GREEN BUCKET – RED BUCKET PARADIGM

AI-Infused Assessment and Evaluation

Feedback First, Grades Second

MARK RYAN

Contents

Preface

Education has long depended on systems that translate complex human learning into simplified symbols—letters, percentages, and rankings. While efficient, these systems often conceal the very processes they intend to represent. At the center of this tension is a distinction that must be clarified with precision: the difference between assessment and evaluation. Both assessment and evaluation are forms of judgment. However, they differ in kind. Assessment is qualitative judgment—it is descriptive, interpretive, and concerned with understanding learning in progress. It examines evidence, identifies patterns, and provides direction. Evaluation is quantitative judgment—it assigns value, producing scores, grades, or rankings based on that evidence. When these two forms of judgment are merged prematurely, feedback becomes constrained, and scores lose their accuracy. This book is grounded in the belief that qualitative and quantitative judgment must be intentionally separated in order to function effectively.

The Green Bucket–Red Bucket Paradigm provides a structure for separating qualitative and quantitative judgment with clarity and precision. Once a rubric has been introduced and understood by learners, that same rubric can be AI-infused and used within the Green Bucket to generate a non-graded, written qualitative assessment that highlights a learner's strengths, identifies areas for growth, and offers specific, actionable feedback. This preserves assessment as a developmental process. The same AI-infused rubric is then applied within the Red Bucket to produce a score, grade, or ranking, accompanied by quantitative explanatory comments that justify the outcome. In this way, artificial intelligence serves as a stabilizing force across both domains, increasing consistency, reducing bias, and ensuring that both qualitative and quantitative judgments remain tightly aligned to shared criteria.

This is not simply a procedural adjustment. It is a shift in how assessment and evaluation are understood. For that reason, the Green Bucket–Red Bucket framework is best described as a paradigm.

A method operates within an existing system. **A paradigm reshapes the system itself.** The prevailing paradigm in education has long combined qualitative and quantitative judgment into a single act: grading. In doing so, it has limited the role of feedback and blurred the purpose of evaluation. This book proposes an alternative structure—one that separates these forms of judgment so that each can function with clarity and precision.

The chapters that follow build this argument systematically.

Chapter 1 examines how traditional grading systems conflate qualitative and quantitative judgment, often reducing complex learning to simplified scores that obscure growth, context, and meaning. It establishes the conceptual foundation for separating assessment from evaluation.

Chapter 2 defines the Green Bucket as a structured space for qualitative assessment, where feedback is prioritized over scoring. It explains how learners benefit from extended, evidence-based commentary that supports revision and deepens understanding.

Chapter 3 turns to the Red Bucket, clarifying how quantitative evaluation functions as a distinct form of judgment. It shows how scores, when separated from formative processes, become more accurate representations of demonstrated learning.

Chapter 4 explores the role of AI-infused rubrics as a bridge between qualitative and quantitative judgment. It explains how artificial intelligence can stabilize criteria, generate aligned feedback, and maintain coherence across both buckets.

Chapter 5 analyses the many forms of bias that influence human judgment, including inconsistency, drift, and perceptual distortion. It demonstrates how AI-supported systems reduce these effects, leading to more reliable and equitable assessment and evaluation.

Chapter 6 focuses on the design of feedback within the Green Bucket, showing how qualitative judgment can be made more precise, actionable, and instructional. It emphasizes feedback as a central mechanism for learning rather than a justification for scoring.

Chapter 7 explains how qualitative evidence is translated into quantitative outcomes within the Red Bucket. It outlines systematic approaches for converting narrative assessment into defensible numerical evaluation.

Chapter 8 provides practical guidance for implementing the paradigm in classroom settings. It addresses workflow, timing, and integration, demonstrating how the dual-bucket system can function efficiently in real instructional environments.

Chapter 9 considers the broader implications of separating assessment and evaluation, particularly in relation to equity, transparency, and accountability. It argues that this paradigm produces systems that are more just and more aligned with the realities of learning.

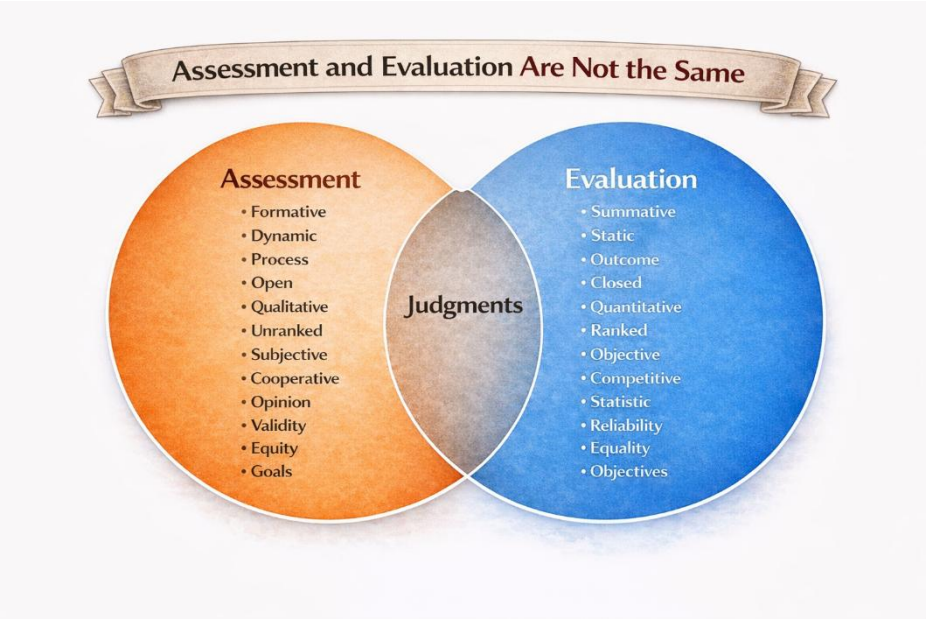
Throughout this book, the distinction between qualitative and quantitative judgment remains central. Assessment and evaluation are not interchangeable; they are complementary processes that must be designed with intention. By separating them—and by leveraging AI to support both—educators can create systems that more accurately represent learning.

This work invites a reconsideration of how judgment functions in education. When qualitative understanding is allowed to develop before quantitative value is assigned, both processes become more meaningful. The result is not simply better grading, but a more coherent and defensible approach to learning itself.

Chapter 1 Why Assessment and Evaluation Must Be Separated

For generations, educators have operated within a system that treats assessment and evaluation as interchangeable. Assignments are given, feedback is offered, and a score is attached—often within the same moment, on the same page, and through the same instrument. This practice is so normalized that it rarely invites scrutiny. Yet beneath this routine lies a fundamental problem: two distinct forms of judgment—qualitative and quantitative—have been fused into a single act. The result is not efficiency, but distortion.

At its core, assessment is an act of **inquiry**. It asks: *What is this learner understanding? Where are they developing? What remains unclear?* These questions cannot be answered with numbers. They require language, interpretation, and attention to context. Assessment is therefore inherently qualitative. It is descriptive rather than reductive, dynamic rather than fixed, and always oriented toward growth. Contemporary

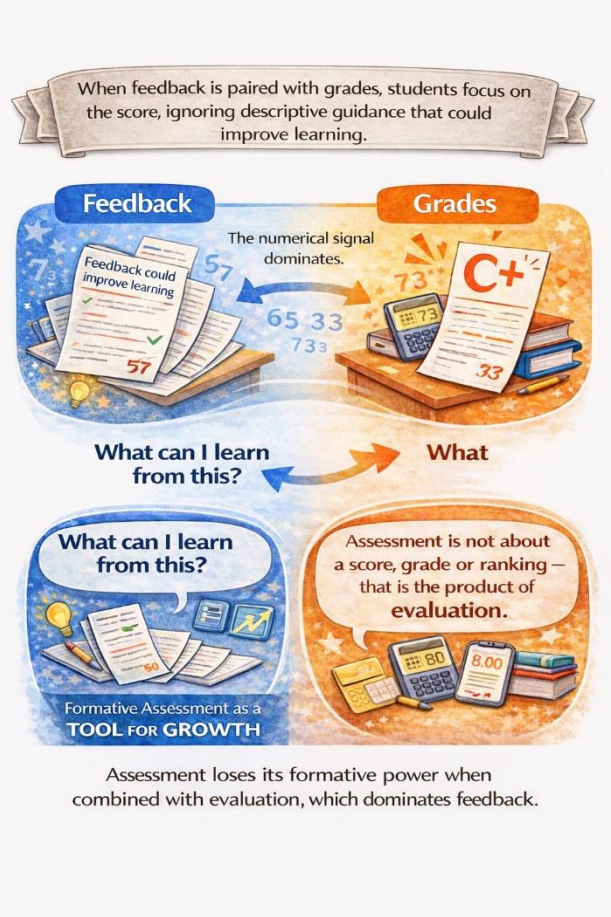


scholarship continues to affirm that formative processes—those that provide feedback during learning—are among the most powerful influences on student achievement and self-regulation. When functioning as intended, assessment expands understanding. It does not conclude it.

Evaluation, by contrast, serves a different purpose. It asks: *How does this performance compare? What level has been reached? What decision must be made?* These questions require aggregation and reduction. Evaluation transforms complex performances into scores, rankings, or categories that can be reported, compared, and acted upon. It is inherently quantitative. Where assessment opens learning, evaluation closes it. Where assessment speaks in narrative, evaluation speaks in numbers.

The difficulty arises when these two systems are collapsed into one. The common phrase “summative assessment” reflects this confusion. It suggests that assessment

itself can be summative—that it can both describe learning and simultaneously reduce it to a final value. This chapter rejects that premise. Summation is not a property of assessment; it is a function of evaluation. The act of summing requires quantification, and quantification belongs exclusively to evaluation. To call something a "summative assessment" is to commit a category error, blending two fundamentally different logics into a single term. Research in recent years has increasingly highlighted the consequences of this conflation. When feedback is paired with grades, students tend to attend to the score while disregarding the descriptive guidance that could improve their learning. The numerical signal dominates. Instead of asking *What can I learn from this?*, students ask *What did I get?* In this shift, assessment loses its formative power. It becomes subordinate to evaluation, serving merely as justification for a grade rather than as a tool for growth.



Assessment loses its formative power when combined with evaluation, which dominates feedback.

This blending also obscures the complexity of learning itself. A score compresses multiple dimensions—understanding, effort, creativity, context—into a single value. While this reduction may be necessary for institutional reporting, it inevitably strips away nuance. Studies in open-access educational research emphasize that qualitative feedback supports deeper learning by preserving this complexity, allowing students to engage with their thinking rather than merely their performance level. In contrast, purely quantitative systems risk promoting surface-level engagement, where success is defined by accumulation of points rather than development of understanding. Equity concerns further intensify the need for separation. When assessment is reduced to numbers, individual learning trajectories are flattened into standardized metrics that may not reflect diverse starting points or contexts. Recent scholarship on equitable assessment practices argues for systems that prioritize descriptive feedback and contextual understanding before any form of measurement is applied. In such models, assessment honors variation, while evaluation applies consistency. These are not competing goals, but sequential ones.

This chapter therefore establishes a foundational principle:

Assessment is qualitative, formative, and diagnostic. Evaluation is quantitative, summative, and classificatory. They are not stages of the same process. They are different systems serving different purposes.

To preserve the integrity of learning, these systems must be separated both conceptually and operationally. Assessment must remain free of numerical intrusion so that it can function as a space for exploration, feedback, and growth. Evaluation must be clearly identified as the moment of reduction, where learning is translated into a form required for reporting and decision-making. The transition from one to the other is not an overlap—it is a conversion.

This distinction is more than semantic. It reshapes how students experience learning. When assessment stands on its own, students engage with feedback as information, not as a final result. They revise, reflect, and refine their thinking without the immediate pressure of ranking. When evaluation is introduced later, it reflects a body of learning that has already been developed through qualitative insight. In this sequence, learning is first understood, then measured.

The traditional system reverses this order. It measures before understanding is complete, attaching numbers to partial learning and calling the result final. In doing so, it narrows the space in which learning can grow.

Chapter 1, then, is not merely a critique—it is a reset. It invites educators to recognize that the long-standing union of assessment and evaluation is not inevitable. It is constructed, and therefore it can be reconstructed. By separating qualitative inquiry from quantitative judgment, we restore clarity to both. Assessment regains its purpose as the language of learning. Evaluation retains its role as the mathematics of decision.

The chapters that follow will build on this foundation, showing how a dual-system approach—what this book terms the Green Bucket and the Red Bucket—can operationalize this separation in practice. Before methods can be designed, however, the paradigm must be clear:

Assessment does not sum. Evaluation must.

Annotated References

Brooks, C. B. C., Burton, R. B., Van der Kleij, F., Ablaza, C., Carroll, A., Hattie, J., & Garcia Salinas, J. (2024). *“It actually helped”: Students’ perceptions of feedback helpfulness prior to and following a teacher professional learning intervention.* Frontiers in Education.
https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1433184/full
Annotation: This article supports the claim that feedback can shape learning when

students perceive it as useful and improvement-oriented. It is especially relevant to your distinction between descriptive assessment and score-based evaluation.

Ferguson, J. H. (2024). *Ungrading in organic chemistry: Students assessing themselves and reflecting on their learning*. Frontiers in Education. https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1394042/full
Annotation: This study is highly useful for Chapter 1 because it directly addresses the problem of grade focus displacing learning focus. It supports the argument that numbers can redirect attention away from qualitative understanding and reflective growth.

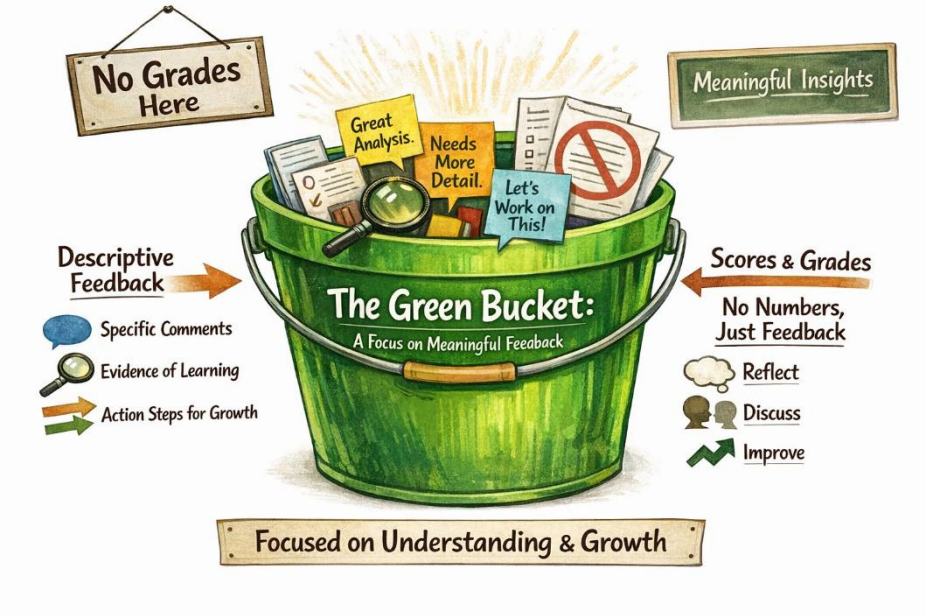
Levy-Feldman, I. (2025). *The role of assessment in improving education and promoting educational equity.* Education Sciences, 15(2), 224. https://www.mdpi.com/2227-7102/15/2/224
Annotation: This open-access article emphasizes that assessment shapes teaching, learning, and equity. It is useful for grounding your chapter's concern that oversimplified grading systems can flatten learner complexity and mask important contextual differences. (MDPI)

Parmigiani, D., Nicchia, E., Murgia, E., & Ingersoll, M. (2024). *Formative assessment in higher education: An exploratory study within programs for professionals in education*. Frontiers in Education. https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1366215/full
Annotation: This article helps establish assessment as a feedback-rich, learning-centered process. It is especially relevant because it also raises the question of whether formative assessment should be linked to grading scales, which directly intersects with the conceptual separation of assessment and evaluation. (Frontiers)

Rabbani, L. M., et al. (2024). *Fostering student engagement with criticism feedback: Importance, contrasting perspectives and key provisions.* Frontiers in Education. https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1344997/full
Annotation: This article is useful for the Green Bucket framework because it treats criticism and feedback as information that advances learning rather than as final judgment. It reinforces the idea that assessment is interpretive, dialogic, and growth-oriented.

Chapter 2 The Green Bucket

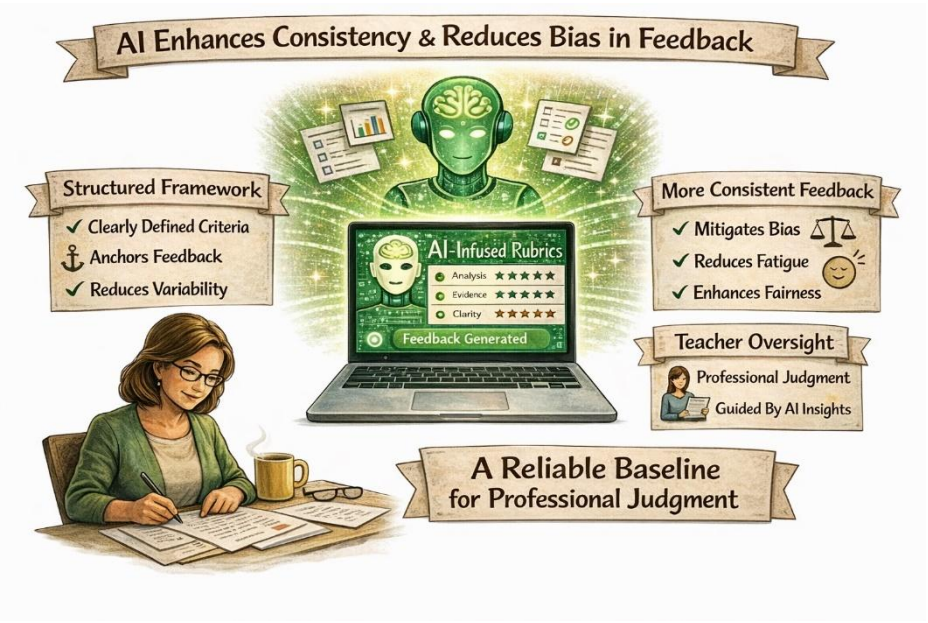
The Green Bucket is a deliberately constructed space in which qualitative judgment is separated from quantitative scoring and given full instructional priority. In this paradigm, assessment is understood as a descriptive act rather than a numerical one. The Green Bucket is not simply a metaphor; it is an operational environment where evidence of learning is examined, interpreted, and communicated through language that is specific, actionable, and oriented toward growth. By design, it withholds grades so that attention remains fixed on feedback. This separation is essential, as decades of classroom practice have shown that when scores are present, they tend to dominate student attention, often eclipsing the very comments intended to guide improvement.



Within the Green Bucket, artificial intelligence serves as a stabilizing and amplifying mechanism for qualitative assessment. Once a rubric has been introduced, clarified, and internalized by learners, that same rubric can be AI-infused to generate consistent, criteria-aligned commentary. The AI does not replace the teacher's professional judgment; rather, it extends it by ensuring that each piece of student work receives thorough, standards-based feedback. This feedback can highlight strengths, identify areas for growth, and suggest next steps with a level of detail that would be difficult to sustain across large numbers of students. In this way, AI supports scalability without sacrificing the integrity of qualitative judgment.

The defining feature of the Green Bucket is its commitment to extended, evidence-based commentary. Feedback is not reduced to brief annotations or general praise. Instead, it is anchored in observable aspects of student work. For example, rather than stating "good job," Green Bucket feedback might note that a student's argument is strengthened by the use of specific textual evidence, while also pointing out where reasoning could be further elaborated. This kind of commentary requires attention to evidence, alignment with criteria, and a clear articulation of what improvement looks like. It transforms feedback from a reaction into an instructional tool. Learners benefit

from this approach in several important ways. First, the absence of grades reduces performance anxiety and redirects cognitive energy toward understanding. When students are not immediately confronted with a score; they are more likely to engage with the feedback itself. They begin to ask questions such as “What does this mean?” and “How can I improve?” rather than “What did I get?” This shift in focus is subtle but significant. It repositions learning as a process rather than an outcome, encouraging students to see their work as iterative and revisable.



Second, evidence-based commentary supports revision as a central component of learning. The Green Bucket assumes that initial attempts are not final products but starting points for development. Feedback, therefore, is constructed with revision in mind. It identifies not only what is present in the work but also what could be added, clarified, or reorganized. In doing so, it provides a pathway forward. Students are not left to guess how to improve; they are given concrete guidance that can be acted upon. Over time, this fosters a habit of revision that deepens understanding and strengthens performance.

Third, the Green Bucket contributes to the development of metacognitive awareness. As students interact with detailed feedback, they begin to recognize patterns in their own work. They learn to anticipate criteria, to assess their own progress, and to make adjustments before submitting future assignments. This self-regulatory capacity is a critical outcome of qualitative assessment. It moves learners toward independence, as they become less reliant on external valuations and more capable of internal judgment.

The role of AI in this process is particularly significant in maintaining consistency and reducing variability in feedback. Human judgment, while essential, can be influenced by time constraints, fatigue, and implicit bias. AI-infused rubrics provide a structured framework that helps mitigate these factors. By anchoring feedback in clearly defined criteria, AI supports a more uniform application of standards across students and

assignments. This does not eliminate the need for teacher oversight; rather, it enhances it by providing a reliable baseline from which professional judgment can operate.

Importantly, the Green Bucket does not exist in isolation. It is part of a larger system in which qualitative assessment and quantitative evaluation are intentionally separated but conceptually linked. The same rubric that guides feedback in the Green Bucket can later be used in the Red Bucket to produce a score or grade. However, the sequencing matters. By encountering qualitative feedback first, students have the opportunity to understand and act on it before any numerical judgment is introduced. This preserves the instructional value of assessment and prevents it from being overshadowed by evaluation.

The implementation of the Green Bucket requires careful attention to classroom structures and expectations. Time must be allocated for students to read, interpret, and respond to feedback. Opportunities for revision must be built into the learning process. Teachers must model how to use feedback effectively, demonstrating how comments can be translated into concrete changes in student work. AI tools must be selected and configured in ways that align with instructional goals and ethical considerations. These elements are not peripheral; they are central to the functioning of the Green Bucket as a meaningful space for learning.

In practice, the Green Bucket transforms the relationship between teacher, student, and assessment. The teacher becomes a provider of insight rather than a distributor of scores. The student becomes an active participant in a cycle of feedback and revision. Assessment becomes a dialogue rather than a declaration. This shift has profound implications for how learning is experienced and understood. It emphasizes growth over ranking, process over product, and understanding over performance.

Ultimately, the Green Bucket functions as a necessary response to the limitations of traditional grading practices. By centering qualitative, AI-infused assessment, it creates a space where feedback can fulfill its intended purpose: to inform, to guide, and to support learning. In doing so, it lays the groundwork for a more coherent and humane approach to assessment—one that recognizes the complexity of learning and honors the role of language in making that complexity visible.

Annotated References

Carless, D., & Winstone, N. E. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education, 28*(1), 150–163. https://doi.org/10.1080/13562517.2020.1782372
Annotation: This article explains how both teachers and students must develop “feedback literacy” to effectively engage with qualitative commentary. It directly supports the Green Bucket model by reinforcing that feedback must be understood, interpreted, and acted upon—not merely delivered.

Holmes, W., Bialik, M., & Fadel, C. (2023). *Artificial intelligence in education: Promises and implications for teaching and learning* (Updated ed.). Center for Curriculum Redesign.
https://curriculumredesign.org/our-work/artificial-intelligence-in-education/
Annotation: This resource examines how AI can enhance teaching and learning processes, including assessment. It aligns with the Green Bucket by highlighting AI's role in generating consistent, criteria-based feedback while maintaining human oversight, supporting scalability without compromising instructional integrity.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2024). *Intelligence unleashed: An argument for AI in education* (Updated open edition). Pearson / UCL Knowledge Lab.
https://discovery.ucl.ac.uk/id/eprint/10020686/
Annotation: This updated open-access work explores how AI can augment—not replace—teacher judgment. It reinforces the chapter's claim that AI-infused rubrics can stabilize qualitative assessment and reduce inconsistency while preserving the teacher's central role.

Chapter 3 The Red Bucket

The Red Bucket, is a conceptual and operational space in which quantitative evaluation is isolated, clarified, and deliberately structured as a distinct form of judgment. Where the Green Bucket centers on qualitative description—language that interprets, guides, and supports learning—the Red Bucket is concerned with measurement. Its purpose is not to shape learning in the moment, but to represent what learning has been demonstrated at a particular point in time. This distinction is essential. When qualitative and quantitative judgments are blended, the numerical signal tends to dominate, often distorting both the meaning of the score and the value of the feedback. By separating these functions, the Red Bucket restores clarity to evaluation.

In the Red Bucket, scores are understood as symbolic representations of performance against defined criteria. They are not neutral; they are constructed through systems of weighting, categorization, and aggregation. For this reason, their validity depends entirely on the integrity of the rubric that produces them. A score does not speak for itself. It reflects the structure of what has been counted, how it has been weighted, and whether those elements align with the intended learning objectives. The Red Bucket makes this structure visible and explicit, requiring educators to define, in advance, what counts as evidence and how that evidence will be translated into a numerical form.

One of the central claims of this chapter is that scores become more accurate representations of demonstrated learning when they are separated from formative processes. During formative assessment, students are still in the process of developing understanding. Errors, revisions, and partial attempts are part of the learning trajectory. Assigning grades during this phase risks capturing moments of incompleteness rather than evidence of attainment. In contrast, the Red Bucket is activated after opportunities for revision, feedback, and clarification have been provided. At this stage, the work submitted represents a more stable performance, allowing evaluation to function as a summary judgment rather than an interruption of learning.

This temporal separation enhances both accuracy and fairness. Students are not penalized for early misunderstandings, and educators are not forced to translate evolving work into premature scores. Instead, the Red Bucket captures a closer approximation of what students can do when given sufficient time, guidance, and opportunity to refine their thinking. The resulting scores are therefore less about compliance or timing and more about demonstrated capability. In this way, the Red Bucket aligns evaluation with the principle that learning is iterative, while still fulfilling institutional requirements for grading and reporting.

A key mechanism within the Red Bucket is the design of rubrics that enumerate and count specific objectives and requirements. Unlike qualitative rubrics, which describe levels of performance in narrative terms, Red Bucket rubrics must operationalize criteria in ways that can be quantified. This does not mean reducing complex thinking to simplistic checklists but rather identifying observable indicators that can be consistently counted across student work. These indicators often include structural and formal elements of an assignment, such as word count, number of references, inclusion of examples, and adherence to formatting conventions like APA style.

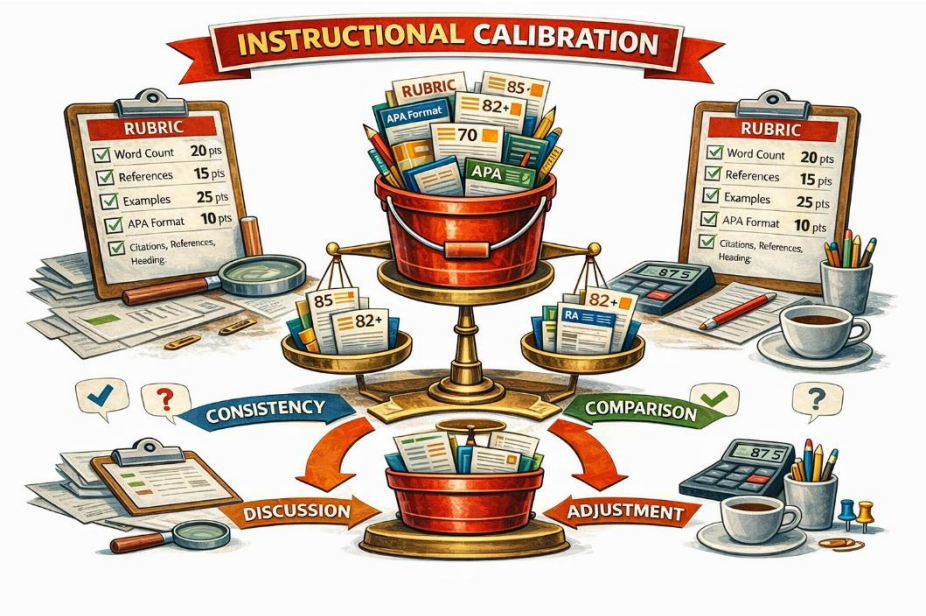
For instance, a research-based assignment may require a minimum of 1,500 words, at least five scholarly references, integration of two theoretical frameworks, and correct use of APA citation and formatting. Each of these elements can be clearly defined, observed, and counted. The rubric assigns point values to each requirement, creating a transparent system in which students understand how their work will be evaluated. This clarity serves two purposes. First, it reduces ambiguity in grading, supporting consistency across different evaluators or across time. Second, it communicates expectations to students in a concrete form, allowing them to align their work with the stated criteria.

The emphasis on enumerated requirements does not diminish the importance of higher-order thinking; rather, it ensures that foundational elements are accounted for in a systematic way. When these elements are clearly specified, they can be evaluated with greater reliability, freeing qualitative judgment to operate where it is most needed—in the interpretation of ideas, arguments, and originality. In the Red Bucket, however, the focus remains on what can be consistently measured. This includes not only the presence of required components, but also their accuracy and completeness.

Another important feature of Red Bucket evaluation is weighting. Not all criteria carry equal importance, and the rubric must reflect these distinctions. For example, adherence to APA format may be assigned a smaller percentage of the total score than the quality of evidence or the coherence of argumentation. By assigning weights, the rubric communicates priorities and ensures that the final score reflects the relative significance of each component. This prevents minor technical errors from disproportionately affecting the overall evaluation, while still holding students

accountable for meeting established standards.

The Red Bucket also supports transparency in grading practices. Because the criteria are explicitly defined and numerically structured, students can see how their scores are derived. This visibility reduces perceptions of arbitrariness and builds trust in the evaluation process. When students understand how points are allocated, they are better positioned to interpret their results and to identify areas for improvement in future work. Although the Red Bucket itself does not provide formative feedback, its clarity contributes indirectly to learning by making expectations and outcomes more legible.



It is important to note that the Red Bucket does not eliminate professional judgment. Decisions about which criteria to include, how to define them, and how to weight them remain deeply pedagogical. What changes is the form that judgment takes. Instead of being embedded in narrative commentary, it is encoded in the structure of the rubric. This shift allows evaluation to function with greater consistency, particularly in contexts where multiple instructors or sections are involved. It also creates opportunities for calibration, as educators can compare how criteria are applied and adjust their interpretations accordingly.

In sum, the Red Bucket is a necessary complement to the Green Bucket. By isolating quantitative evaluation, it clarifies the role of scores as representations of demonstrated learning rather than drivers of it. Through carefully designed rubrics that enumerate and weight specific requirements, the Red Bucket produces evaluations that are more accurate, transparent, and aligned with instructional goals. This separation does not fragment the assessment process; it organizes it. Each bucket serves a distinct purpose, and together they create a system in which both feedback and evaluation can function with integrity.

Annotated References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2024). *Standards for educational and psychological testing (updated guidance)*. https://www.testingstandards.net/
Annotation: These standards provide a comprehensive framework for validity, reliability, and fairness in educational measurement. They support the Red Bucket's emphasis on the idea that scores are constructed representations dependent on clearly defined criteria and weighting systems.

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education, 9*, 1399377. https://www.frontiersin.org/articles/10.3389/feduc.2024.1399377/full

This open-access article focuses on validity in assessment design, emphasizing that scores are only meaningful when grounded in clearly defined constructs and criteria. It strongly supports the chapter's claim that scores are symbolic representations dependent on rubric integrity and alignment with learning objectives.

UNESCO. (2023). *Guidance for generative AI in education and research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

This policy document emphasizes transparency, fairness, and clarity in evaluation systems. It supports the Red Bucket's focus on making criteria explicit and visible so that students understand how scores are derived.

European Commission. (2024). *Ethics guidelines for trustworthy AI in education*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Annotation: These guidelines emphasize fairness, accountability, and transparency in algorithmic systems. They support the Red Bucket's structured and equitable approach to quantitative evaluation.

Chapter 4 Qualitative Assessment to Quantitative Evaluation

Time to investigate the space between the Green Bucket and the Red Bucket, examining how AI-infused rubrics function as a bridge between qualitative interpretation and quantitative evaluation. Where earlier chapters established the necessity of separating descriptive feedback from numerical scoring, this chapter considers how coherence can be restored across these domains without collapsing their distinct purposes. The central claim is that artificial intelligence, when anchored in well-constructed rubrics, can carry the integrity of criteria across both forms of judgment, ensuring that what is described qualitatively is the same as what is counted quantitatively.

At the core of this bridging function is the rubric itself. A rubric is not simply a scoring guide; it is a structured representation of what counts as evidence of learning. In the Green Bucket, the rubric operates as a language tool. It provides the vocabulary through which feedback is expressed—terms that describe levels of clarity, depth, organization, or conceptual understanding. In the Red Bucket, the same rubric is translated into measurable indicators. Criteria become countable elements, weighted categories, or discrete thresholds. The challenge historically has been that these two uses of the rubric often drift apart. What a teacher says in feedback does not always align with what is ultimately scored.

AI-infused rubrics address this problem by stabilizing the relationship between description and measurement. Once a rubric is clearly defined, artificial intelligence can be trained or configured to apply that rubric consistently across student work. In the qualitative domain, AI generates feedback that is explicitly tied to rubric criteria. Rather than offering generalized praise or vague suggestions, it produces commentary that references specific dimensions of performance. This creates a tight coupling between the rubric and the feedback, ensuring that students receive information that is both actionable and directly relevant to the expectations of the task.

At the same time, the AI can map these same criteria onto quantitative structures. Because the rubric has already been operationalized, the transition from description to

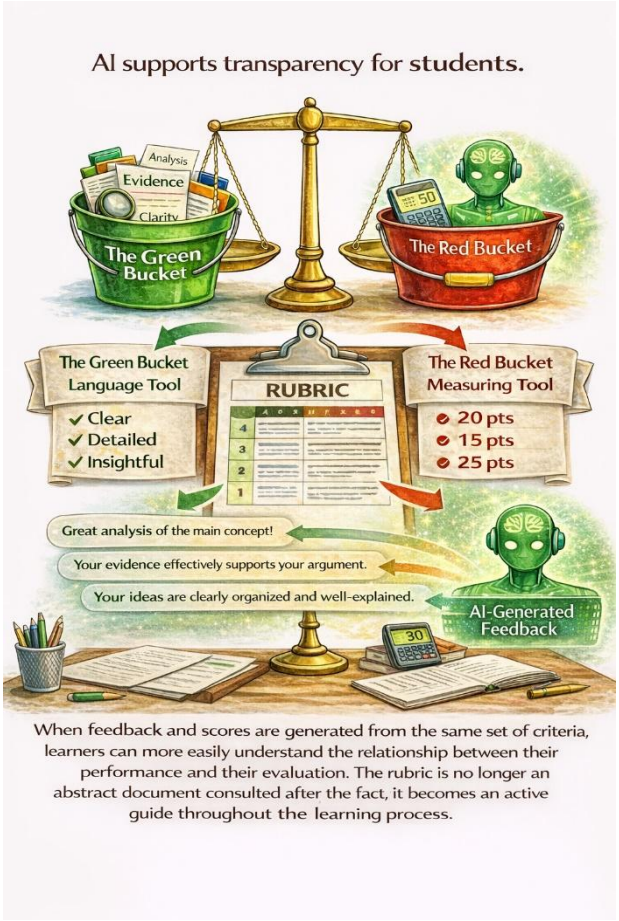
measurement does not require reinterpretation. The system can count the presence of required elements, evaluate adherence to defined standards, and assign values based on predetermined weights. The key advantage here is coherence. The criteria that generate feedback are the same criteria that generate scores. This eliminates a long-standing source of confusion for learners, who often experience a disconnect between what they are told and how they are graded.

Another important function of AI in this context is the reduction of variability. Human judgment, while indispensable, is inherently influenced by context. Time constraints, cognitive load, and implicit bias can all affect how criteria are interpreted and applied. AI-infused rubrics provide a stabilizing layer that mitigates these fluctuations. By consistently referencing the same criteria in the same way, the system creates a baseline of reliability. This does not replace professional judgment; rather, it allows educators to operate from a more stable foundation. Teachers can review, adjust, and refine AI-generated outputs, but they do so within a framework that has already ensured alignment with the rubric.

The bridging role of AI also extends to scalability. In environments where multiple instructors are responsible for assessing similar assignments, maintaining consistency becomes increasingly difficult. AI-infused rubrics offer a mechanism for calibration across sections and instructors. Because the system applies criteria uniformly, it creates a shared standard that can be examined and refined collectively. Discrepancies can be identified not as individual inconsistencies, but as opportunities to revisit the rubric itself. In this way, the rubric becomes a living document, continuously improved through interaction with both human and machine judgment.

Equally significant is the way AI supports transparency for students. When feedback and scores are generated from the same set of criteria, learners can more easily understand the relationship between their performance and their evaluation. The rubric is no longer an abstract document consulted after the fact; it becomes an active guide throughout the learning process. Students can see how qualitative feedback points directly to the elements that will later be measured. This clarity has important implications for motivation and self-regulation. When expectations are stable and visible, learners are better positioned to direct their efforts and monitor their own progress.

It is important to note that the effectiveness of AI-infused rubrics depends entirely on the quality of the rubric design. Artificial intelligence does not create coherence on its own; it amplifies whatever structure it is given. If criteria are vague, misaligned, or poorly defined, the system will reproduce those weaknesses at scale. For this reason, the development of the rubric remains a deeply pedagogical act. Educators must carefully determine which aspects of learning are best addressed qualitatively and which can be meaningfully quantified. They must also ensure that the language of the rubric can be translated into measurable indicators without losing its conceptual integrity.



Within this framework, the bridge between the Green and Red Buckets is not a merging of functions, but a synchronization of them. Qualitative judgment retains its role in guiding learning through descriptive feedback, while quantitative judgment continues to represent performance through scores. AI-infused rubrics ensure that these two forms of judgment are aligned at the level of criteria. The same expectations govern both processes, even as the forms of expression differ.

This positions artificial intelligence not as a replacement for teacher expertise, but as an instrument for coherence. By stabilizing criteria, generating aligned feedback, and maintaining consistency across contexts, AI-infused rubrics make it possible to preserve the integrity of both qualitative and quantitative judgment while eliminating the disconnect that has long separated them. In doing so, they transform the assessment system from a set of parallel practices into an integrated structure, where description and measurement operate in concert, each reinforcing the clarity and purpose of the other.

Annotated Bibliographies

Bin Dahmash, N. F. (2025). The analytic use of rubrics in writing classes by language students in an EFL context: Students' writing model and benefits. Frontiers in Education, 10, 1588046. https://doi.org/10.3389/feduc.2025.1588046

This open-access study is useful for Chapter 4 because it shows how analytic rubrics clarify requirements, help students identify strengths and weaknesses, and support longer-term gains such as self-efficacy and task management. It supports the discussion of the rubric as an active guide rather than a document consulted only after grading.

Fajardo-Ramos, D. C., Chiappe, A., & Mella-Norambuena, J. (2025). Human-in-the-loop assessment with AI: Implications for teacher education in Ibero-American universities. Frontiers in Education, 10, 1710992. https://doi.org/10.3389/feduc.2025.1710992

This scoping review fits the chapter's emphasis on AI as a stabilizing layer rather than a replacement for teacher judgment. It is especially relevant to the argument that educators still refine, review, and interpret AI-supported outputs within a coherent rubric framework.

Hochstetter-Diez, J., Negrier-Seguel, M., Diéguez-Rebolledo, M., Candia-Garrido, E., & Vidal, E. (2025). From mapping to action: SmartRubrics, an AI tool for competency-based assessment in engineering education. Sustainability, 17(13), 6098. https://doi.org/10.3390/su17136098

This open-access article directly supports the chapter's claim that AI can help construct and standardize rubric-based assessment. It is especially helpful for the sections on scalability, calibration, and the translation of criteria into more systematic assessment structures.

Holmes, W., Bialik, M., & Fadel, C. (2024). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign. https://curriculumredesign.org/our-work/artificial-intelligence-in-education/
This work highlights how AI can connect qualitative insights with quantitative measures through structured systems like rubrics. It aligns with the chapter's focus on bridging descriptive and numerical evaluation.

Ilieva, G., Yankova, T., Ruseva, M., & Kabaivanov, S. (2025). A framework for generative AI-driven assessment in higher education. Information, 16(6), 472. https://doi.org/10.3390/info16060472

This source is well aligned with the chapter's bridge concept because it focuses on framework design for AI-driven assessment in higher education. It is a strong reference for the claims about coherence, structured criteria, and the need to design AI assessment systems around clear pedagogical purposes.

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. Frontiers in Education, 9, 1399377. https://doi.org/10.3389/feduc.2024.1399377

This open-access article is one of the best matches for the chapter's concern with validity. It supports the point that AI does not create coherence by itself; rather, the validity of AI-supported assessment depends on the quality of the constructs, criteria, and inferences built into the rubric and assessment design.

Lo, J., Wong, C., Ng, A., Wong, P., Cheung, D., & Lai, P. (2026). Stretching AI's reach: Assessing an AI-driven feedback system for extended academic writing. Computers and Education: Artificial Intelligence, 10, 100511. https://doi.org/10.1016/j.caeai.2025.100511

This open-access study is especially valuable for your chapter because it examines a rubric-based AI feedback system for extended academic writing. It supports the discussion of both the promise and the limits of AI-generated feedback: coverage and efficiency improved, but generic comments remained a concern, which reinforces your point that strong rubric design and teacher oversight still matter.

Chapter 5 Bias Mitigation

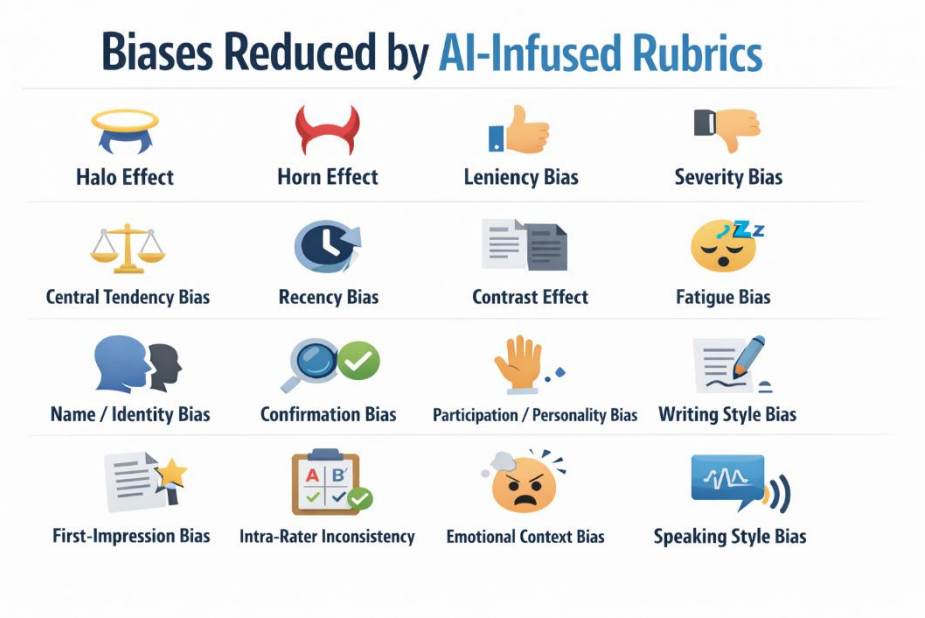
This chapter deepens the examination of judgment by analyzing the many biases—often subtle and unintentional—that influence evaluation, while highlighting how AI can mitigate their impact. While educators often strive for fairness, human cognition introduces variability that is both predictable and difficult to control. This chapter organizes these influences into three overarching dynamics: inconsistency, drift, and perceptual distortion, while also identifying specific bias forms that operate within and across these categories. Understanding these patterns is essential for building assessment systems that are both reliable and equitable.

Inconsistency refers to the uneven application of criteria, even by the same evaluator. A central example is *Intra-Rater Inconsistency* – the same evaluator applying criteria unevenly over time. This may occur within a single grading session or across multiple days. Closely related is *Emotional Context Bias* – personal emotions affecting judgment at the time of grading, where frustration, stress, or even positive mood can influence scoring. *Fatigue Bias* – declining accuracy or consistency due to evaluator tiredness further compounds this issue, especially in high-volume grading contexts. Together, these forms of inconsistency reveal that human judgment is not fixed; it fluctuates based on internal states as much as external criteria.

Drift represents a gradual shift in evaluative standards over time. While inconsistency captures moment-to-moment variation, drift reflects a directional change. For example, an evaluator may begin with strict standards and unconsciously relax them, or the reverse. Several biases contribute to this phenomenon. *Leniency Bias* – consistently grading more generously than warranted and *Severity Bias* – consistently grading more harshly than warranted illustrate how evaluators may settle into habitual patterns that diverge from intended standards. Similarly, *Central Tendency Bias* – avoiding extreme scores and clustering ratings around the middle reflects a drift toward safety, where evaluators hesitate to assign very high or very low marks. These patterns often emerge without awareness, yet they systematically alter outcomes.

Perceptual distortion encompasses biases that arise when irrelevant or extraneous factors influence judgment. One of the most well-documented is the *Halo Effect* – allowing one positive trait (e.g., neatness) to overly influence the overall evaluation. Its counterpart, the *Horn Effect* – letting one negative trait disproportionately lower the overall judgment, demonstrates how a single feature can dominate perception. These biases highlight the brain's tendency to generalize from limited information, creating coherence at the expense of accuracy.

Other forms of perceptual distortion operate through comparison and context. The *Contrast Effect* – evaluating a student relative to others rather than against criteria



illustrates how judgments shift depending on what has been recently observed. A strong performance may appear weaker if it follows an exceptional one, and stronger if it follows a poor one. Similarly, *Recency Bias* – overweighting the most recent performance instead of the full body of work skews evaluation toward what is most immediately available in memory, rather than what is most representative.

Social and identity-related factors also play a significant role. *Name / Identity Bias* – judgments influenced by a student's name, background, or identity markers reveals how implicit associations can shape expectations and interpretations. *Participation / Personality Bias* – letting likability or participation level influence academic evaluation further demonstrates how non-academic traits can seep into academic judgments. These biases are particularly concerning because they can reinforce existing inequities, often without the evaluator's awareness.

Cognitive expectations introduce another layer of distortion. *Confirmation Bias* – interpreting work in ways that confirm prior expectations or beliefs leads evaluators to see what they anticipate, rather than what is objectively present. Once an impression is formed, it can be difficult to dislodge. This is closely tied to *First-Impression Bias* – early judgments shaping all subsequent evaluations, where initial perceptions anchor later interpretations. Together, these biases create a self-reinforcing loop, making it challenging to reassess work with fresh objectivity.

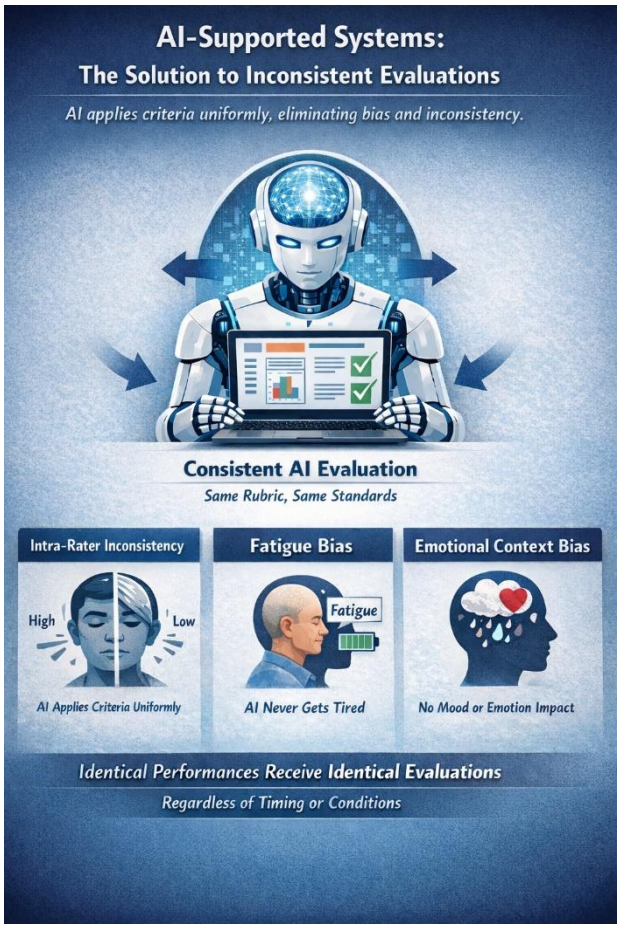
Finally, stylistic and presentation factors can overshadow substantive understanding. *Writing Style Bias* – favoring certain writing styles over actual content mastery may advantage students who align with preferred conventions, regardless of conceptual depth. Similarly, *Speaking Style Bias* – judging based on delivery or accent rather than substance can influence oral assessments, privileging fluency or familiarity over clarity of thought. These biases underscore the importance of distinguishing between form and content in evaluation.

When considered collectively, these biases illustrate that human judgment is not merely imperfect but systematically influenced by identifiable patterns. Even with well-designed rubrics and professional training, these effects persist because they are rooted in fundamental cognitive processes. This raises a critical challenge: how can assessment systems achieve consistency and fairness when the human element is inherently variable?

AI-supported systems offer a compelling response to this challenge. By design, AI applies criteria uniformly, eliminating many sources of inconsistency. Once a rubric is encoded, the system evaluates each submission using the same parameters, unaffected by fatigue, mood, or emotional context. This directly addresses Intra-Rater Inconsistency, Fatigue Bias, and Emotional Context Bias, ensuring that identical performances receive identical evaluations regardless of timing or conditions.

AI also stabilizes evaluation by preventing drift. Unlike human evaluators, AI does not gradually shift its standards. It does not become more lenient, more severe, or more centered over time. As a result, biases such as Leniency Bias, Severity Bias, and Central Tendency Bias are significantly reduced. Each evaluation is anchored to a fixed standard, preserving fairness across all submissions.

Perceptual distortions are similarly mitigated. AI systems can be designed to focus exclusively on relevant features, minimizing the influence of extraneous factors. This reduces the impact of the Halo Effect, Horn Effect, and Contrast Effect, as well as biases related to writing or speaking style. Moreover, when identity markers are removed or anonymized, the influence of Name / Identity Bias and related social biases can be substantially diminished.



Importantly, AI systems are not free from bias; however, their biases differ in nature. Because AI operates through explicit models and data, its biases are systematic and therefore detectable. This allows for ongoing calibration and refinement. In contrast, human biases are often inconsistent and opaque, making them more difficult to identify

and correct. AI introduces the possibility of continuous improvement through analysis, rather than reliance on individual awareness alone.

This chapter emphasizes that the goal is not to replace human judgment but to strengthen it. In a hybrid model, AI provides a consistent baseline evaluation, while human educators review, interpret, and contextualize results. This approach preserves professional expertise while reducing the influence of cognitive bias. It also allows educators to focus more on meaningful feedback and instructional design, rather than the mechanics of scoring.

In sum, Chapter 5 reframes bias as an inherent feature of human cognition rather than a failure of intention. By identifying specific forms—ranging from inconsistency and drift to a wide array of perceptual distortions—it becomes possible to design systems that address these limitations directly. AI-supported evaluation offers a path toward greater reliability and equity, creating conditions in which assessment more accurately reflects student learning.

Annotated References

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. (2023). *Automatic assessment of text-based responses in post-secondary education: A systematic review*. arXiv. https://arxiv.org/abs/2308.16151

Annotation:
Large-scale systematic review (93 studies) on AI-based assessment systems. Demonstrates how human grading is labor-intensive and variable, while AI systems improve consistency and scalability, directly supporting the inconsistency and drift framework.

Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., Wuttig, A., Sommerfeldt, H., & Schütt, S. (2024). *Beyond human subjectivity and error: A novel AI grading system*. arXiv. https://arxiv.org/abs/2405.04323

Annotation:
Empirical study comparing AI grading with human experts. Finds AI grading had 44% lower deviation from benchmark scores than humans, demonstrating reduced subjectivity and stronger reliability—direct evidence for the claim that AI mitigates intra-rater inconsistency and fatigue bias.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2023). *Inherent trade-offs in the fair determination of risk scores*. arXiv. https://arxiv.org/abs/1609.05807

Annotation:
Seminal open-access work on algorithmic fairness. Shows that bias in AI is systematic and mathematically analyzable, supporting that argument that AI bias can be detected and calibrated, unlike opaque human bias.

Li, M., Enkhtur, A., Yamamoto, B. A., Cheng, F., & Chen, L. (2023). *Potential societal biases of ChatGPT in higher education: A scoping review*. arXiv. https://arxiv.org/abs/2311.14381
Annotation:
Comprehensive review of bias in generative AI within higher education. Identifies risks of identity bias and embedded societal bias, reinforcing the point that AI is not bias-free but differs in being traceable and correctable.

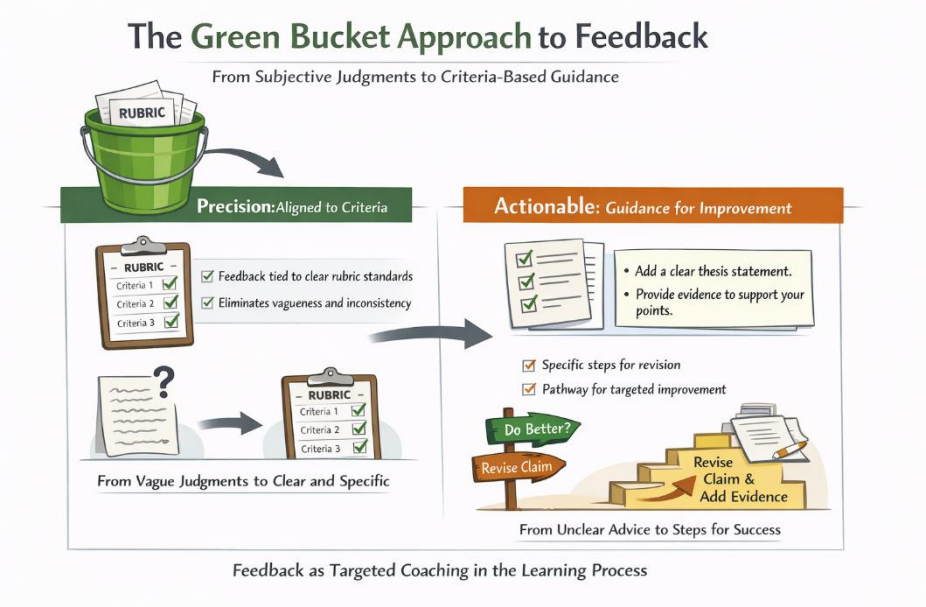
Mok, R., Akhtar, F., Clare, L., Li, C., Ida, J., Ross, L., & Campanelli, M. (2024). *Using AI large language models for grading in education: A hands-on test for physics*. arXiv. https://arxiv.org/abs/2411.13685
Annotation:
Experimental evaluation of LLM grading versus human grading. Finds AI approaches human-level grading when guided by rubrics, supporting the claim that structured criteria (rubrics) are key to reducing perceptual distortion and inconsistency.

Chapter 6 Designing Feedback in the Green Bucket

Feedback design within the AI-infused Green Bucket is the central mechanism through which learning advances. In this model, there is no score—the rubric-aligned commentary, generated and structured through AI, functions as the instruction itself rather than a justification for evaluation. The Green Bucket becomes a protected instructional space where qualitative judgment is translated into precise, immediately usable guidance. By separating feedback from grading, and by stabilizing it through AI-supported systems, the distortions introduced by scores and human variability are removed, restoring clarity, consistency, and instructional focus.

At its core, the AI-infused Green Bucket transforms subjective judgment into disciplined, criterion-based communication at scale. Educators inherently make evaluative judgments, but without structure, those judgments can be uneven, vague, or inconsistent across learners and contexts. AI imposes precision by enforcing strict alignment between feedback and a transparent rubric. Each comment is generated or validated against defined criteria, ensuring that observations are directly tied to expectations. This eliminates ambiguity and allows learners to see exactly what needs improvement and why, with consistency that does not fluctuate across time, fatigue, or context.

Precision alone, however, is insufficient. For feedback to function as instruction, it must also be actionable. AI-infused Green Bucket feedback is designed to move beyond identification of strengths and weaknesses toward explicit direction for revision. Rather than stating that an argument is “unclear,” the system specifies what is missing—such as a defined claim, relevant evidence, or logical sequencing—and indicates how to address it. This transforms feedback into targeted coaching, embedded directly within the learning process. The learner is not left to interpret general impressions but is given a clear pathway for improvement grounded in rubric-defined performance.

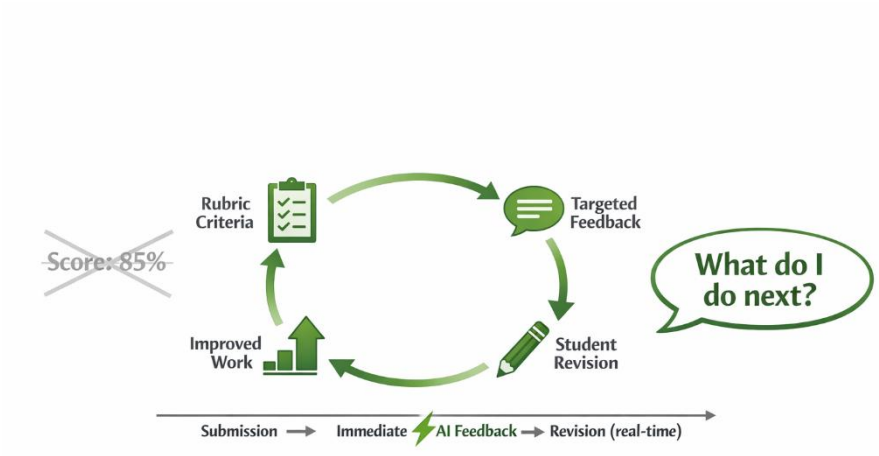


Conciseness remains a defining constraint, particularly in digitally mediated environments where cognitive load must be managed carefully. AI systems enforce disciplined brevity, often

constraining feedback to a focused format of approximately 150 words. This limitation is not reductive; it is clarifying. It ensures that feedback prioritizes the most instructionally significant elements and communicates them efficiently. Overextended commentary can dilute key signals, whereas concise, structured feedback sharpens attention and increases the likelihood of engagement and revision.

The absence of a score is a deliberate and necessary feature of the AI-infused Green Bucket. When numerical values are present, they tend to dominate learner attention, often displacing the feedback itself. By removing the score entirely, the system redirects attention to improvement. The learner's orientation shifts from "What did I receive?" to "What do I do next?" AI reinforces this shift by consistently delivering guidance without attaching evaluative labels, ensuring that every interaction remains forward-looking and instructionally focused.

Immediacy is significantly amplified through AI integration. Feedback is delivered at or near the point of submission, eliminating the delays that traditionally separate performance from response. This timing is critical: feedback intersects with the learner's active cognitive state, allowing misconceptions to be corrected before they stabilize and effective strategies to be reinforced while still in use. AI enables this immediacy at scale, transforming feedback from retrospective commentary into a real-time instructional intervention.



From the learner's perspective, the result is a system of continuous, reliable guidance. AI applies the rubric uniformly, ensuring that feedback is stable, predictable, and directly tied to expectations. Variability introduced by human factors—such as fatigue, mood, or inconsistency—is removed. This consistency builds trust in the process and reduces uncertainty about how to improve. Learners engage in a continuous cycle of attempt, feedback, and revision, with each iteration informed by precise, criterion-based input delivered at the moment it is most useful.

Iteration is fundamental to the Green Bucket model. Feedback is not a terminal event but part of a recursive process. Learners receive guidance, revise their work, and resubmit—often multiple times. AI supports this cycle by maintaining alignment to the rubric across iterations, tracking changes in performance, and reinforcing progress toward criteria. This structured repetition mirrors the development of expertise, where improvement emerges through deliberate practice informed by timely, specific feedback.

The design of feedback also requires disciplined attention to clarity and tone. While AI ensures structural alignment and consistency, the language of feedback must remain accessible and direct. Feedback avoids unnecessary complexity while maintaining academic rigor, ensuring that learners can interpret and act on it without confusion. The result is communication that is both precise and usable, reinforcing its function as instruction rather than evaluation.

Importantly, the AI-infused Green Bucket relocates the locus of quality. In traditional systems, quality is often determined at the point of grading. Here, quality is developed within the feedback cycle itself. By the time work transitions to the evaluative phase, it has already been shaped through multiple rounds of targeted, rubric-aligned guidance. Assessment becomes confirmatory rather than diagnostic—it verifies competence that has already been developed through structured iteration.

In online and asynchronous environments, where direct interaction may be limited, this design becomes even more consequential. The AI-infused Green Bucket replaces physical proximity with continuous, structured instructional response. Every submission receives immediate, meaningful feedback, maintaining the continuity of learning regardless of setting. The system ensures that learners are never operating without guidance, even in the absence of real-time human interaction.

Ultimately, the AI-infused Green Bucket positions feedback as the core of instruction. Through the integration of AI, feedback becomes precise, actionable, concise, rubric-aligned, ungraded, immediate, and iterative—delivered consistently across learners and contexts. When these elements are systematically applied, feedback ceases to be an adjunct to teaching and instead becomes the primary driver of learning, guiding performance through clear, timely, and disciplined instructional communication.

Annotated References

Britton, E. R. (2023). Developing teacher feedback literacy through self-study. *Teaching and Teacher Education*.
Abstract page (working):
https://www.sciencedirect.com/science/article/abs/pii/S107529352300017X

This study shows how educators refine feedback practices through structured reflection. It supports your emphasis on discipline and precision in feedback design, demonstrating that effective feedback is not intuitive but developed through systematic alignment with criteria and purpose.

Dawson, P., Yan, Z., Lipnevich, A. A., Tai, J., Boud, D., & Mahoney, P. (2023). Measuring what learners do in feedback: The Feedback Literacy Behaviour Scale. *Assessment & Evaluation in Higher Education, 49*(3), 348–362.
Direct PDF: https://blogs.deakin.edu.au/cradle/wp-content/uploads/sites/188/2023/08/Measuring-what-learners-do-in-feedback-the-feedback-literacy-behaviour-scale.pdf

This is one of the strongest current empirical studies on feedback. It shifts the focus from perceptions to observable learner behaviors, showing that feedback is only effective when students *act on it*. This directly supports the Green Bucket model's emphasis on actionable, iterative feedback cycles and positions feedback as a process rather than a static response.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning: A review of research and practice. Education Sciences, 13(4), 410. https://www.mdpi.com/2227-7102/13/4/410

This peer-reviewed review examines how generative AI tools function in instructional contexts, with particular attention to feedback generation. The authors demonstrate that AI systems can deliver immediate, individualized, and consistent responses to student work. This directly supports the Green Bucket model's emphasis on immediacy and uniformity. The study also highlights the importance of structured prompts and clearly defined criteria, reinforcing the need for rubric alignment in AI-generated feedback.

Little, T., Dawson, P., Boud, D., & Tai, J. (2024). Can students' feedback literacy be improved? A scoping review of interventions. *Assessment & Evaluation in Higher Education*.
Open-access PDF:
https://repository.mdx.ac.uk/download/e90080bb662ea6e88d162bf3fd15b911e85f7eec6cfb9e55bd274a0ebe40b0f4/1894943/2024-Little%20etal-Can%20students%20feedback%20literacy%20be%20improved.pdf

This review confirms that structured interventions significantly improve students' ability to use feedback, including confidence, interpretation, and action-taking. It validates the Green Bucket's design as a systematic feedback environment, rather than ad hoc commentary, and supports the idea that feedback must be explicitly structured to influence learning.

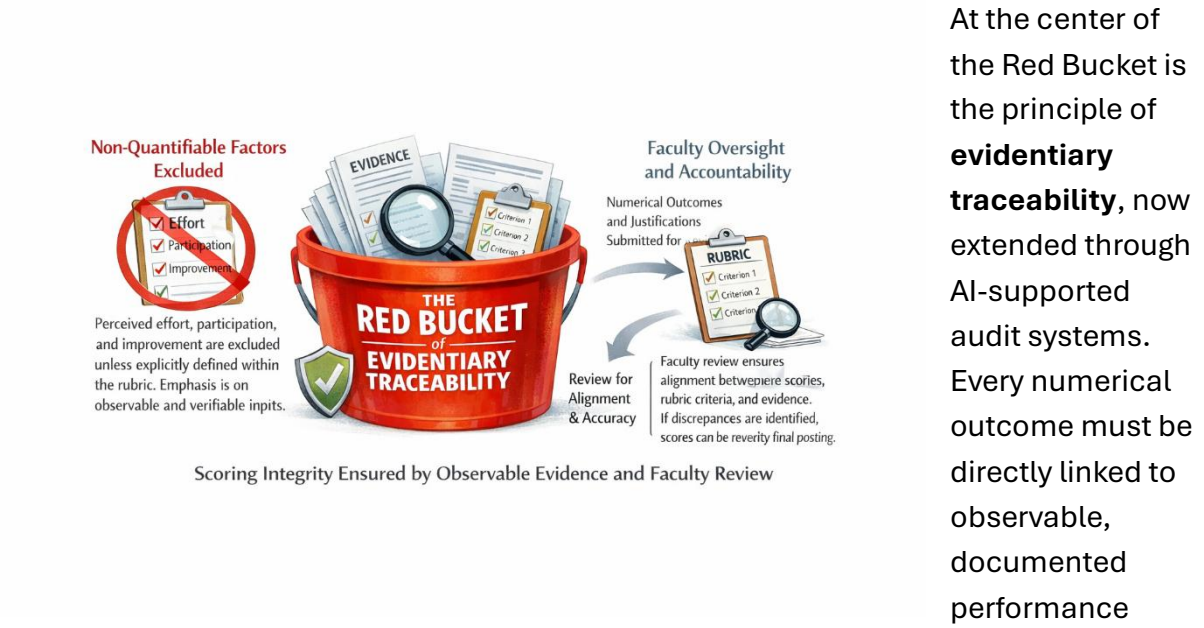
Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4475995

This widely cited open-access paper provides practical and research-informed strategies for integrating AI into teaching and learning. It positions AI as a tool for iterative feedback and revision, supporting the idea that learners benefit from rapid cycles of response and improvement. The work aligns closely with the Green Bucket's design by showing how AI can function as a continuous feedback mechanism when guided by explicit expectations.

Chapter 7 Generating Scores in the Red Bucket

The generation of scores occurs exclusively within the Red Bucket. In an AI-infused assessment system, this phase formalizes the translation of qualitative evidence into quantitative outcomes through disciplined, rule-governed processes embedded within the rubric itself. Scores are not independent judgments; they are structured derivatives of prior narrative assessment, mediated through AI-supported rubric architectures. The qualitative feedback produced in the Green Bucket—anchored in explicit, AI-infused rubric criteria—serves as the sole evidentiary basis for numerical conversion. Each score emerges from a transparent alignment between documented performance and predefined performance levels, ensuring that quantification remains procedural rather than interpretive. The Red Bucket introduces no new evaluation; it executes a controlled transformation of established qualitative judgment into a standardized, defensible metric suitable for reporting, comparison, and institutional use.

The purpose is not to reduce learning to a number, but to formalize judgment in a way that satisfies institutional requirements while preserving fidelity to evidence. Within an AI-infused rubric system, quantification is not an independent act; it is a derivative process, algorithmically constrained and fully grounded in prior qualitative analysis.



Scoring Integrity Ensured by Observable Evidence and Faculty Review

At the center of the Red Bucket is the principle of **evidentiary traceability**, now extended through AI-supported audit systems. Every numerical outcome must be directly linked to observable, documented performance aligned with rubric criteria that are both human-designed and machine-readable. Unlike traditional grading systems—where scores may reflect impressionistic or holistic judgments—the AI-infused Red Bucket requires that all numerical assignments emerge from discrete, criterion-based observations that can be tracked, verified, and reproduced. Each element of the rubric functions as a measurable dimension, and the aggregation of these dimensions produces the final quantitative result. The score is

therefore not an abstraction, but a structured representation of demonstrated learning with a verifiable audit trail.

The process begins with the stabilization of qualitative evidence. Narrative feedback generated in the Green Bucket is parsed and organized by AI-assisted systems according to rubric criteria. This involves categorizing comments, observations, and identified strengths or gaps under clearly defined performance indicators. AI does not generate new judgment; it organizes existing human or AI-supported feedback into aligned structures. The goal is to eliminate ambiguity by ensuring that each qualitative statement corresponds to a specific criterion. For example, a comment regarding argument clarity is automatically aligned to descriptors such as “thesis articulation” or “logical coherence,” transforming narrative feedback into analyzable, rubric-bound units.

Once aligned, qualitative evidence is translated into performance levels. Each rubric criterion contains predefined descriptors associated with calibrated levels of proficiency. In an AI-infused rubric, these descriptors are encoded so that alignment between evidence and level can be assisted—though not replaced—by machine processes. The evaluator confirms the appropriate level based solely on documented performance. By anchoring decisions to explicit descriptors and supported alignment mechanisms, variability is reduced and inter-rater reliability is strengthened across contexts.

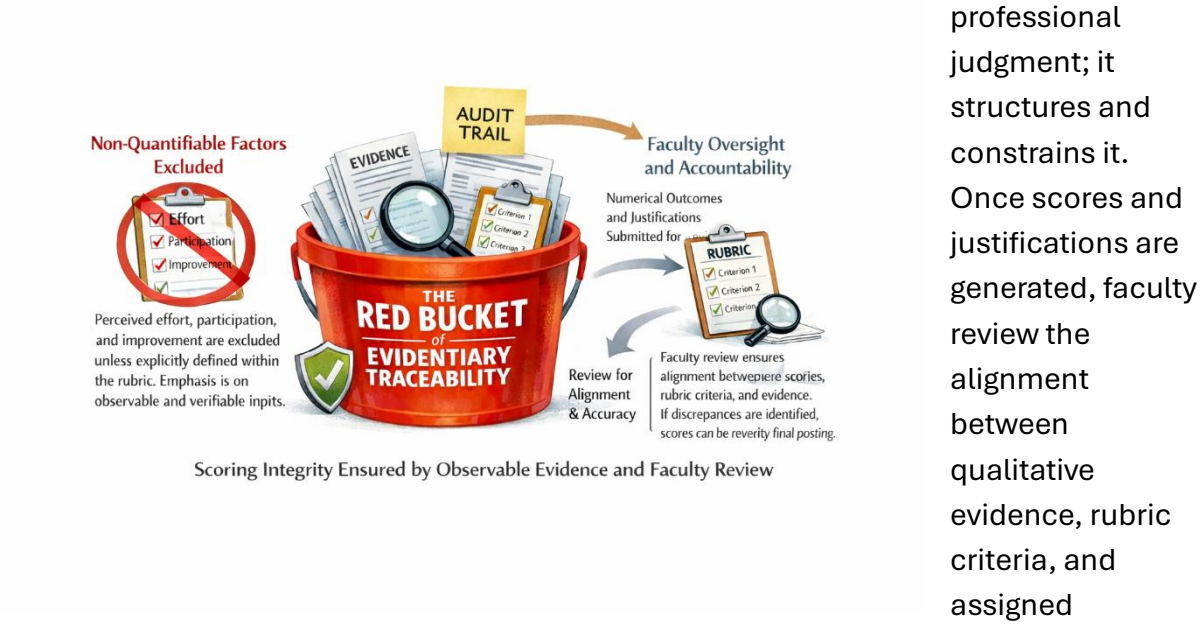
The next stage is numerical encoding. Each performance level within the rubric corresponds to a predefined numerical value embedded within the AI-infused rubric system. Whether using a 0–4 scale, a 1–5 scale, or a weighted index, the values are fixed, transparent, and consistently applied. AI systems ensure that once a performance level is selected, the numerical conversion is automatic and error-free. This removes discretionary scoring behavior and preserves procedural integrity.

Aggregation follows encoding. Individual criterion scores are combined into an overall quantitative outcome through rule-based models embedded in the rubric system. These may include weighted averages or proportional scoring structures aligned to instructional priorities. AI supports this stage by enforcing consistency in aggregation and instantly recalculating outcomes when revisions occur. Criteria designated as more critical exert proportionally greater influence, ensuring that the final score reflects the designed hierarchy of learning priorities.

A defining feature of the AI-infused Red Bucket is the requirement for concise, rubric-based justification. Each numerical outcome is accompanied by a brief explanation—typically no more than 50 words—generated or supported by AI but strictly constrained to rubric-aligned evidence. This justification references only quantifiable outcomes

(e.g., performance levels across criteria) rather than narrative interpretation. The constraint enforces precision and prevents drift back into subjective commentary.

Importantly, the Red Bucket excludes non-quantifiable factors unless explicitly encoded within the rubric. Perceived effort, participation, or improvement over time are not included unless they are operationalized as observable criteria. AI systems reinforce this boundary by limiting inputs to defined rubric fields, preventing the inclusion of implicit or affective judgments. This design safeguards the integrity of the scoring process and reduces bias. Faculty oversight remains central. AI does not replace professional judgment; it structures and constrains it. Once scores and justifications are generated, faculty review the alignment between qualitative evidence, rubric criteria, and assigned outcomes. AI supports this review by surfacing discrepancies, highlighting misalignments, and maintaining a complete audit trail. If inconsistencies are identified, revisions can be made before final posting. This layered process ensures accountability without sacrificing efficiency.



Scoring Integrity Ensured by Observable Evidence and Faculty Review

Technology enables the Red Bucket to function at scale. AI-infused rubric systems encode criteria, automate alignment, perform numerical conversion, and generate audit-ready summaries. These systems reduce cognitive load and enforce procedural consistency, but they operate within a tightly bounded framework defined by human-designed rubrics. The role of AI is not to evaluate independently, but to ensure that evaluation adheres strictly to evidence and structure.

The Red Bucket also resolves the longstanding problem of grade defensibility. Every score is supported by a transparent chain of evidence: qualitative feedback, rubric alignment, performance level assignment, numerical encoding, and aggregation. AI systems preserve this chain as an auditable record, allowing any score to be traced back to its evidentiary origins. This traceability strengthens trust across students, educators, and institutions.

Equally important is the maintained separation between the Green and Red Buckets. Feedback remains instructional and non-quantified in the Green Bucket, while quantification occurs only after qualitative analysis is complete. AI enforces this sequencing by preventing premature scoring and ensuring that numerical outputs are generated only from stabilized evidence. This prevents scores from distorting interpretation and preserves the integrity of the learning process.

In practice, the AI-infused Red Bucket establishes a disciplined bridge between narrative understanding and institutional reporting. It translates rich, descriptive evidence into structured quantitative outcomes without severing the connection to learning. The score is not the endpoint of evaluation; it is a compressed, rule-governed representation of documented performance.

Ultimately, the Red Bucket is a necessary but tightly controlled component of evaluation. Through AI-infused rubrics, systematic alignment, calibrated encoding, and evidence-based justification, qualitative judgment is transformed into quantitative form with precision and accountability. Scores are no longer arbitrary conclusions; they are verifiable outputs of a transparent, structured, and auditable process grounded entirely in observable learning.

Annotated Reference Section

AlGhamdi, E. (2025).
Leveraging prompt-based large language models for automated scoring and feedback generation in higher education. *Computers & Education, 243*, 105511.
https://doi.org/10.1016/j.compedu.2025.105511

This open-access study demonstrates how LLMs can perform scoring when guided by structured prompts aligned to rubric criteria. It highlights that scoring quality depends on tight alignment between prompts and predefined evaluation dimensions, reinforcing the Red Bucket principle that AI must operate within rule-governed structures rather than free-form judgment.

Campbell, K. K., Holcomb, M. J., Vedovato, S., Young, L., Danuser, G., Dalton, T. O., Jamieson, A. R., & Scott, D. J. (2025).
Applying state-of-the-art artificial intelligence to grading in simulation-based education: Assessment, feedback, and ROI. *Discover Artificial Intelligence, 5*, Article 202.
https://doi.org/10.1007/s44163-025-00417-3

This peer-reviewed, open-access study provides direct empirical evidence of AI-driven scoring systems aligned to structured rubrics. The authors demonstrate that AI grading achieved 83% agreement with human raters and significantly improved efficiency while maintaining rubric fidelity. This directly supports the Red Bucket model: scores are

derived from structured rubric criteria, not independent judgment, and can be audited and reproduced with high consistency.

Eyal, L. (2025).
Developing and validating an AI-TPACK assessment tool for teacher educators.
Education Sciences, 15(11), 1452.
https://doi.org/10.3390/educsci15111452

Eyal provides a validated AI-supported assessment instrument grounded in structured criteria and measurable constructs. The study demonstrates how AI-assisted scoring can achieve validity only when tied to clearly defined performance dimensions, directly supporting the Red Bucket's emphasis on criterion-based quantification and reproducibility.

Gao, Y., et al. (2025).
A multimodal interactive framework for science assessment in the era of generative artificial intelligence. *Journal of Research in Science Teaching*.
https://doi.org/10.1002/tea.70009

Gao and colleagues propose a structured AI-supported evaluation framework emphasizing alignment between qualitative evidence and measurable criteria. Their model demonstrates how AI systems categorize and translate complex student outputs into structured evaluation dimensions. This aligns precisely with the Red Bucket's process of stabilizing qualitative evidence before quantitative encoding.

Tensen, D. (2025).
Using AI to generate formative feedback in doctoral education. *Assessment & Evaluation in Higher Education*.
https://doi.org/10.1080/02602938.2025.2536558

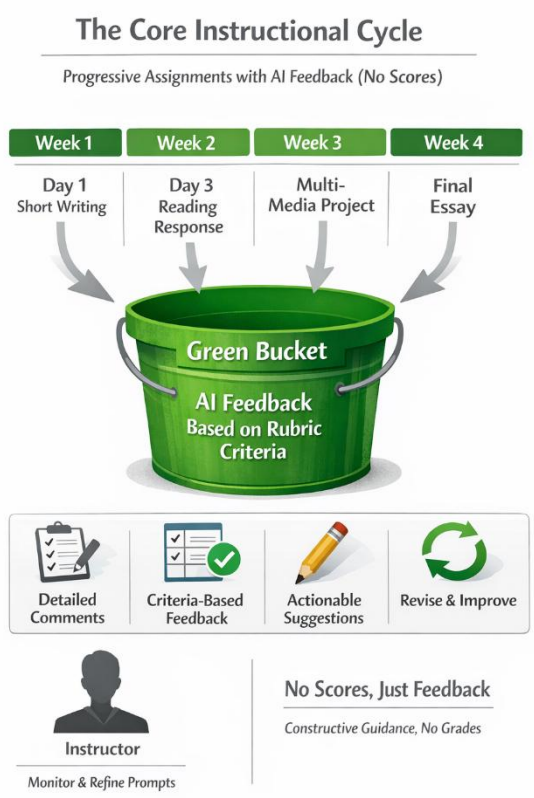
Tensen explores AI-assisted feedback systems and highlights the importance of structured alignment between feedback and evaluation criteria. While focused on feedback, the study shows that meaningful quantification depends on prior stabilization of qualitative evidence—directly supporting the Green-to-Red Bucket transition.

Chapter 8 Implementation of Dual-Bucket Program

This section provides practical guidance for implementing the AI-infused dual-bucket paradigm within the constraints of a one-course-per-month online college environment. The emphasis is not on theoretical positioning, but on operational clarity—how instructors structure workflow, manage timing, and integrate AI-supported processes so that qualitative feedback and quantitative scoring remain distinct, efficient, and instructionally coherent.

At the center of implementation is the separation of functions between the Green Bucket and the Red Bucket. In practice, this separation must be reflected in both the design of the course and the sequence of instructional activity. The Green Bucket governs all formative interaction with student work. It is where student submissions and re-submissions are reviewed, rubric-aligned feedback is generated, and revision is guided. The Red Bucket, by contrast, is activated only after the feedback cycle has stabilized. It is responsible solely for translating established qualitative evidence into quantitatively processed scores through rule-based rubric conversion. Maintaining this boundary is essential; efficiency emerges not from speed alone, but from the elimination of redundancy and role confusion.

In a one-course-per-month model, compressed timelines demand a tightly structured weekly cadence. A four-week course can be organized into iterative cycles that align directly with the dual-bucket system. Beginning in the first week, expectations are established and the rubric architecture is made explicit. Students engage in submissions designed to reveal baseline performance. During this initial phase, AI-supported feedback within the Green Bucket delivers immediate, criterion-aligned commentary that clarifies expectations without introducing scores. The purpose is calibration: students learn how their work is interpreted through the rubric before any conversion to quantitative outcomes occurs.



Different assignments due on distinct days throughout week one and into weeks two, three, and four represent the core instructional cycle. Students submit progressively more complex work, and each submission is processed at first exclusively within the Green Bucket. AI systems generate structured feedback tied directly to rubric criteria, ensuring consistency across all learners. The instructor's role is to monitor alignment, refine prompts when necessary, and intervene only when feedback requires adjustment for accuracy or nuance. Because the feedback is rubric-bound, it remains focused and actionable, allowing students to revise efficiently. Importantly, no scores are introduced during the Green Bucket phase. This preserves the integrity of feedback as instruction rather than evaluation.

Workflow efficiency is achieved through batching and standardization. Rather than responding to each submission individually in an unstructured manner, assignments are processed through predefined rubric channels. AI tools can be configured to recognize criteria, generate commentary, and organize responses in a consistent format. This reduces variability and ensures that feedback remains anchored to expectations. Instructors review outputs at a systems level rather than rewriting responses, allowing attention to shift from production to verification. The result is a scalable process that maintains quality without requiring unsustainable effort.

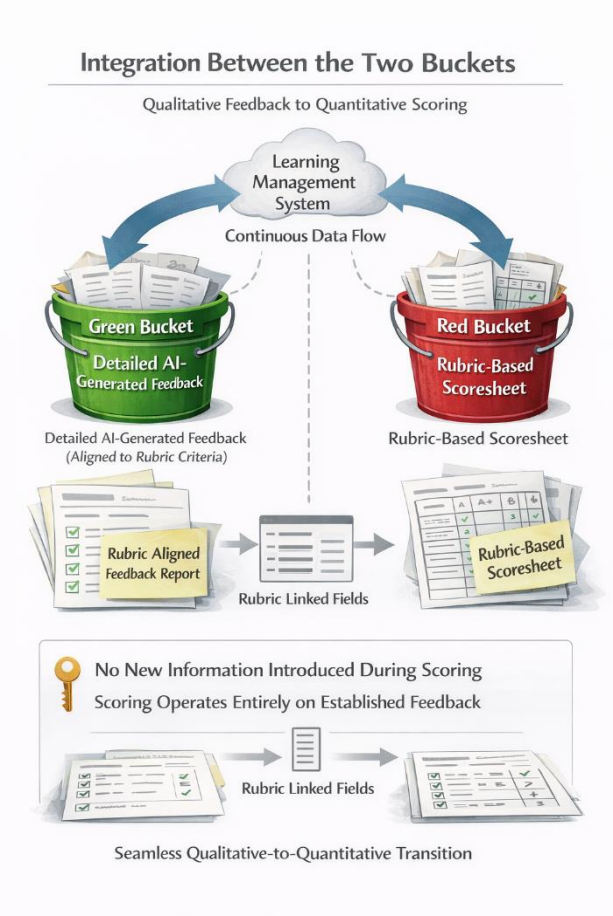
Timing within this model is intentionally structured. Feedback is delivered immediately to guide revision while the work is still cognitively active for the learner, establishing a consistent and predictable rhythm. AI-generated responses are produced instantaneously and are drawn directly from rubric criteria that students have already read and conceptualized, ensuring alignment, transparency, and instructional continuity.

Once a Green Bucket assignment is reviewed, the student receives approximately 150 words of qualitative, rubric-aligned commentary. Using that feedback, the student determines whether to revise and resubmit to the Green Bucket. There is no limit to the number of revision cycles; students may continue refining their work until they

determine it meets the defined criteria. Only when the student is satisfied does the work move forward to the Red Bucket.

By this stage, multiple cycles of feedback and revision have produced a structured body of qualitative evidence aligned to the rubric. Once the student deposits the work into the Red Bucket, the shift to scoring is a procedural conversion of that existing evidence. AI-supported systems map documented performance to predefined rubric levels, generating numerical outcomes that are directly traceable to prior feedback. Each score includes a brief justification derived from the established commentary, maintaining transparency and defensibility in the quantification process.

Integration between the two buckets depends on data continuity. All feedback generated in the Green Bucket must be stored in a way that preserves its alignment to rubric criteria. This creates an evidentiary record that the Red Bucket can access without reinterpretation. Learning management systems can support this integration by structuring assignments around rubric-linked fields, enabling seamless transition from qualitative input to quantitative output. The key is that no new information is introduced during scoring; the Red Bucket operates entirely on what has already been established.



Instructor workload, often a central concern in accelerated courses, is managed through redistribution rather than reduction. Effort shifts from repetitive grading tasks to oversight of system performance. Instead of writing extensive comments for each student, the instructor ensures that AI-generated feedback remains accurate, relevant, and aligned. This supervisory role is more sustainable because it operates at the level of patterns rather than individual instances. When adjustments are needed, they are applied to the rubric or prompt structure, improving all subsequent outputs simultaneously.

Student experience within this model is defined by clarity and continuity. Because feedback is consistent and rubric-aligned, learners can track their progress across iterations without interpreting conflicting signals. The absence of immediate scoring reduces fixation on outcomes and redirects attention toward revision. By the time

scores are introduced, they reflect a stabilized representation of performance rather than a snapshot influenced by timing or inconsistency. This alignment between process and outcome reduces confusion and supports sustained engagement.

Effective implementation also depends on precise communication. Weekly live online sessions should explicitly introduce the dual-bucket structure, clarifying that rubric-aligned feedback functions as the central mechanism for learning, while numerical scores represent a final procedural conversion rather than an evaluative starting point. When this distinction is clearly communicated, students engage more consistently in iterative revision within the Green Bucket, and the use of transparent rubrics produces higher-quality, evidence-aligned Red Bucket outcomes.

Finally, efficiency in this system is not achieved by accelerating every component, but by sequencing them correctly. The Green Bucket prioritizes instructional interaction, while the Red Bucket formalizes outcomes. By isolating these functions and supporting them with AI-driven processes, the dual-bucket paradigm becomes not only feasible but effective within the compressed timeline of a one-course-per-month model. The result is a coherent workflow in which feedback drives learning, scoring reflects evidence, and instructional effort is directed where it has the greatest impact.

Annotated Bibliography

Alghamdi, L. H. H., & Alghizzi, T. M. M. (2025). Educators' reflections on AI-automated feedback in higher education: A structured integrative review of potentials, pitfalls, and ethical dimensions. *Frontiers in Education, 10*, 1704820. https://doi.org/10.3389/feduc.2025.1704820

This open-access review is highly relevant to Chapter 8 because it examines how faculty view AI-generated feedback in higher education, including its usefulness, risks, and ethical constraints. It supports the emphasis on instructor oversight, system monitoring, and the need to keep AI feedback aligned, accurate, and instructionally defensible.

Boud, D., & Dawson, P. (2023). What feedback literate teachers do: An empirically-derived competency framework. *Assessment & Evaluation in Higher Education, 48*(2), 158–171. https://doi.org/10.1080/02602938.2021.1910928

This article provides an empirically derived framework for teacher feedback literacy. It fits your chapter especially well because the Green Bucket depends on well-designed feedback processes, clear criteria, and consistent instructional use of commentary rather than improvised grading.

Carless, D., & Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education, 28*(1), 150–163. https://doi.org/10.1080/13562517.2020.1782372

This article is central to the chapter's logic because it explains how teacher-designed feedback systems and student use of feedback work together. It supports the discussion of revision cycles, student decision-making, and the need for clear communication about how formative feedback functions before any scoring begins.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*, Article 22. https://doi.org/10.1186/s41239-023-00392-8

This systematic review supports the chapter's broader claim that AI can be used in higher education to improve instructional efficiency, feedback workflows, and process consistency. It is useful for grounding the chapter's implementation model in the current research landscape rather than presenting the dual-bucket system as purely conceptual.

Martin, F., Kim, S., & Bolliger, D. U. (2025). Assessment types, strategies, and feedback in online higher education courses in the age of artificial intelligence: Perspectives of instructional designers. *TechTrends*. Advance online publication. https://doi.org/10.1007/s11528-025-01115-8

This source is especially strong for this chapter because it addresses online higher education directly and examines assessment, feedback, and AI in course design practice. It supports the treatment of compressed online course structures, workflow planning, and the need for deliberate assessment sequencing in accelerated environments.

Topping, K. J., Gehringer, E., Khosravi, H., Gudipati, S., Jadhav, K., & Susarla, S. (2025). Enhancing peer assessment with artificial intelligence. *International Journal of Educational Technology in Higher Education, 22*, Article 17. https://doi.org/10.1186/s41239-024-00501-1

This article is helpful because it explores how AI can structure and strengthen feedback processes. Even though its focus includes peer assessment, it is relevant to your chapter's concern with standardization, criterion-based commentary, and scalable feedback architectures that reduce inconsistency. (Springer)

Weidlich, J., Gotsch, F., Schudel, K., Marusic-Würscher, C., et al. (2025). Teacher, peer, or AI? Comparing effects of feedback sources in higher education. *Computers and Education Open, 8*, 100300. https://doi.org/10.1016/j.caeo.2025.100300

This study is valuable for Chapter 8 because it compares different feedback sources in higher education and therefore speaks directly to the instructional place of AI-generated commentary. It helps support the argument that AI can be used within feedback workflows, but that its function must be understood in relation to other feedback sources and pedagogical goals.

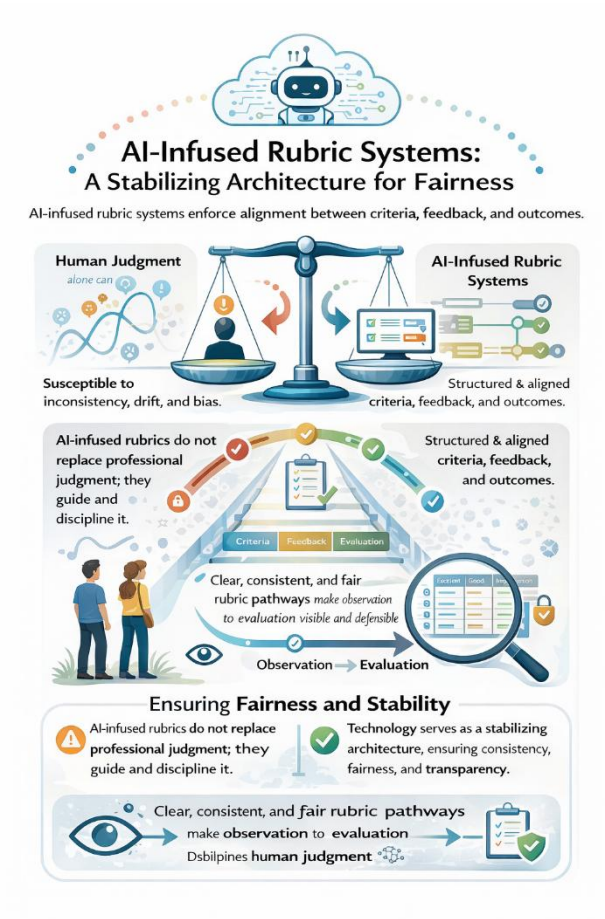
Final Thoughts

The work presented here rests on a simple but far-reaching premise: learning and the measurement of learning are not the same act, and when they are treated as if they are, both are diminished. For generations, educational systems have compressed complex human development into single numerical indicators, asking scores to carry meanings they were never designed to hold. In doing so, they have obscured growth, flattened context, and redirected attention away from the processes that actually produce learning. This book has argued that clarity can be restored—not by abandoning evaluation, but by placing it in its proper role and sequence.

At the center of this reorientation is the deliberate separation between qualitative assessment and quantitative evaluation. When these functions are disentangled, each becomes more precise, more useful, and more aligned with its purpose. Qualitative assessment regains its instructional integrity. It becomes a space where judgment is expressed through language rather than numbers, where learners encounter clear, criterion-based feedback that they can act upon immediately. In this environment, revision is not a penalty or an afterthought; it is the mechanism through which learning advances. The absence of scores is not a lack, but a protection—preserving attention on improvement rather than performance.

Quantitative evaluation, when it is finally introduced, is no longer burdened with the task of motivating, guiding, and explaining. It becomes what it should have been all along: a structured summary of demonstrated performance. Because it is derived from an accumulated body of qualitative evidence, it carries greater legitimacy. Because it is rule-governed and traceable to explicit criteria, it carries greater transparency. And because it is separated from the formative process, it avoids distorting the very learning it seeks to represent.

The integration of AI-infused rubric systems makes this separation not only conceptually sound, but operationally viable at scale. Where human judgment alone is susceptible to inconsistency, drift, and bias, structured systems enforce alignment between criteria, feedback, and outcomes. They do not replace professional judgment; they discipline it. They ensure that what is said to one learner is comparable to what is said to another, that expectations remain stable over time, and that the pathway from observation to evaluation is visible and defensible. In this sense, technology serves not as a shortcut, but as a stabilizing architecture for fairness.



Equally important is the way this model redefines time within the learning process. Feedback arrives when it can still be used. Learners engage in multiple cycles of submission and revision without artificial limits. Evidence accumulates gradually, rather than being captured in isolated snapshots. Evaluation occurs only after this process has matured, transforming what has been observed into a form required by institutions without interrupting the work of learning itself. The result is a system that is both efficient and humane—structured enough to meet institutional demands, yet flexible enough to reflect how learning actually unfolds.

The implications extend beyond technique. A system built on clear separation, transparent criteria, and traceable judgment is inherently more equitable. It reduces the influence of hidden expectations and subjective interpretation. It makes visible the basis on which decisions are made. It allows learners to understand not only where they stand, but how they arrived there and how they might improve. In doing so, it shifts the culture of assessment from one of judgment imposed to one of evidence examined.

What emerges is not simply a new method, but a different orientation toward teaching and learning. Instruction becomes the design of conditions under which learners can refine their work through informed iteration. Evaluation becomes the disciplined reporting of what that work demonstrates. Between them lies a clear boundary—one that protects the integrity of both.

If this boundary is maintained, the effects are cumulative. Feedback becomes more meaningful. Scores become more accurate. Systems become more transparent. And most importantly, learning is no longer subordinated to the mechanics of grading. It is allowed to proceed as a process—visible, revisable, and grounded in evidence—while evaluation follows as a coherent and defensible conclusion.