, PhD

Table of Contents

Preface 3

Core Concepts 4

A-Range Scale Shifts..... 7

Grade Inflation Policy Gap..... 10

AI and Oral Exams..... 14

Rubrics and AI Reliability..... 17

Two-Tier Model 20

Program-Level Grading Health..... 23

Conclusion — Designing the Future of Grading..... 26

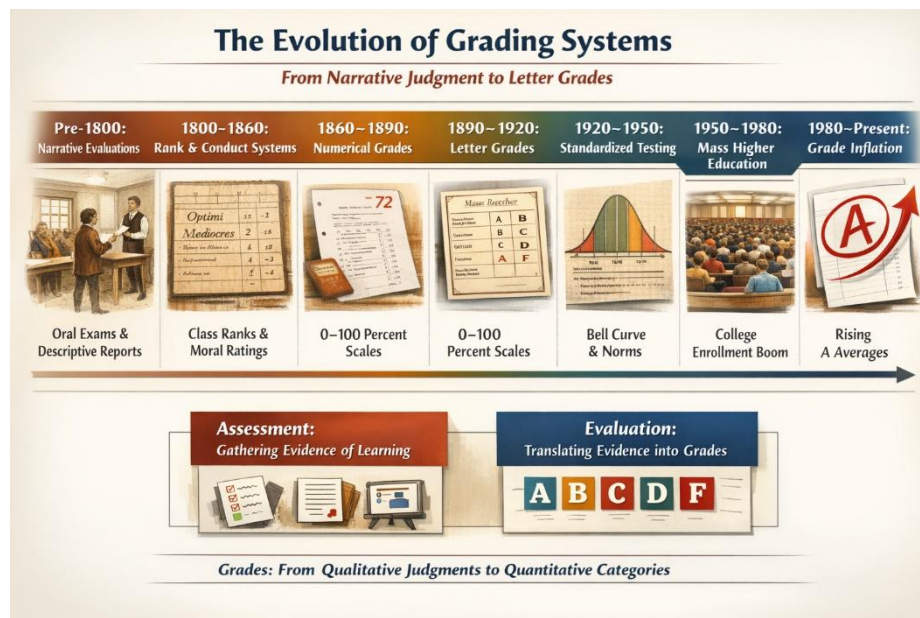
Preface

This book examines grading in higher education as a system of measurement rather than a reflection of individual instructor leniency or student motivation alone. It argues that grade outcomes are shaped primarily by structural features of evaluation design, including scale cutoffs, rubric construction, assessment formats, and institutional policy environments. Reframing grade inflation as a systems issue rather than a cultural or motivational one, the book analyzes how grading scale architecture and evaluation design produce GPA distributions and influence the signaling value of academic transcripts.

The text provides educators, administrators, and researchers with a design-oriented framework for understanding how grade outcomes emerge and how grading systems can be restructured to preserve interpretive validity, differentiation, and credibility. Key topics include the distinction between assessment and evaluation, the effects of shifting A-range thresholds, the persistence of inflation in the absence of policy response, AI's disruption of text-based evaluation, AI-supported rubric reliability, two-tier examination models grounded in Hegelian theory, and the function of program-level feedback systems. The Grading Architecture Framework and Structural Drivers model offer practical tools for structural reform grounded in measurement principles rather than perception or tradition.

Core Concepts

Grading systems emerged primarily as administrative technologies: efficient, portable mechanisms for communicating student standing across expanding schools, districts, and postsecondary pathways. As institutions grew in size and complexity, narrative evaluations—though rich in descriptive detail—proved difficult to scale. They required substantial time, varied widely by instructor, and resisted straightforward comparison across contexts. Letter grades and percentage-based schemes addressed these challenges by providing a standardized shorthand that could move reliably across classrooms, transcripts, institutions, and selection systems such as admissions or scholarships. In this sense, grading systems functioned less as pedagogical tools than as information infrastructures, designed to enable coordination within large educational bureaucracies. This historical function continues to shape contemporary



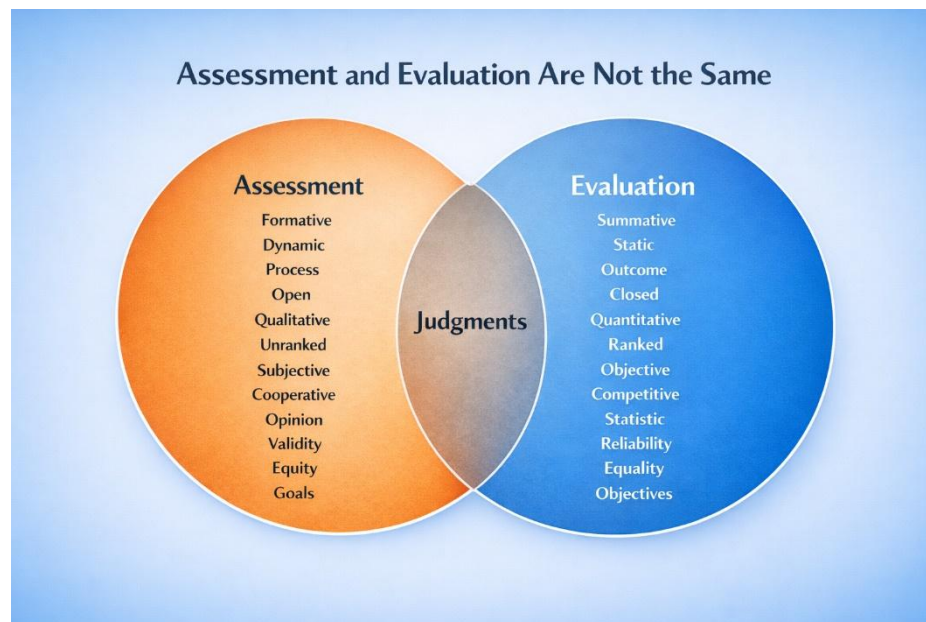
debates about grading reform. Grades are often treated as if they directly represent learning itself, yet they are better understood as symbolic summaries produced through institutional conventions. They compress varied

forms of evidence—tests, projects, participation, and judgments of quality—into a single categorical or numerical signal. That signal must be interpretable not only by teachers and students, but also by registrars, employers, accreditation bodies, and policy systems. As a result, grading practices are influenced as much by needs for comparability, efficiency, and record-keeping as by theories of learning or motivation.

Modern reform discussions frequently return to this origin point to argue that grades are not neutral reflections of achievement but socially constructed indicators embedded in organizational systems (Normann et al., 2023). Changes to grading scales, reporting formats, or evaluation criteria therefore do more than adjust classroom practice; they alter the architecture of how educational performance is represented, compared, and acted upon across institutional boundaries.

A useful starting point is the conceptual distinction between **assessment** and **evaluation**. **Assessment** is the process of gathering and interpreting evidence about

what students know and can do—through observation, dialogue, practice tasks, drafts, formative checks, and feedback cycles. It is inherently information-rich and can be



iterative, responsive, and contextual. **Evaluation**, by contrast, is the judgment process that converts evidence into a score, category, or symbol (e.g., A–F; 0–100; 1–4).

Evaluation requires rules: criteria, cut points, weighting, aggregation, and decisions about what counts (and what does not). In other words, assessment produces *evidence*; evaluation produces a *verdict*—a grade—based on chosen structures and priorities (Levy-Feldman, 2025).

When assessment and evaluation are conflated, grades are misread as direct measures of learning rather than **decisions shaped by evidence, criteria, and scale design**. Two students can demonstrate comparable mastery yet receive different grades if assignments are weighted differently, if late-work penalties are embedded, if participation is counted, or if the A-range threshold is moved from 96% to 94%. These are not trivial technicalities; they are definitional features of evaluation. The grade is the outcome of a measurement-and-aggregation system, not a transparent window into cognition or understanding.

Research shows that the modal college grade in the United States has shifted dramatically over the past several decades, with A now the most commonly awarded grade in American colleges and universities—a marked change from earlier periods when lower grades were more prevalent (Schorr, 2025). This pattern suggests that grade inflation should be understood as a structural phenomenon embedded in institutional and policy contexts, rather than merely a reflection of declining academic rigor or individual instructor leniency. Narratives that frame grade inflation as a story about personal or moral failure overlook evidence that grade distributions can shift substantially under changing conditions. Large-scale longitudinal analyses in higher education have documented significant upward shifts in the proportion of high-honors

grades and mean GPAs over time, consistent with inflationary dynamics that cannot be reduced to individual character explanations (Ciftci, 2024). Likewise, evidence from the COVID-19 era indicates that course grades rose in ways that outpaced plausible changes in underlying student achievement, suggesting that altered institutional conditions and practices changed evaluative outcomes (Tillinghast et al., 2023).

From a structural perspective, grade inflation becomes a question of signal integrity: if grades are meant to communicate achievement, what happens when the signal drifts because evaluation rules, incentives, and constraints drift? Recent work treating grades explicitly as signals shows how shifts in grading standards can weaken comparability across time and context, especially when external disruptions or local policy changes alter what grades represent (Goldhaber & Goodman Young, 2023). Put simply, grade inflation is best analyzed as an institutional property of how evaluation systems are designed and governed—how evidence is selected, converted, and communicated—rather than as a referendum on student or teacher virtue.

References

Ciftci, S. K. (2024). Grade inflation effects of capacity expansion in higher education: A longitudinal study in undergraduate teacher education programs from 2003 to 2022.

Humanities and Social Sciences Communications, 11, Article 1385.

<https://doi.org/10.1057/s41599-024-03387-6> (Nature)

Goldhaber, D., & Goodman Young, M. (2023). *Course grades as a signal of student achievement: Evidence on grade inflation before and after COVID-19* (CALDER Research Brief No. 35). CALDER / American Institutes for Research.

<https://caldercenter.org/sites/default/files/2024-11/CALDER%20Brief%2035-1123.pdf> (caldercenter.org)

Levy-Feldman, I. (2025). The role of assessment in improving education and promoting educational equity. *Education Sciences*, 15(2), 224.

<https://doi.org/10.3390/educsci15020224> (MDPI)

Normann, D. A., Haug, B. S., & Søvik, I. M. (2023). Reduced grading in assessment: A scoping review. *Studies in Educational Evaluation*, 79, 101299. [Reduced grading in assessment: A scoping review - ScienceDirect](#) (ScienceDirect)

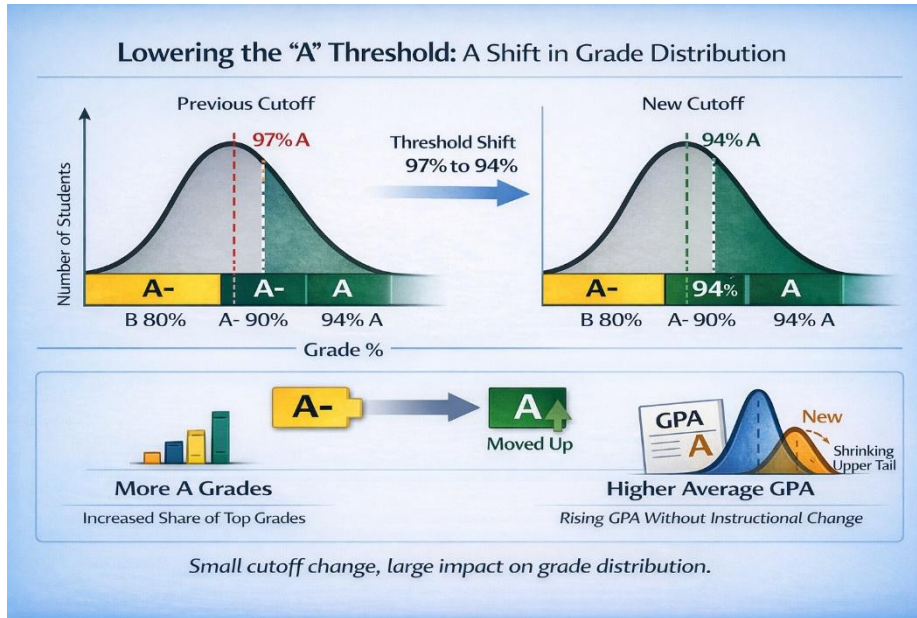
Schorr, C. (2025, September 30). *Addressing the grade inflation collective action problem*. <https://www.ed.gov/about/homeroom-blog/addressing-grade-inflation-collective-action-problem>

Tillinghast, J. A., Mjelde, J. W., & Yeritsyan, A. (2023). COVID-19 and grade inflation: Analysis of undergraduate GPAs during the pandemic. *SAGE Open*, 13(4).

<https://doi.org/10.1177/21582440231209110> (SAGE Journals)

A-Range Scale Shifts

Across many institutions, the numeric threshold for an A has drifted downward (e.g., from 97% toward 94%), a change that appears minor in absolute percentage points but can be large in its distributional consequences. The reason is not simply “easier



grading” as a moral stance; it is that grading scales function like classification rules applied to a score distribution. When the distribution is top-heavy—common in selective programs, upper-division courses, and graduate

contexts—small boundary shifts reclassify a nontrivial share of students from the next-highest category into the top category, raising the share of A grades even if instruction, assignments, and learning remain unchanged. This is a measurement effect produced by scale architecture rather than an instructional effect (Paul et al., 2022).

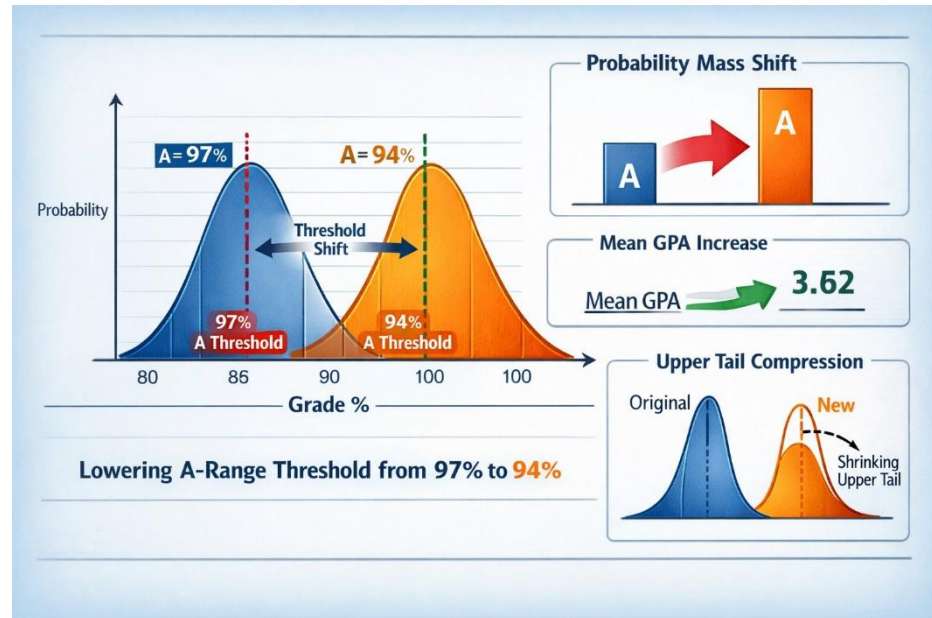
In statistical terms, the A cutoff is a decision threshold on a continuous measure. When many observations cluster near the upper bound (a “ceiling” pattern), the local density around the cutoff is high. Lowering the threshold by two or three points moves the boundary into a denser region of the distribution, sweeping more cases into the A category. During periods of widespread disruption and policy change—such as the COVID-19 era—evidence from multiple contexts shows upward shifts in awarded grades and GPAs that are consistent with structural, policy, and evaluative mechanisms affecting grade outcomes (Karadağ, 2021; Tillinghast et al., 2023). Importantly, this mechanism does not require any change in student mastery; it only requires that many students are already near the top.

Lowering the A cutoff from 97% to 94% has effects that extend beyond a simple three-percentage-point shift because performance distributions in academic settings are typically compressed near the top. In these regions, many students cluster within a narrow score band, so a small cut-score adjustment captures a disproportionately large group. Measurement research shows that cut scores are classification decisions, and minor boundary changes in dense score regions can substantially alter category membership, grade distributions, and summary indicators like GPA without

corresponding learning gains. Thus, the shift reflects a structural reclassification effect driven by ceiling compression rather than an instructional change

Once the A category expands, the mean GPA rises mechanically. This raises a second-order consequence: differentiation declines. As the A becomes the modal grade,

transcripts compress at the top, reducing the capacity of GPA to signal relative performance among high achievers. Large-scale analyses of honors classifications and GPA



patterns over time illustrate how the proportion of top outcomes can grow substantially across cohorts, which is consistent with a system drifting toward grade concentration at the upper end (Ciftci & Karadağ, 2024). The practical problem is not that students “should” receive lower grades; it is that the instrument loses resolution precisely where high-stakes selection often occurs (competitive scholarships, graduate admissions, and labor-market screening).

From a design perspective, then, the grading scale is not a neutral container for evaluation. It is an active parameter that can shift the mean and reshape variance. In this sense, moving the A-range threshold from 97% to 94% operates as a statistical lever: it increases the probability mass assigned to the top category, inflates the mean GPA, and compresses the upper tail—often without any corresponding instructional change. Downstream, this compression can alter incentives and weaken the interpretive validity of grades as indicators of mastery. Evidence on longer-run consequences of inflated grading standards further suggests that easier “top-grade” access can change how grades function as signals, with potential labor-market implications (Denning et al., 2025).

References (Open Access, 2021–2026)

Ciftci, S. K., & Karadağ, E. (2024). *Grade inflation effects of capacity expansion in higher education: A longitudinal study in undergraduate teacher education programs from 2003 to 2022*. **Humanities and Social Sciences Communications**.

<https://www.nature.com/articles/s41599-024-03387-6>

Denning, J. T., Eide, E., Murphy, R., & Pope, D. (2025). *Easy A's, less pay: The long-term effects of grade inflation* (Working paper).

https://www.econweb.umd.edu/~pope/Grade_Inflation.pdf

Karadağ, E. (2021). *Effect of COVID-19 pandemic on grade inflation in higher education in Türkiye*. **PLOS ONE**, **16**(8), e0256688.

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0256688>

Paul, C. A., Siverling, E. A., Heckler, A. F., & others. (2022). *Percent grade scale amplifies racial or ethnic inequities in introductory physics*. **Physical Review Physics Education Research**, **18**, 020103.

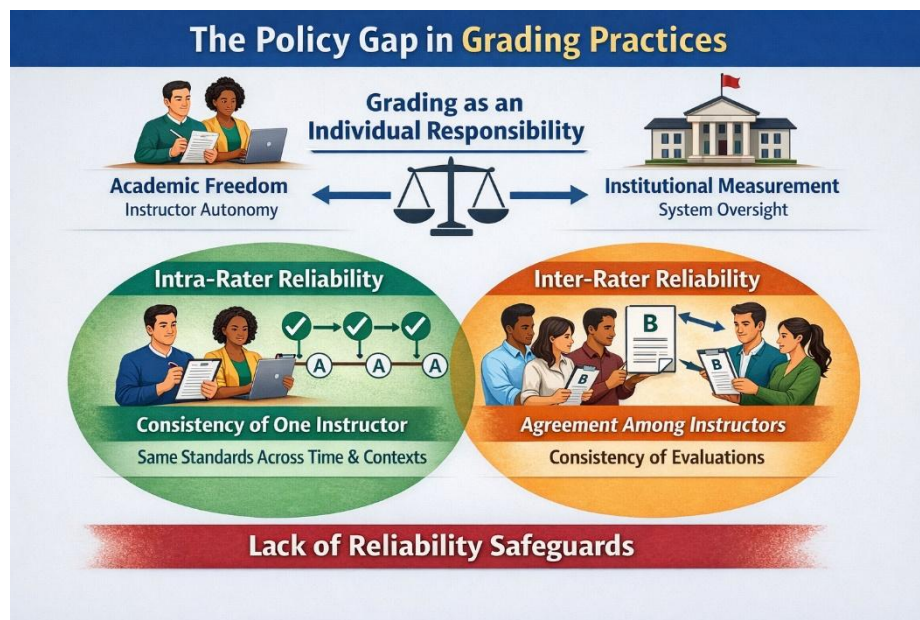
<https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.18.020103>

Tillinghast, J. A., Mjelde, J. W., & Yeritsyan, A. (2023). *COVID-19 and grade inflation: Analysis of undergraduate GPAs during the pandemic*. **SAGE Open**, **13**(4).

<https://journals.sagepub.com/doi/10.1177/21582440231209110>

Grade Inflation Policy Gap

Grade inflation has been extensively documented across higher education, yet coherent policy responses remain inconsistent and fragmented. Longitudinal empirical studies show that GPA averages have risen across institutional types and fields, with the proportion of students graduating with high honors increasing markedly over multi-decade spans, even when quality metrics outside grades show limited concurrent improvement. For example, in one longitudinal analysis of teacher education programs, the share of graduates with GPAs above 3.50 grew substantially over two decades, providing quantitative evidence of systemic upward drift in grades over time (Ciftci & Karadag, 2024). This pattern matters because grades function as institutional signals used for progression decisions, scholarship awards, graduate admissions, and employment screening. When top grades become common, the signal weakens, differentiation declines, and transcripts lose interpretive precision (Velazquez et al., 2025).



A central reason for the policy gap lies in autonomy norms surrounding grading. Grading is often framed as an extension of academic freedom, positioning it as an individual instructor responsibility

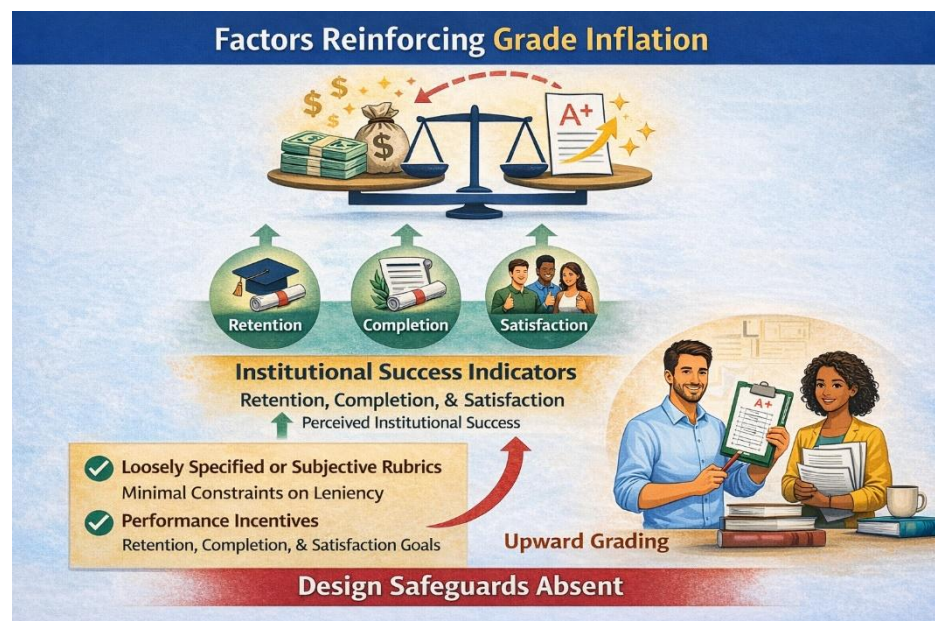
rather than as part of an institutional measurement system. One consequence is the absence of reliability safeguards at both levels: **intra-rater reliability**, the consistency with which the same instructor applies standards across time and contexts, and **inter-rater reliability**, the degree of agreement among different instructors evaluating the same work. Without structures that support consistency, grading can drift within a single instructor's practice and vary widely across sections or evaluators, weakening the interpretive stability of GPA. Research on grading practice underscores how inconsistency among raters and over time complicates the task of treating grades as comparable indicators of learning (Pack et al., 2024). Governance systems therefore rarely include calibration mechanisms such as cross-section moderation, shared performance standards, blind re-marking, or program-level grade audits. In the absence of these design features, grade distributions emerge from decentralized and variably

consistent decisions, allowing upward drift to develop as a systemic measurement effect rather than a deliberate policy choice.

A second sustaining condition is limited measurement literacy at the policy level. Institutions routinely track grade averages, completion rates, and pass rates, but far less attention is given to the interpretive validity of grades as measures of learning or mastery. It is critical to recall that validity is assessed, not evaluated: validity refers to the degree to which evidence supports the interpretations made from documented results, not faculty or administrative evaluation of grade trends. When grades rise, leaders may interpret this as positive progress based on evaluation criteria, yet rising grades do not automatically constitute validity evidence. Without external anchors—such as standardized evaluations, performance benchmarks, or licensure outcomes— institutions lack the empirical basis to know whether grades continue to represent mastery or have shifted due to scale effects, rubric drift, or design changes (Velazquez et al., 2025).

Third, accountability structures often reinforce the very conditions that enable inflation.

Retention, completion, and student satisfaction are frequently emphasized as institutional success indicators. Because higher grades can correlate with improved persistence and perceived



success, upward grading may indirectly align with performance incentives. Courses with loosely specified rubrics or heavy reliance on subjective assessment formats tend to produce fewer constraints on grading decisions, reducing friction against leniency. In the absence of design safeguards, inflation becomes structurally compatible with institutional goals.

Temporal dynamics further weaken reform incentives. Students benefit immediately from higher grades through scholarships and admissions opportunities, while the long-term erosion of grade signaling value unfolds gradually. Employers and graduate programs may compensate by attending more to institutional prestige, standardized

tests, or experiential credentials, but these adjustments occur downstream and outside institutional accountability loops. Because the costs of signal degradation are delayed and diffuse, they rarely generate urgent policy response.

Addressing the policy gap does not mean removing instructor discretion; it means recognizing grading as a measurement system that warrants intentional design oversight. Institutions can use program-level dashboards that align course grades with independent indicators of performance, allowing leaders to test whether grade interpretations remain evidence-based. Moderation protocols, shared exemplars, and cross-course calibration processes can bring coherence to standards without flattening instructional individuality. Transparent rubrics, sufficient assessment sampling, and archiving of graded work create the conditions for periodic validity checks. When grade distributions are reported alongside direct evidence of mastery, rising GPAs become diagnostic signals for inquiry rather than assumed proof of improvement.

Extending this design perspective, transcripts of student work—whether written or derived from recorded performance tasks—can document a structured pre-submission phase in which students apply a university-developed rubric and use AI tools for an initial scoring pass and narrative feedback on strengths and gaps. This shifts rubrics from purely summative instruments to formative diagnostic frameworks. Students engage criteria earlier, revise with targeted guidance, and approach instructor grading with clearer alignment to expectations. The instructor’s role continues to involve evaluation in assigning scores and assessment through individualized feedback, but the process leading to those judgments becomes more transparent, grounded in evidence, and oriented toward supporting learning.

Ultimately, grade inflation persists less because it is invisible than because it is structurally ungoverned. Without systems designed to assess validity evidence and maintain interpretive stability, grading drifts in response to institutional pressures, incentive structures, and local norms. The policy challenge is therefore not cultural correction alone, but the construction of measurement-capable governance that ensures grades remain credible indicators of learning rather than artifacts of scale design.

References

- Ciftci, S. K., & Karadag, E. (2024). *Grade inflation effects of capacity expansion in higher education: A longitudinal study of teacher training programs (2003–2022)*. Humanities and Social Sciences Communications. <https://www.nature.com/articles/s41599-024-03387-6>
- Pack, A., Barrett, A., & Escalante, J. (2024). *Large language models and automated essay scoring: Insights into validity and reliability*. Computer-Assisted Education AI

(open access).

<https://www.sciencedirect.com/science/article/pii/S2666920X24000353>

Almaghaslah, D. (2025). *Can ChatGPT-generated MCQs reduce grade inflation in pharmacy education?* *Frontiers in Pharmacology*.

<https://www.frontiersin.org/articles/10.3389/fphar.2025.1516381/full>

Diagnosing grade inflation: A curriculum analytics approach to quality assurance in higher education (2025). Open-access manuscript.

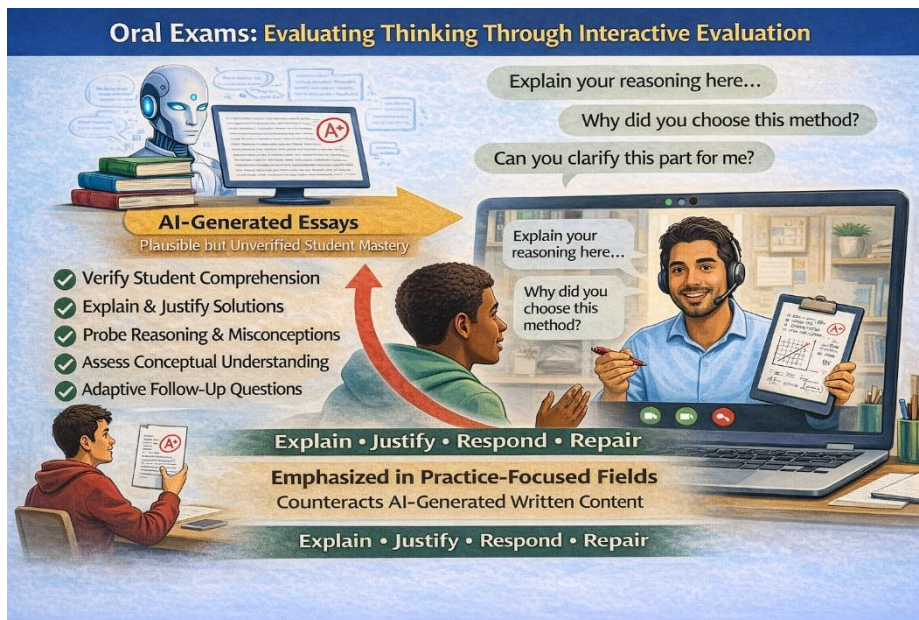
https://www.researchgate.net/publication/396269448_Diagnosing_Grade_Inflation_A_Curriculum_Analytics_Approach_to_Quality_Assurance_in_Higher_Education

“How consistent are humans when grading programming assignments?” (2024). ArXiv preprint. <https://arxiv.org/abs/2409.12967>

AI and Oral Exams

AI has destabilized the integrity of text-based assessment because the “product” (an essay, short answer, or problem explanation) can be generated or heavily revised by large language models with minimal trace. As a result, the central question is shifting from “*Did the student write this?*” to “*Can the student demonstrate the underlying reasoning in real time?*” Recent syntheses of generative AI in academic integrity argue that detection tools are an unreliable primary safeguard and that institutions should prioritize assessment redesign that foregrounds authentic performance and defensible evidence of learning (Bittle & Seifert, 2025).

Oral exams address this shift by evaluating thinking as a performance: students must explain choices, justify steps, respond to probing follow-ups, and repair misconceptions under questioning. Because the interaction is adaptive, the assessment can quickly move beyond memorized phrasing into conceptual



understanding, making it harder to outsource cognition to a pre-generated script. In practice-focused disciplines, oral assessment has been explicitly recommended as a way to verify comprehension in an era when AI

can produce plausible written work that may not reflect student mastery (Eachempati et al., 2025).

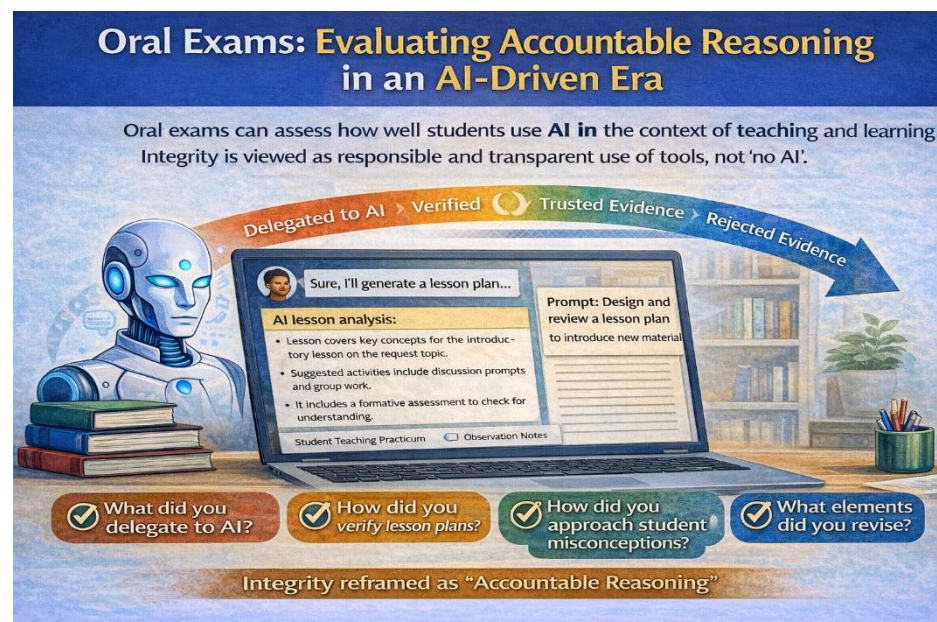
Importantly, oral exams are not simply “talking instead of writing.” They are a format change that allows instructors to sample reasoning at multiple depths: (a) foundational concepts, (b) application to a novel case, and (c) metacognitive reflection on limitations, uncertainty, and tradeoffs. This aligns with the broader “beyond detection” direction in the assessment literature: shifting from AI-policing to designing tasks where process evidence and situated judgment are central (Kickbusch et al., 2025). When oral exams are structured—using prompts mapped to learning outcomes, shared rubrics, and consistent timing—they can function as defensible measures rather than informal conversations.

A major concern is reliability and fairness. Oral assessments can introduce variability if questioning differs substantially across students or if scoring criteria are implicit. A systematic review of oral assessments in higher education emphasizes that validity and academic integrity benefits are strongest when orals are integrated with careful design choices: transparent criteria, examiner training, and moderation processes that reduce inconsistent judgment (Nallaya et al., 2024). In other words, oral exams work best when treated as a measurement system, not a spontaneous “gut check.” Here’s a clearer, more scholarly rewrite:

Strengthening grading reliability does not require removing instructor judgment; it requires designing evaluation as a structured measurement process. One promising approach is the integration of artificial intelligence as a calibration support tool. When students submit work alongside the official assignment rubric and a verbatim transcript from an oral examination, AI systems can perform a first-pass rubric alignment, mapping evidence in the transcript directly to performance criteria. This creates an auditable trail linking observed performance to scoring descriptors.

Because the rubric is standardized and the transcript preserves the full performance record, AI-assisted scoring introduces consistency at two levels. Intra-rater reliability improves because instructors can compare their judgments against the same rubric-tagged evidence across time, reducing drift in how standards are applied. Inter-rater reliability increases because multiple evaluators — human or AI-supported — reference identical textualized evidence rather than memory, impression, or partial notes. The transcript functions as a shared evidentiary artifact, while the rubric serves as the classification framework.

In this model, AI does not replace professional evaluation; it operationalizes the rubric, flags discrepancies, and highlights where human raters diverge from rubric-aligned evidence. The result is a more



transparent, repeatable, and evidence-linked grading process that preserves instructor authority while substantially increasing scoring reliability.

Oral formats also respond to the pragmatic reality that diversified assessment reduces single-point vulnerability. Practitioner-oriented reviews in open-access venues increasingly recommend combining orals with projects, reflections, and supervised checkpoints to decrease dependence on writing alone (Evangelista, 2025). This does not require eliminating writing; it requires reserving high-stakes decisions for moments where identity, reasoning, and ownership of ideas are directly observable.

Finally, oral exams can be future-facing rather than merely defensive. If programs openly acknowledge that AI will be part of professional practice, oral exams become a venue to evaluate how students *use* tools: asking them to explain what they delegated to AI, how they verified outputs, what evidence they trusted, and what they rejected. That stance reframes integrity from “no AI” to “accountable reasoning.” Research on generative AI’s impact on authentic assessment underscores that the most durable response is designing assessments where the learner must demonstrate judgment, context sensitivity, and explanation—capacities that remain legible in oral defense (Kofinas & Katsanidou, 2024).

References

Bittle, K., & Seifert, T. (2025). *Generative AI and academic integrity in higher education*. *Information*, 16(4), 296. <https://www.mdpi.com/2078-2489/16/4/296>

Eachempati, P., Komattil, R., & colleagues. (2025). *Should oral examination be reimaged in the era of AI?* *Advances in Physiology Education*. <https://journals.physiology.org/doi/full/10.1152/advan.00191.2024>

Evangelista, E. D. L. (2025). *Ensuring academic integrity in the age of ChatGPT*. [ERIC full-text PDF]. <https://files.eric.ed.gov/fulltext/EJ1460216.pdf>

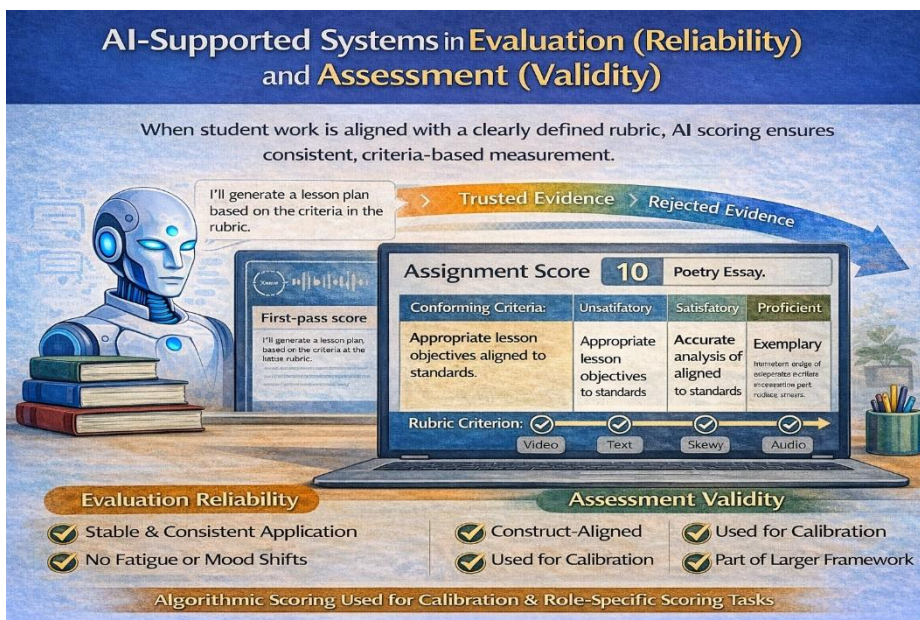
Kickbusch, S., Ashford-Rowe, K., Kemp, A., Boreland, J., & Huijser, H. (2025). *Beyond detection: Redesigning authentic assessment in an AI-mediated world*. *Education Sciences*, 15(11), 1537. <https://doi.org/10.3390/educsci15111537>

Nallaya, S., Gentili, S., Weeks, S., & Baldock, K. (2024). *The validity, reliability, academic integrity and integration of oral assessments in higher education: A systematic review*. *Issues in Educational Research*, 34(2), 629–646. <http://www.iier.org.au/iier34/nallaya.pdf>

Rubrics and AI Reliability

Rubrics were developed to make academic judgment more transparent, criterion-referenced, and consistent. By articulating performance levels and linking them to learning outcomes, rubrics reduce idiosyncratic grading and help align scoring with instructional intent. Yet decades of measurement research show that rubrics alone do not eliminate variability. Human raters differ in severity, interpretation of descriptors, attention to evidence, and susceptibility to fatigue or context effects. Even the same instructor may shift over time, a phenomenon known as intra-rater drift. As assessment increasingly functions as a high-stakes signal for progression, certification, and accountability, variability becomes a measurement problem rather than merely a pedagogical one (Brookhart, 2021).

AI-supported scoring systems introduce a design response to this challenge. When student work—text, audio transcripts from oral tasks, or structured performance



evidence—is aligned with a clearly specified rubric, AI models can be trained or configured to map observable features to rubric criteria with high procedural consistency. Unlike human raters, algorithmic systems do not

experience fatigue, mood shifts, or shifting interpretive frames across time. This makes them well suited for first-pass scoring or calibration roles, where the goal is stable application of predefined criteria rather than holistic judgment. Research on automated scoring emphasizes that such systems are most defensible when tightly coupled to construct-relevant features and when used as part of a broader validity argument rather than as stand-alone arbiters (Zhai et al., 2022).

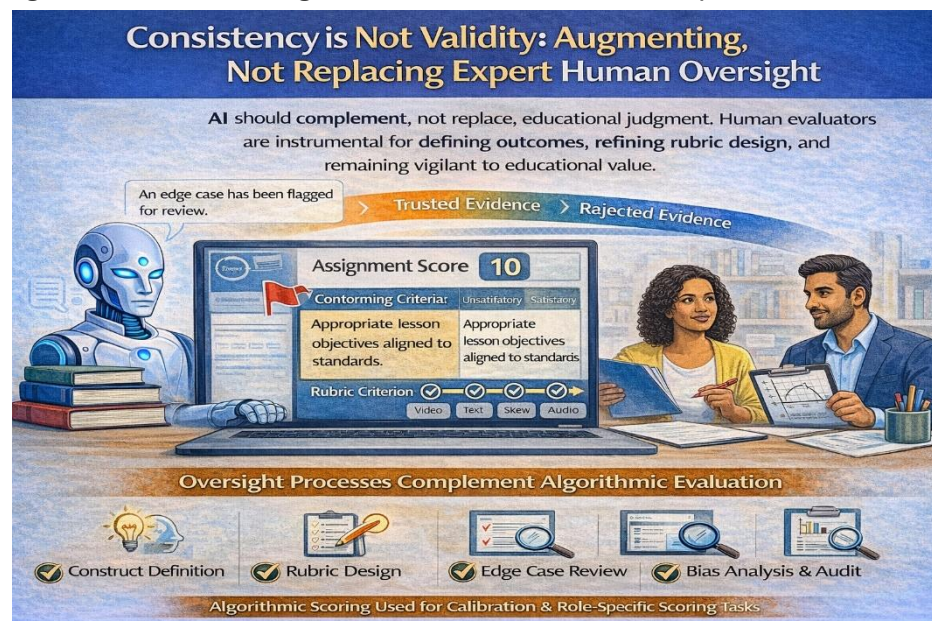
A key contribution of AI scoring is the reduction of intra-rater drift. In traditional settings, instructors' standards can gradually shift across a semester as they encounter stronger or weaker work, implicitly recalibrating expectations. AI systems, by contrast, apply the same learned decision boundaries to early and late submissions alike. This stability supports fairness across cohorts and sections, particularly in large programs where

multiple instructors interpret shared rubrics. Studies of AI in educational measurement note that algorithmic consistency can strengthen reliability indices, especially when human raters use AI outputs as reference points during moderation and calibration sessions. (Vatankhah et al., 2026).

However, consistency is not synonymous with validity. A system can apply a rubric flawlessly yet still measure the wrong construct if the rubric itself is poorly aligned with learning goals. For this reason, contemporary scholarship frames AI scoring as augmentation rather than replacement of human expertise. Human evaluators remain essential for construct definition, rubric design, review of edge cases, and periodic audits of model behavior. Oversight processes—such as double scoring of samples, bias analysis, and review of discrepant cases—ensure that AI outputs remain tethered to educational intent and do not encode unintended inequities (Holmes & Tuomi, 2022). In this model, AI handles the repetitive application of criteria, while humans safeguard meaning. (Kaldaras, 2024).

AI systems also enhance transparency when designed with evidence tracing. Some platforms can highlight which textual segments, discourse moves, or performance

indicators triggered specific rubric levels. This creates an auditable link between observed evidence and assigned scores, supporting both accountability and formative use. Students



can see how their work aligns with criteria, and instructors can quickly identify patterns of misunderstanding across cohorts. Such features move rubrics from static scoring guides to dynamic feedback instruments, strengthening the learning function of evaluation alongside its certifying role.

The most defensible implementation, therefore, is a hybrid architecture. AI performs large-scale, criterion-aligned scoring to stabilize application of standards; human raters review samples, resolve ambiguities, and conduct validity checks. Programs can schedule periodic recalibration, comparing human and AI judgments to detect drift or construct misalignment. Rather than displacing professional judgment, AI systems

operationalize rubric logic with consistency, freeing educators to focus on interpretation, feedback, and instructional response. In this sense, AI does not automate evaluation; it helps transform grading into a more explicitly designed measurement process in which reliability is engineered and validity is continuously monitored.

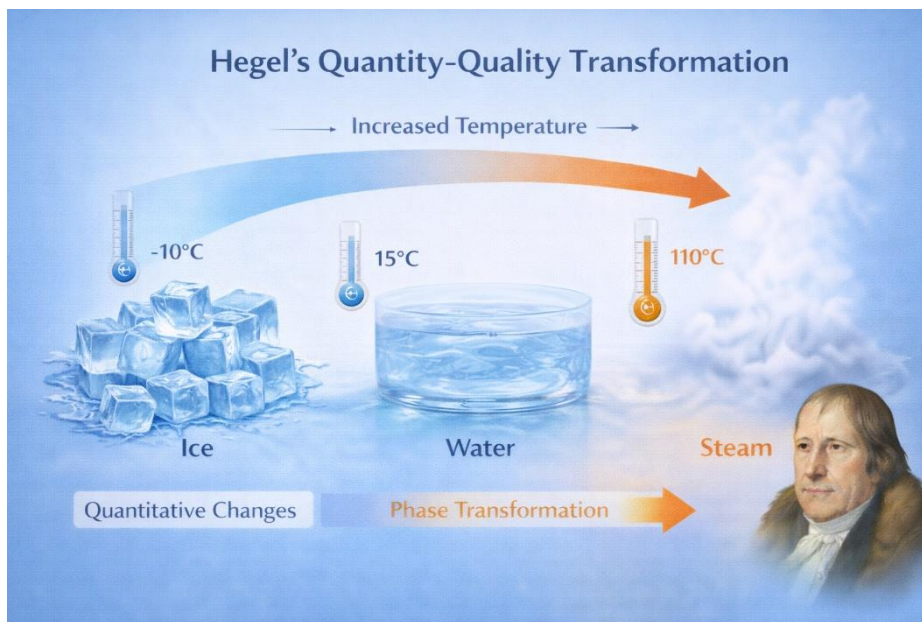
References

- Brookhart, S. M. (2021). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, 6, 1–9. <https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2018.00022/full>
- Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, 57(4), 542–570. <https://doi.org/10.1111/ejed.12533>
- Kaldaras, L. (2024). *Developing valid assessments in the era of generative artificial intelligence: Construct alignment and intended use evidence*. *Frontiers in Education*. <https://doi.org/10.3389/feduc.2024.1399377>
- Vatankhah Barenji, R., Salimi, N., & Khoshgoftar, S. (2026). *An LLM-Powered Assessment Retrieval-Augmented Generation (RAG) for Higher Education* (preprint). arXiv. <https://doi.org/10.48550/arXiv.2601.06141>
- Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automated scoring of complex student responses. *Computers & Education: Artificial Intelligence*, 3, 100065. <https://doi.org/10.1016/j.caeai.2022.100065>

Two-Tier Model

A two-tier exam structure—offering a shorter and a longer pathway—functions as an evaluation design strategy, not an assessment variation. For example, students might choose between a 1,500-word final essay or a 3,000-word essay, or between a 10-minute and a 20-minute oral exam. Both options are anchored to the same course outcomes and scored with shared performance criteria. The difference lies in the quantity of cognitive performance required, not in what is being measured. Research on flexible evaluation formats shows that structured choice can increase engagement and reduce performance stress when expectations are explicit and access barriers are minimized (Barua & Lockee, 2025). Effectiveness depends on both pathways measuring identical constructs while differing in depth, integration, and transfer demands.

Hegel's **Law of the Transformation of Quantity into Quality** explains how gradual increases in a measurable factor can produce a fundamental change in outcome. In pedagogical terms, increases in cognitive demand function the same way. When



students choose between a shorter and a longer exam, they are not merely selecting duration; they are selecting the depth of intellectual engagement possible within the task.

A longer exam allows for expanded argumentation, integration of wider evidence, and sustained reasoning. This extended cognitive space enables learners to develop more complex syntheses, examine counterarguments, and demonstrate higher-order evaluative judgment. The additional time and mental effort — quantitative increases — create the conditions for qualitatively different performances that reveal deeper conceptual connections.

Structured choice, therefore, sharpens the resolution of learning evidence by aligning the format of evaluation with the level of thinking students are ready to demonstrate.

Hegel's idea that quantity becomes quality helps explain why longer exams can reveal different kinds of learning, not just more of the same. When students choose a longer exam, they have

more time to build arguments, connect ideas, integrate broader evidence, and sustain reasoning.

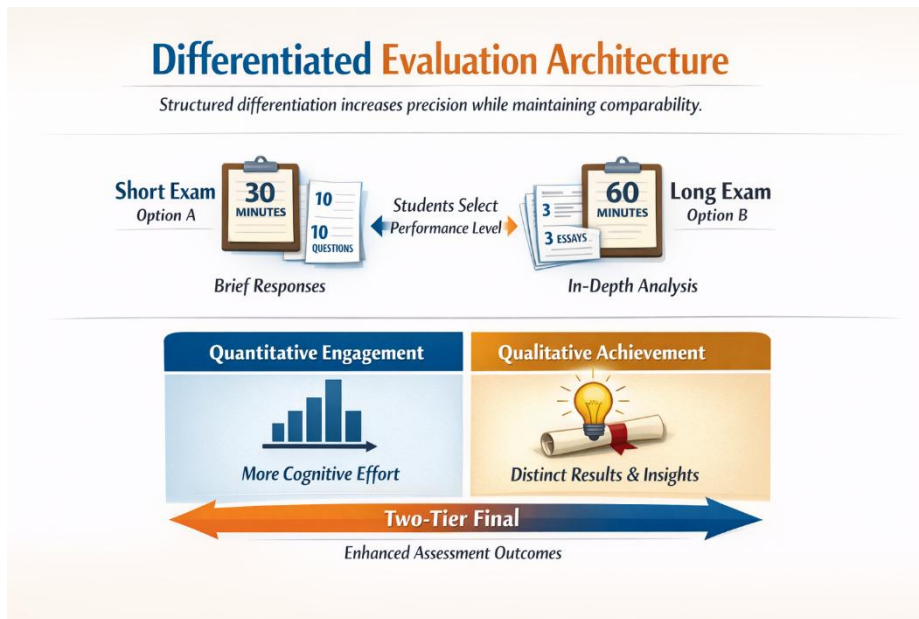
These quantitative increases in cognitive work can produce a

qualitative shift toward deeper synthesis, stronger counterarguments, and clearer conceptual understanding.



Student choice also reflects disposition: interest and motivation shape how deeply learners engage cognitively. When the format aligns with what students find meaningful, their commitment increases. Research on choice shows tiered evaluation can strengthen effort without lowering rigor, especially when expectations are transparent and tied to shared learning goals. The longer pathway signals greater performance complexity, not easier grading (Davies et al., 2025). Design integrity must preserve validity and comparability. Systematic reviews identify clearly defined criteria and instructional support as essential conditions (Heil & Ifenthaler, 2023). In a two-tier model, this requires one rubric aligned to outcomes, common scoring dimensions across tiers, and tier-specific descriptors only where complexity legitimately differs. This maintains measurement scale integrity.

The distribution of top-tier performances becomes differentiated not by lowering standards, but by variation in performance length and complexity. At the same time, participation conditions still matter. Flexibility mechanisms may widen gaps if time, preparation resources, or external obligations constrain access (Supriya et al., 2024). Longer pathways thus require equitable scaffolding, while shorter pathways must remain a complete and standards-aligned demonstration of course outcomes.



Research on adaptive and differentiated evaluation systems suggests structured differentiation can increase precision while maintaining comparability (Machkour et al., 2025). A two-tier final thus operates as

differentiated evaluation architecture: students select performance length, and the system captures quantitatively greater cognitive engagement that produces qualitatively distinct achievement evidence.

References

- Barua, L., & Lockee, B. (2025). Flexible assessment in higher education: A comprehensive review of strategies and implications. *TechTrends*, 69, 301–309. <https://doi.org/10.1007/s11528-025-01039-3>
- Davies, C. A., Cordier, R., Graham, P., Littlefair, D., Speyer, R., & Melo, D. (2025). *Interventions to improve connectedness, belonging, and engagement in secondary schools: A systematic review and meta-analysis*. **Education Sciences**, 15(5), 582. <https://doi.org/10.3390/educsci15050582>
- Heil, J., & Ifenthaler, D. (2023). Online assessment in higher education: A systematic review. *Online Learning*, 27(1), 187–218. <https://doi.org/10.24059/olj.v27i1.3398>
- Jopp, R., Pallant, J. L., & Russell, H. (2023). Choose your own adventure: Understanding why students prefer certain types of assessment. *Journal of University Teaching & Learning Practice*, 20(7), Article 11. <https://doi.org/10.53761/1.20.7.11>
- Machkour, M., El Jihaoui, M., Lamalif, L., Faris, S., & Mansouri, K. (2025). Toward an adaptive learning assessment pathway. *Frontiers in Education*, 10, 1498233. <https://doi.org/10.3389/educ.2025.1498233>
- Supriya, K., Bang, C., Ebie, J., Pagliarulo, C., Tucker, D., Villegas, K., Wright, C., & Brownell, S. E. (2024). Optional exam retakes reduce anxiety but may exacerbate score disparities between students with different social identities. *CBE—Life Sciences Education*, 23(3), ar30. <https://doi.org/10.1187/cbe.21-11-0320>

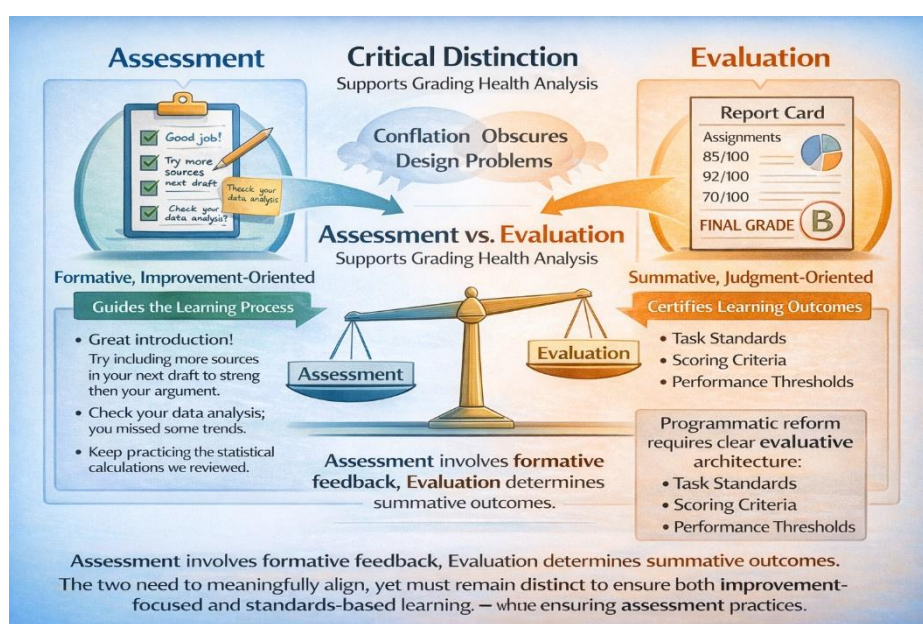
Program-Level Grading Health

Program-level grading health refers to the extent to which evaluation outcomes—summative, criterion-referenced numerical judgments—accurately represent what a curriculum intends students to learn. Grades at this level are not isolated instructor decisions; they are system signals reflecting curricular alignment, instructional coherence, and the design of evaluative measures. When grade patterns show compression at the upper end, wide cross-section variability, or weak differentiation, the issue is often structural rather than individual, indicating misalignment between intended learning, teaching activity, and evaluative design (Clark & Karvonen, 2021).

A critical distinction supports grading health analysis: assessment involves formative, feedback-oriented commentary that guides learning, whereas evaluation produces

summative determinations such as scores, grades, or credentialing decisions.

Assessment informs learning processes; evaluation certifies learning outcomes.

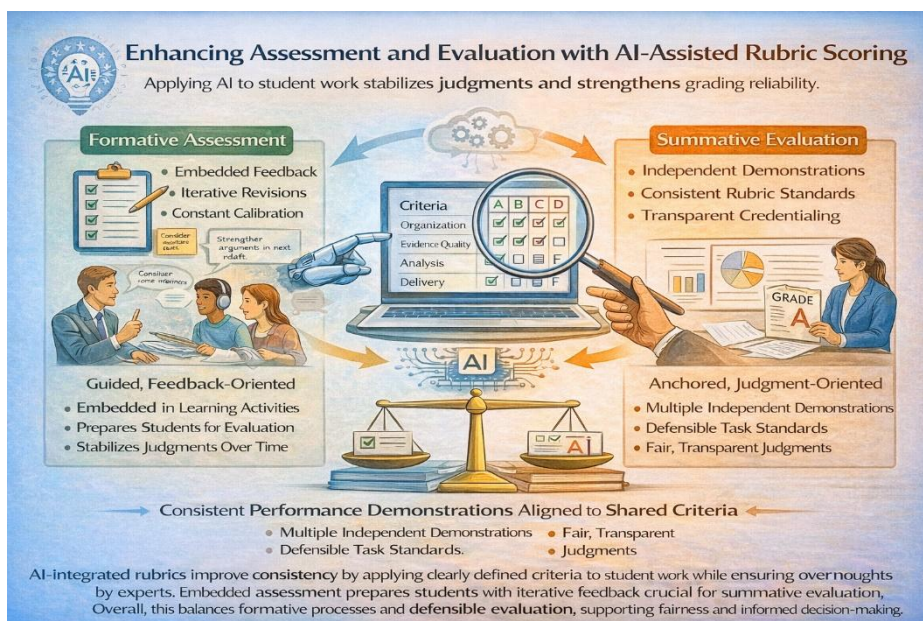


Conflating the two obscures design problems. For example, rich formative feedback may be present while summative evaluative criteria remain unclear or inconsistently applied. Program-level reform therefore focuses primarily on the architecture of evaluation systems—task standards, scoring criteria, and performance thresholds—while ensuring assessment practices meaningfully prepare students to meet those standards (Charlton & Newsham-West, 2024).

Constructive alignment provides a core framework for evaluating grading health. Intended learning outcomes, instructional experiences, formative assessment, and summative evaluation should form a coherent chain. When evaluation tasks do not directly sample the knowledge and skills emphasized in instruction, grades lose interpretive validity. Program-level assessment planning research shows that mapping outcomes across courses strengthens vertical and horizontal coherence, ensuring that summative evaluations reflect cumulative learning rather than fragmented local

expectations (Charlton & Newsham-West, 2024). In such systems, grades communicate developmental progression rather than isolated course performance.

As has been noted, using AI to apply rubrics to transcripts of student performance in video, oral, or written assignments offers a practical way to address the persistent



problem of limited intra-rater and inter-rater reliability. Algorithmic scoring anchored to clearly defined criteria can help stabilize judgments, reduce drift over time, and increase consistency across evaluators

while still allowing professional oversight. Instructionally embedded assessment theory further clarifies the relationship between formative assessment and summative evaluation. When assessment opportunities are integrated into learning activities, students receive iterative feedback that prepares them for later evaluative demonstrations (Clark & Karvonen, 2021). These formative processes do not replace evaluation; they strengthen its evidentiary base and interpretive validity. A sound grading system rests on multiple independent demonstrations aligned to shared criteria, rather than single high-stakes events or loosely defined tasks, supporting both fairness and defensible decision-making.

End-of-Course (EOC) surveys can function as diagnostic tools for grading health when interpreted appropriately. These surveys do not evaluate student achievement; instead, they provide perception data about clarity, alignment, workload, and feedback processes. Such information reveals whether evaluation designs were transparent and whether assessment practices supported preparation for evaluative tasks. Research on student feedback systems emphasizes that survey timing, item design, and interpretation protocols influence the usefulness of these data for instructional and program reform (Williams, 2024). Used systematically, EOC results help identify whether grading issues stem from unclear expectations, insufficient practice opportunities, or inequitable participation conditions.

Curriculum analytics research also supports program-level monitoring of grade distributions as quality indicators. Patterns such as ceiling clustering or rapid upward shifts in high grades may signal evaluation designs with limited performance differentiation or insufficiently discriminating criteria (Velazquez et al., 2025). Importantly, such analyses do not presume declining standards; instead, they prompt review of evaluative architecture—rubric specificity, performance levels, and task complexity. Combined with survey diagnostics, these data enable evidence-based reform.

A healthy program-level grading system thus integrates: (1) aligned evaluative criteria across courses, (2) formative assessment structures that prepare students for summative demonstrations, (3) transparent communication of standards, and (4) ongoing diagnostic feedback from student experience data. In this model, grades become valid indicators of learning achievement because evaluation design, instructional preparation, and student experience operate as a coherent system (Clark & Karvonen, 2021; Velazquez et al., 2025).

References

- Charlton, N., & Newsham-West, R. (2024). *Enablers and barriers to program-level assessment planning*. *Higher Education Research & Development*, 43(5), 1074–1088. <https://www.tandfonline.com/doi/full/10.1080/07294360.2024.2307933>
- Clark, A. K., & Karvonen, M. (2021). *Instructionally embedded assessment: Theory of action for an innovative system*. *Frontiers in Education*, 6, Article 724938. <https://www.frontiersin.org/articles/10.3389/feduc.2021.724938/full>
- Velazquez, L., Atenas, B., Cruz Hernández, N., Castro Palacio, J. C., & Monsoriu, J. A. (2025). *Diagnosing grade inflation: A curriculum analytics approach to quality assurance in higher education*. *Studies in Higher Education*. <https://doi.org/10.1080/03075079.2025.2572513>
- Williams, A. (2024). *Delivering effective student feedback in higher education: An evaluation of the challenges and best practice*. *International Journal of Research in Education and Science*, 10(2), 473–501. <https://files.eric.ed.gov/fulltext/EJ1426687.pdf>
- Bart, W. M., Abulela, M. A. A., & Khalaf, M. A. (2026). *Investigating course-level effects on student evaluations of teaching in higher education*. *Education Sciences*, 16(1), 94. <https://www.mdpi.com/2227-7102/16/1/94>

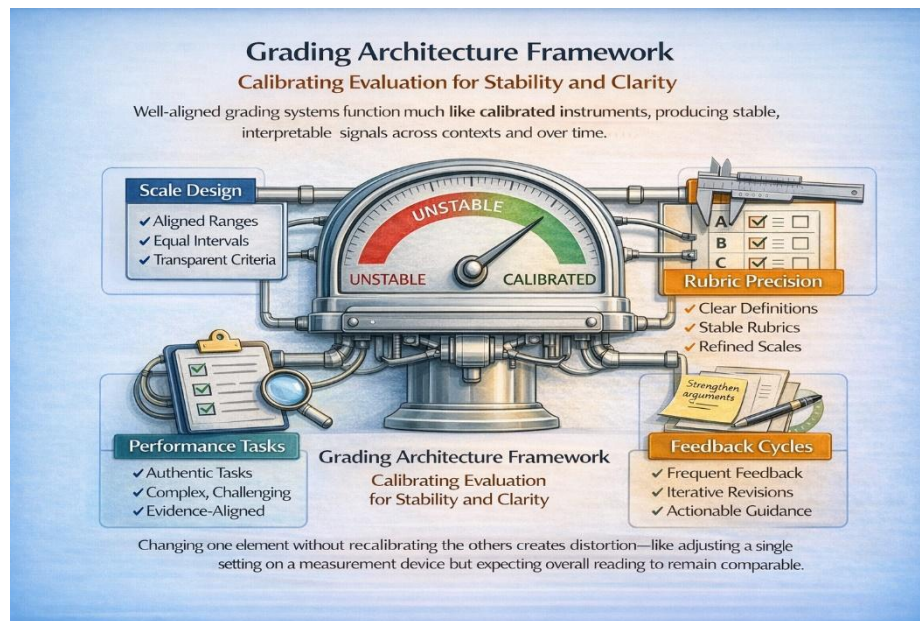
Conclusion — Designing the Future of Grading

Grading—and the long-documented inflationary drift of grades—has been an enduring issue in education for decades. In higher education, grading is often treated as a downstream byproduct of instruction, yet it functions more accurately as a measurement system shaped by structural design choices. Contemporary national data show that the A has become the most frequently awarded grade on many campuses, a pattern widely interpreted as evidence of systemic grade inflation rather than sudden, universal increases in student mastery (ACT, 2022). When grading systems drift without calibration, GPA shifts from serving as a stable indicator of learning to reflecting evolving institutional norms and scale practices. Under these conditions, grades signal changes in design and policy at least as much as changes in performance, underscoring the need to treat grading as an engineered measurement architecture rather than an incidental instructional outcome.

Viewed structurally, GPA outcomes are not simply the product of student effort or instructor leniency. They emerge from interacting design decisions: scale structures, rubric specificity, task formats, and institutional policies. Evidence from large datasets continues to document persistent upward grade trends over time, including ceiling effects that mask variation at the top of the scale and complicate interpretation across settings (Sanchez & Moore, 2022). Internationally, longitudinal work likewise links rising GPAs to system-level changes (e.g., program expansion) and disruption periods, reinforcing that inflation is often produced by macro-conditions rather than individual failings (Ciftci & Karadag, 2024).

Reframing grading as a system clarifies the path forward: healthy evaluation environments resemble calibrated instruments that produce stable, interpretable signals across contexts and over time. The Grading Architecture Framework, as argued in this booklet, positions scale design, performance tasks, rubric precision, and feedback cycles as interdependent components. Changing one element without

recalibrating the others creates distortion—much like adjusting a single setting on a



measurement device and expecting the overall reading to remain comparable.

Reliability is a central concern in any measurement architecture. When criteria are loosely specified, interpretive validity weakens;

when judgments vary across raters or time, the evidentiary base for grades erodes. Studies of constructed-response scoring repeatedly show that rater variability is not a minor nuisance but a policy-relevant threat to score dependability—especially when decisions rely on small differences (Huang & Whipple, 2023). In GPA systems, that same logic applies: if “A-range” judgments are unstable, distribution shifts can be mistaken for learning gains.

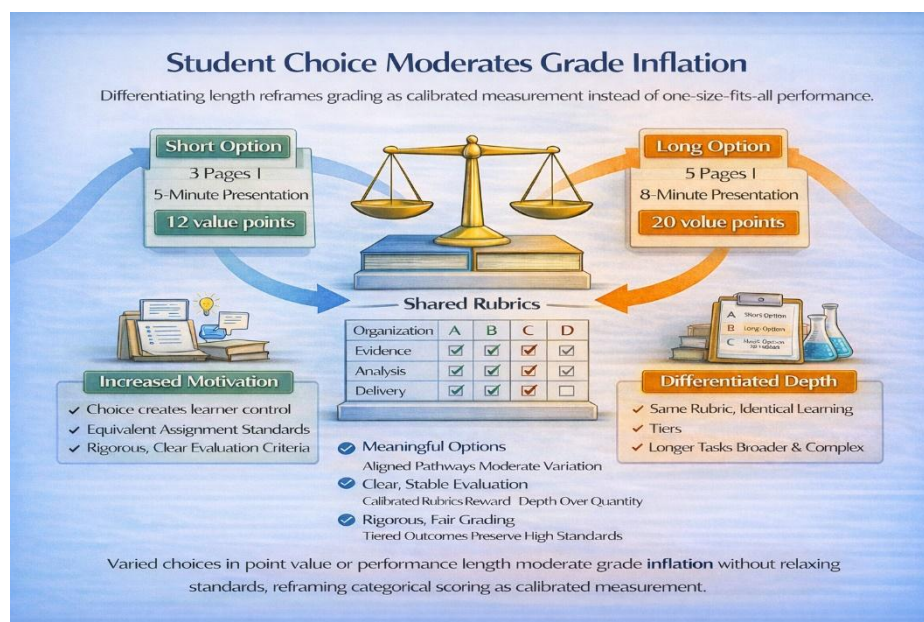
Artificial intelligence enters this landscape not as a shortcut, but as an alignment technology. When transparently designed and anchored to clearly articulated learning outcomes, AI-supported rubrics can stabilize the application of criteria across artifacts (written, oral, or video) and across scorers—while preserving human review for interpretive oversight. In parallel, research on AI-generated feedback in academic writing suggests that AI can produce clear, specific guidance that students often find useful, even when learning gains are comparable to human feedback—pointing to a pragmatic role for AI in strengthening consistency and reducing workload without removing educators from the loop (Escalante et al., 2023).

Equally important is the instructional function of the system. When students can interact with criteria through AI-supported feedback before summative judgment, assessment becomes a structured rehearsal space: learners clarify expectations, iterate, and enter high-stakes demonstrations with stronger alignment. The goal is not higher grades, but clearer grades—improved resolution between levels of performance.

One emerging response involves structured student choice in assignment point value or

performance length. When learners select between distinct pathways—such as shorter or longer demonstrations aligned to the same learning outcomes—satisfaction may rise because students perceive greater control, clarity, and fairness in evaluation. At the same time, grade inflation pressures can be moderated. Not all students opt for the most extensive pathway, which naturally redistributes performance tiers without lowering standards. Differentiation occurs through variation in depth and complexity rather than relaxed criteria.

When both options are scored with shared rubrics and identical outcome alignment, rigor is preserved while authentic motivation plays a role. This design reframes grading as calibrated measurement rather than a one-size-fits-all performance event.



Voluntary student completion of End-of-Course Surveys at response rates above 50 percent is essential for credible program evaluation. When most voices are represented, findings more accurately reflect classroom experience rather than extreme opinions. Colleges share responsibility for cultivating a scholarly culture where continuous improvement is normal, evidence-informed, and oriented toward strengthening curricula and teaching. Transparent communication about how survey data inform change increases trust and participation. Grading reform should also operate through program-level feedback loops. Institutions can monitor score compression at upper ranges, identify inflationary drift early, and recalibrate standards before GPA signals lose meaning, preserving rigor, comparability, and the interpretive value of academic records. This matters because inflation has consequences: when grades systematically rise without corresponding gains, downstream decisions (placement, scholarships, labor-market signals) become noisier, and students may face unintended long-term effects (Denning et al., 2025). Grading, when engineered as measurement architecture, can better communicate learning with fidelity, fairness, and trust.

References

ACT. (2022). *Grade inflation continues to grow in the past decade* (ACT Research Report Series). ACT, Inc. <https://www.act.org/content/act/en/research/pdfs/R2134-Grade-Inflation-Continues-to-Grow-in-the-Past-Decade-Final-Accessible.html>

Ciftci, S. K., & Karadag, E. (2024). Grade inflation effects of capacity expansion in higher education: A longitudinal study in undergraduate teacher education programs from 2003 to 2022. *Humanities and Social Sciences Communications*, 11, Article 856. <https://doi.org/10.1057/s41599-024-03387-6>

Denning, J. T., Nesbit, R., Pope, N., & Warnick, M. (2025). *Easy A's, less pay: The long-term effects of grade inflation* (Working paper).

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, Article 57. <https://doi.org/10.1186/s41239-023-00425-2>

Huang, J., & Whipple, P. B. (2023). Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: Implications for educational assessment policy. *Humanities and Social Sciences Communications*, 10, Article 860. <https://doi.org/10.1057/s41599-023-02385-4>

Sanchez, E., & Moore, R. (2022). *Grade inflation continues to grow in the past decade* (ACT Research Report). ACT, Inc.

