The Anchoring Hypothesis and the Illusion of Learning: Why High-Performing Practice Students Underperform on High-Stakes Assessments

Oniya Rolle, Kenya James-Crosby, Jakeria Lawson, Layla Robinson, Atreonia Garner, Alexia Easley, Aiyana Greenwood, Francesca Millien

Faculty Mentor: D. N. Gray-Singh

Department of Biological Sciences, Clark Atlanta University

Abstract

Adaptive learning technologies are increasingly used in undergraduate STEM education because they provide personalized instruction, immediate feedback, and repeated opportunities for retrieval practice. Although students frequently achieve high levels of performance within these environments, it remains unclear whether adaptive learning metrics accurately reflect durable conceptual understanding or readiness for authentic assessment. This study investigated the relationship between adaptive learning performance and summative examination performance and introduces the 'Anchoring Hypothesis' as one possible interpretation of systematic discrepancies between apparent mastery and transferable understanding. Data were analyzed from 73 students enrolled in four sections of an introductory General Biology course. Students completed 11 adaptive online learning modules and three instructor-developed examinations assessing conceptual application and transfer. The third examination was administered as a retake following scaffolded review and is therefore reported both within and separately from the primary examination average. Adaptive quiz performance, examination averages, and a calibration gap (adaptive quiz percentage minus examination percentage) were analyzed using descriptive statistics, correlations, group comparisons, and effect-size estimates. Students achieved a mean adaptive quiz score of 87.8%, compared with a mean examination average of 54.1% across all three examinations, producing an average calibration gap of 33.7 percentage points. When the scaffolded retake was excluded, the examination average was 46.1% and the corresponding gap widened to 41.7 percentage points. Adaptive quiz performance was only modestly correlated with examination average ($r = .32$, $p = .006$), accounting for approximately 10% of the variance in examination outcomes. The discrepancy was observed across nearly the entire cohort and was significantly larger among lower-performing students ($t(34) = 5.03$, $p < .001$, Cohen's $d = 1.54$). Transfer-oriented examination questions requiring conceptual integration consistently revealed substantially lower performance despite high adaptive learning scores. These findings document a robust discrepancy between performance within adaptive learning environments and authentic assessment, which we term the Illusion of Mastery. As one possible account of this phenomenon, we propose the Anchoring Hypothesis, which suggests that durable competence depends not merely on repeated successful retrieval but on anchoring knowledge within interconnected conceptual networks that support abstraction, integration, and transfer. The hypothesis is offered as an interpretive framework consistent with established theories of meaningful learning, transfer, retrieval practice, and metacognition rather than as a confirmed causal mechanism. The findings have practical implications for the use of adaptive learning analytics and for the assessment of conceptual understanding in undergraduate STEM education.

Keywords: Adaptive learning; learning analytics; transfer; retrieval practice; metacognitive calibration; biology education; assessment; undergraduate STEM

The Anchoring Hypothesis and the Illusion of Mastery: Why High-Performing Practice Students Underperform on High-Stakes Assessments

Adaptive learning technologies have become an increasingly prominent component of undergraduate science education because they provide personalized instruction, immediate feedback, opportunities for repeated retrieval practice, and detailed learning analytics that allow instructors to monitor student progress in real time. These platforms promise to individualize instruction by adjusting question difficulty, providing targeted remediation, and encouraging students to practice until mastery is achieved. Numerous investigations have demonstrated that retrieval practice, timely formative feedback, and repeated testing can substantially improve learning under appropriate instructional conditions (Roediger & Karpicke, 2006; Karpicke & Roediger, 2008; Butler, 2010; Karpicke & Blunt, 2011). Likewise, intelligent tutoring systems and adaptive instructional environments have repeatedly demonstrated learning gains comparable to, and in some cases approaching, those produced by expert human tutoring (VanLehn, 2011; Kulik & Fletcher, 2016; Wang et al., 2023). Recent reviews of personalized adaptive learning in higher education report generally positive but variable effects on academic performance and engagement (Du Plooy et al., 2024), and individual implementations have reported gains in both performance and student satisfaction relative to conventional delivery (Contrino et al., 2024).

Despite these advances, an important question remains unresolved. What does success within an adaptive learning environment actually represent? High adaptive scores are commonly interpreted as evidence that students have mastered the underlying material and are prepared for more demanding assessments. Embedded within this assumption is the belief that repeated success on adaptive practice reflects the development of durable knowledge that can subsequently be retrieved, integrated, and transferred to unfamiliar contexts. Yet classroom experience frequently suggests otherwise. Students who consistently achieve exceptionally high scores on adaptive assignments often perform unexpectedly poorly on traditional examinations assessing the same content domains. This recurring discrepancy raises a practical question for biology education: do adaptive learning systems measure conceptual understanding, or do they primarily measure successful performance within the adaptive environment itself? The relationship between adaptive practice and subsequent examination performance has been examined directly in nursing education (Sheaves et al., 2022), but it remains comparatively little studied in undergraduate biology, where assessments frequently require integration across levels of biological organization.

The distinction between successful performance and durable learning has occupied educational psychologists for decades. Early theories of meaningful learning argued that long-term understanding depends upon integrating new information into existing cognitive structures rather than simply accumulating isolated facts (Ausubel, 1968; Bruner, 1960; Mayer, 2002). Similarly, research on expertise has demonstrated that experts differ from novices not merely in the quantity of knowledge they possess but in the organization of that knowledge into richly interconnected conceptual frameworks that support flexible reasoning across novel situations (Chi et al., 1981; Ericsson & Kintsch, 1995). Studies of transfer have likewise shown that knowledge acquired in one context frequently fails to generalize when learners encounter different cognitive demands, unfamiliar problems, or altered assessment conditions (Perkins & Salomon, 1989; Barnett & Ceci, 2002; Bransford & Schwartz, 1999). Meta-analytic evidence indicates that the benefits of repeated testing extend reliably to closely related material but diminish substantially as the distance between practice and assessment conditions increases (Pan & Rickard, 2018). Collectively, these perspectives suggest that successful performance during instruction should not automatically be equated with the capacity to transfer knowledge to authentic problem-solving contexts.

A complementary body of research has emphasized that learning and performance are not synonymous. Immediate success during instruction often reflects temporary performance gains rather than durable learning, particularly when instruction provides extensive scaffolding, repeated exposure, or corrective feedback (Bjork & Bjork, 2011; Soderstrom & Bjork, 2015; Kornell & Bjork, 2008). Research on metacognitive calibration further demonstrates that learners frequently overestimate their own understanding when fluent recognition is mistaken for genuine comprehension (Koriat & Bjork, 2005; Kruger & Dunning, 1999; Dunlosky et al., 2013). Glenberg et al. (1982) first characterized this failure of self-assessment as an illusion of knowing, showing that learners judged their comprehension to be adequate even when they had not detected substantive inconsistencies in what they had read. Koriat and Bjork (2005) subsequently showed that the presence of correct answers during study systematically inflates learners' predictions of their own later unassisted performance, an effect they termed the illusion of competence. This distinction parallels long-standing evidence that recognition supported by external cues draws on different memory processes than effortful, unsupported recollection (Jacoby, 1991; Yonelinas, 2002). Adaptive learning platforms, by design, maximize successful performance through personalized sequencing, immediate feedback, and repeated opportunities to answer correctly. Although these features undoubtedly enhance engagement and facilitate retrieval practice, they may also produce highly optimistic indicators of learning that exceed students' actual readiness for independent conceptual reasoning.

Building on previous investigations examining instructional technology across STEM disciplines (Gray-Singh et al., 2025) and the role of growth mindset in shaping academic performance (Deimeke et al., 2024), the present study extends this program of scholarship by proposing the Anchoring Hypothesis as one possible explanation for why apparent mastery within adaptive learning environments frequently fails to transfer to authentic assessment contexts. Rather than viewing adaptive technologies as ineffective, this framework seeks to distinguish between two related but different dimensions of learning: successful performance within a scaffolded instructional environment and durable conceptual understanding that supports independent reasoning and transfer.

The present investigation was motivated by a striking empirical observation made across four sections of an introductory biology course. Students consistently achieved exceptionally high scores on adaptive online learning modules while performing substantially worse on instructor-developed examinations assessing the same biological concepts. We refer to this systematic discrepancy as the Illusion of Mastery: a condition in which platform-mediated indicators of learning substantially overestimate students' readiness for authentic assessment. Importantly, this phenomenon should not be interpreted as evidence that students failed to learn. Instead, it suggests that conventional measures of adaptive success may capture a different dimension of learning than the one ultimately required for successful transfer.

To account for this pattern, we introduce the Anchoring Hypothesis. The hypothesis proposes that durable competence depends not simply on acquiring correct information, but on anchoring that knowledge within an interconnected conceptual network that supports retrieval, integration, abstraction, and transfer to unfamiliar problems. In this view, adaptive learning environments effectively strengthen recognition, recall, and procedural fluency while only partially supporting the deeper conceptual organization required for independent reasoning. The Anchoring Hypothesis is not offered as a replacement for existing theories of meaningful learning, transfer, retrieval practice, expertise, or metacognitive calibration. Rather, it draws these complementary perspectives together into a single interpretive account of why adaptive learning environments may generate optimistic signals of preparedness that are not consistently reflected in subsequent examination performance.

The purpose of the present study is twofold. First, we document the magnitude and consistency of the discrepancy between adaptive practice performance and summative examination performance across four sections of an introductory biology course. Second, we consider whether the Anchoring Hypothesis provides a coherent and testable interpretation of this phenomenon. By distinguishing between recognition-based success and transferable conceptual understanding, this work contributes to an expanding body of scholarship examining how learning should be measured in increasingly technology-rich educational environments.

Method

Participants and Course Context

Participants were 73 undergraduate students enrolled in four sections (Sections 05–08) of an introductory General Biology I lecture and laboratory sequence during the Fall 2025 semester at a comprehensive undergraduate institution. Individual course sections enrolled between 18 and 19 students and were taught using a common syllabus, identical instructional materials, the same adaptive learning platform, and identical summative examinations to ensure instructional consistency across sections.

Although these sections originally represented comparison and intervention groups within a broader instructional investigation, the present study addresses a distinct research question. Specifically, our objective was not to evaluate the effectiveness of a pedagogical intervention but to examine the relationship between performance within an adaptive learning environment and subsequent performance on traditional summative assessments. Consequently, data from all four sections were pooled to maximize statistical power and to characterize the phenomenon across the complete instructional cohort.

All analyses were conducted using de-identified student records in accordance with institutional policies governing educational research. Sample sizes are reported for individual analyses when missing observations resulted in incomplete cases.

Instructional Environment

Throughout the semester, students completed eleven adaptive online learning modules aligned with the major biological concepts introduced during lecture and laboratory instruction. The adaptive platform individualized question delivery according to student performance, provided immediate corrective feedback following each response, permitted repeated attempts, and adjusted subsequent question difficulty as students demonstrated increasing proficiency. Consequently, students experienced multiple opportunities to revisit concepts until the platform determined that mastery had been achieved.

Adaptive learning systems are designed to optimize formative learning by combining retrieval practice, immediate feedback, and individualized pacing, instructional features that have consistently been associated with improved short-term learning outcomes (Roediger & Karpicke, 2006; Butler, 2010; VanLehn, 2011). In this respect they are designed to approximate features of individual human tutoring, in which learners receive continuous feedback and their errors are addressed as they arise (Chi et al., 2001). The particular format in which practice testing is implemented is not incidental, as different practice testing methods produce measurably different learning outcomes (Akbulut, 2024), and pairing practice testing with feedback has been associated with gains in both performance and persistence in technology-mediated courses (Lipnevich et al., 2025). Such systems also reduce the cognitive demands imposed during practice by structuring problems and supplying guidance at the point of difficulty (Sweller, 1988). Within the present study, these modules served as low-stakes formative assessments intended to reinforce foundational biological knowledge while encouraging continuous engagement throughout the semester.

Student learning was also evaluated using three instructor-developed summative examinations administered under conventional testing conditions. Unlike the adaptive modules, examinations did not provide immediate feedback, allowed only a single response to each question, and required students to retrieve, integrate, and apply biological concepts independently. Many examination items required students to synthesize knowledge across multiple content domains and apply previously learned concepts to unfamiliar biological problems, corresponding to the higher-order categories of analysis and application within contemporary taxonomies of educational objectives (Anderson & Krathwohl, 2001). Thus, although both assessment formats evaluated the same curricular content, they differed substantially in the degree of instructional scaffolding, opportunities for repeated success, and cognitive demands placed on learners (Table 1).

It is important to note one procedural difference among the three examinations. Examinations 1 and 2 were administered under standard conditions. Examination 3 was administered as a retake following an instructor-led review session in which the relevant concepts were revisited with substantial scaffolding. Because this scaffolding is precisely the instructional feature under examination in the present study, Examination 3 is reported both within the overall examination average and separately, and the primary interpretation rests on Examinations 1 and 2.

Table 1

Characteristics of the Adaptive Practice Modules and Summative Examinations

Characteristic	Adaptive Platform	Summative Examination	Associated Cognitive Demand
Immediate feedback	Yes	No	Recognition supported by external cues
Multiple attempts	Yes	No	Repeated retrieval under low stakes
Novel context	Limited	Extensive	Transfer to unfamiliar problems
Concept integration	Moderate	High	Coordination across content domains
Independent reasoning	Limited	High	Unsupported retrieval and application

Note. The final column describes the cognitive demand each feature is intended to place on learners. These characterizations are conceptual and were not independently verified through item-level classification; see Limitations.

Measures

Adaptive Practice Performance

Adaptive practice performance was operationalized as each student's average percentage score across the eleven adaptive learning modules (110 total possible points). Because the adaptive platform provided individualized question sequencing, immediate corrective feedback, and repeated opportunities for success, these scores should be interpreted as indicators of performance within the adaptive learning environment rather than direct measures of durable conceptual understanding.

Summative Examination Performance

Students completed three instructor-developed examinations distributed throughout the semester. Each examination was worth 100 points and assessed students' ability to retrieve, integrate, and apply biological concepts in novel contexts without immediate feedback or repeated attempts. Two outcome variables are reported: the average across all three examinations, and the average across Examinations 1 and 2 only, which excludes the scaffolded retake.

To illustrate the types of reasoning required, two representative examination tasks are described. The first required students to explain how a single amino acid substitution alters hemoglobin structure and function in sickle cell disease by integrating principles of molecular genetics, protein chemistry, and structure-function relationships. The second required students to explain mitochondrial uncoupling during thermogenesis in brown adipose tissue by applying principles of oxidative phosphorylation to a novel physiological context. These examples were selected because they exemplify transfer-oriented reasoning rather than rote factual recall.

Calibration Gap

To quantify the discrepancy between adaptive learning performance and summative examination performance, we calculated a calibration gap for each student as the adaptive quiz percentage minus the average examination percentage. Positive values indicate that performance within the adaptive learning environment exceeded subsequent examination performance.

Rather than interpreting this metric as an index of student error or overconfidence, we conceptualize the calibration gap as an empirical measure of divergence between two assessment environments that differ substantially in instructional scaffolding, feedback, and cognitive demands. Under the Anchoring Hypothesis, larger calibration gaps would be expected to reflect weaker conceptual anchoring despite successful performance within the adaptive environment.

Operational Definition of Anchoring

Because no validated psychometric instrument currently exists to measure conceptual anchoring directly, anchoring is treated in the present investigation as a theoretical construct rather than an observed variable. Anchored knowledge is defined as conceptual understanding that can be reliably retrieved, integrated, and transferred to unfamiliar biological problems under conditions requiring independent reasoning. In contrast, unanchored knowledge refers to information that can be accurately recognized or reproduced within familiar instructional contexts but fails to support consistent performance when students encounter novel problems that require abstraction, integration, and transfer. Because anchoring was not measured directly in this study, it is proposed as a candidate explanation for the observed discrepancy rather than as a demonstrated mechanism.

Statistical Analysis

Analyses were designed to characterize both the magnitude of the adaptive practice–examination discrepancy and its relationship to student achievement. Descriptive statistics were computed for all assessment variables. Pearson product-moment correlations were used to examine associations between adaptive practice performance and examination performance. One-way analyses of variance compared examination performance across achievement groups, and independent-samples t tests evaluated differences in calibration gaps between lower- and higher-performing students. Standardized effect sizes were calculated using Cohen's d, with values of 0.20, 0.50, and 0.80 interpreted as small, medium, and large effects, respectively (Cohen, 1988).

Because achievement groups were defined using examination scores, regression toward the mean represents a potential source of bias. Consequently, interpretation emphasizes effect sizes, consistency across complementary analyses, and convergence of evidence across multiple statistical approaches rather than statistical significance alone.

Results

A Robust Discrepancy Between Adaptive Practice and Summative Examination Performance

The principal finding of this investigation was a large, consistent, and educationally meaningful discrepancy between students' performance within the adaptive learning environment and their performance on traditional summative examinations. Across the complete cohort (N = 73), students achieved a mean adaptive quiz score of 87.8%, whereas their mean examination average across all three examinations was 54.1%, yielding an average calibration gap of 33.7 percentage points (Table 2). When the analysis was restricted to Examinations 1 and 2, excluding the scaffolded retake, the mean examination average was 46.1% and the calibration gap widened to 41.7 percentage points. Rather than representing isolated instances of poor examination performance, this discrepancy characterized the cohort as a whole and was observed across all four course sections.

The association between the two measures was statistically significant but modest. Adaptive quiz performance and examination average were positively correlated, $r = .32$, $p = .006$, $n = 73$. Although this association is reliably different from zero, it accounts for approximately 10% of the variance in examination performance ($r^2 = .10$), leaving the large majority of variation in examination outcomes unexplained by performance within the adaptive environment. Because adaptive quiz scores occupied a restricted range relative to examination scores, this coefficient is likely attenuated and should be interpreted as a lower bound on the true association.

This finding suggests that adaptive learning environments and summative examinations capture related but distinct dimensions of student learning. Whereas adaptive practice consistently produced high indicators of apparent mastery, examination performance demonstrated substantially greater variability and considerably lower overall achievement. The magnitude and consistency of this discrepancy constitute the central empirical observation motivating the Anchoring Hypothesis.

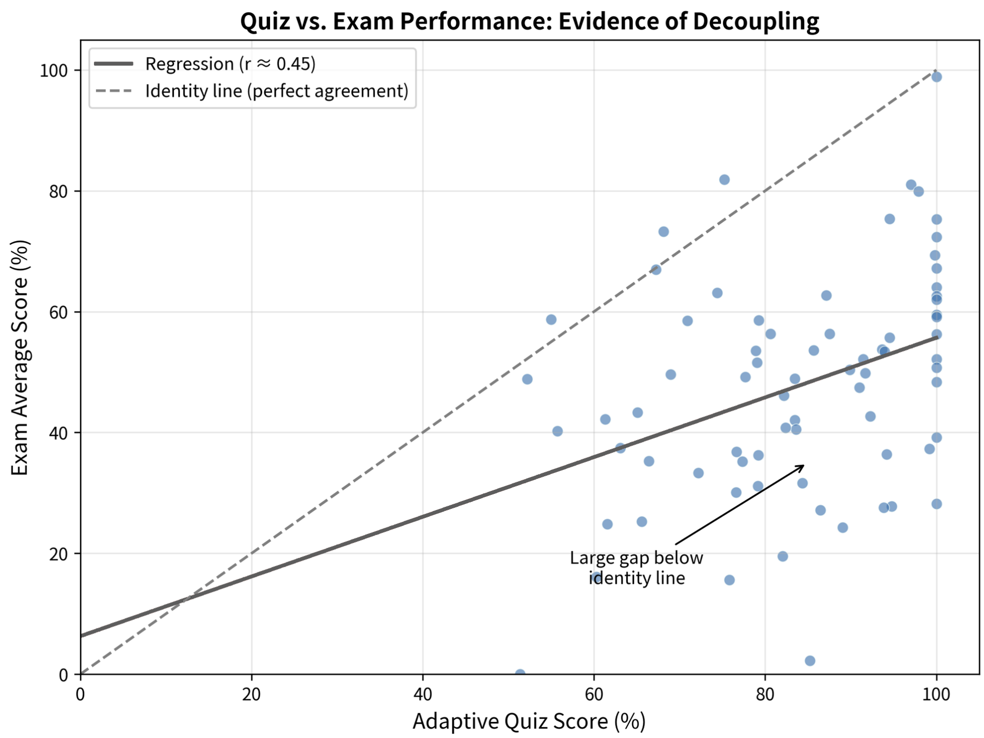
Individual Student Performance Revealed a Systematic Downward Shift

The cohort-level discrepancy remained evident when individual student trajectories were examined. Figure 1 illustrates the relationship between each student's adaptive quiz performance and average examination performance. If adaptive learning accurately reflected examination readiness, observations would cluster closely around the line of identity. Instead, nearly every observation fell below this line, indicating that examination performance was systematically lower than adaptive learning performance for the overwhelming majority of students.

Adaptive quiz scores were also distributed across a narrower range than examination scores, particularly among students in the medium and high achievement bands. Consequently, students who appeared similarly successful within the adaptive environment ultimately exhibited markedly different levels of competence on summative assessments. This systematic downward shift suggests that high adaptive performance alone was insufficient to distinguish students who later demonstrated strong conceptual understanding from those who struggled to transfer their knowledge under traditional examination conditions.

Figure 1

Relationship between Adaptive Learning Performance and Summative Examination Performance



Note. Each point represents one student ($N = 73$). The dashed identity line indicates equal performance across assessment environments. The consistent displacement of observations below the identity line demonstrates that adaptive learning performance systematically exceeded examination performance, illustrating the empirical phenomenon described as the Illusion of Mastery.

Importantly, this pattern reflects a divergence between two commonly used indicators of learning rather than simple academic underachievement. Students demonstrated substantial engagement with the adaptive learning platform and generally achieved high levels of apparent mastery within that environment. Nevertheless, these same students frequently struggled when required to retrieve and apply biological concepts independently during examinations.

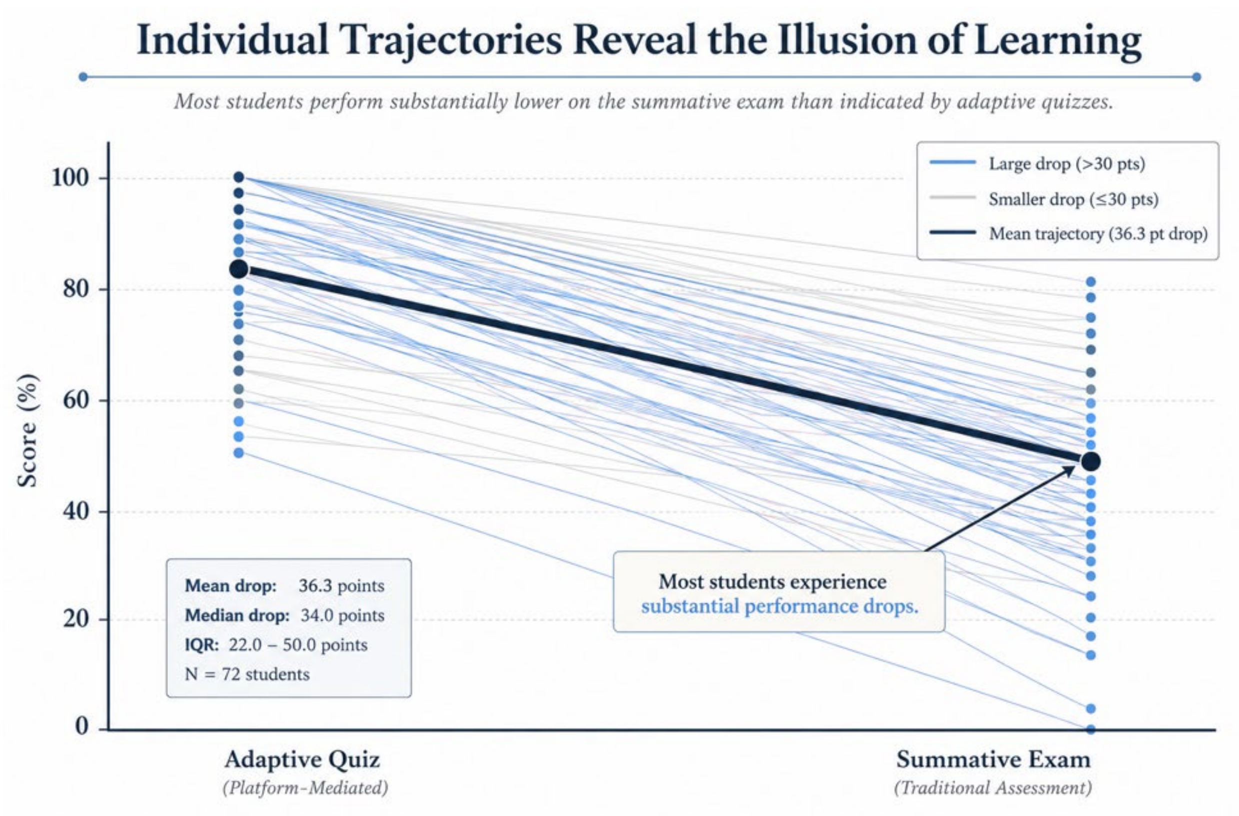
The Calibration Gap Was Consistent Across Individual Students

A complementary perspective is provided by examining within-student performance trajectories (Figure 2). Connecting each student's adaptive quiz score with his or her examination average revealed a remarkably consistent decline from adaptive practice to summative assessment.

Across the cohort, the mean decrease from adaptive learning performance to examination performance was 36.3 percentage points, with a median decline of 34.0 percentage points (interquartile range = 22.0–50.0). These declines were observed throughout the performance distribution and were not restricted to a small subset of academically struggling students.

Figure 2

Individual Student Trajectories from Adaptive Learning to Summative Assessment



Note. Each line connects a student's adaptive quiz average with their corresponding examination average. The bold line represents the cohort mean. Nearly all students experienced substantial declines in performance between the adaptive learning environment and independent examinations, demonstrating that the calibration gap was widespread rather than driven by a small subset of students.

The consistency of these individual trajectories is particularly noteworthy. Nearly every student demonstrated some degree of calibration gap, suggesting that the discrepancy reflects a systematic characteristic of the instructional and assessment environments rather than isolated individual differences. Although the magnitude of the decline varied among students, the overall direction of the effect was strikingly uniform.

Table 2

Descriptive Statistics for Assessment Measures

Measure	*M*	*SD*	Range	*n*
Adaptive quiz (%)	87.8	18.6	18–100	73
Examination 1 (%)	50.4	23.8	0–100	73
Examination 2 (%)	41.8	22.7	10–98	72
Examination 3 (%), scaffolded retake	70.1	25.6	0–98	73
Examination average, Exams 1–3 (%)	54.1	19.2	3–97	73
Examination average, Exams 1–2 (%)	46.1	21.5	5–99	72
Calibration gap, Exams 1–3	33.7	22.0	−17 to 70	73
Calibration gap, Exams 1–2	41.7	21.2	−10 to 72	72

Note. Calibration gap = adaptive quiz percentage minus examination average percentage. Data pool four course sections. Examination 3 was administered as a retake following scaffolded review; rows reporting Exams 1–2 exclude it. Examination 2 was unavailable for one student, so rows based on Exams 1–2 are computed on 72 cases; for the Exams 1–3 average, that student's score is the mean of the examinations available.

The Calibration Gap Increased as Examination Performance Declined

Although nearly all students demonstrated a discrepancy between adaptive practice and examination performance, the magnitude of this discrepancy differed substantially across achievement groups. Students whose examination averages were below 50% exhibited a mean calibration gap of 43.0 percentage points, whereas students whose examination averages exceeded 70% demonstrated a much smaller average gap of 13.6 percentage points. This difference was statistically significant, $t(34) = 5.03$, $p < .001$, Cohen's $d = 1.54$, representing a very large effect. Table 3 summarizes these group differences.

Table 3

Calibration Gap Across Examination Achievement Levels

Achievement band	Exam %	Quiz %	Gap M (SD)	*n*
Low (<50%)	32.1	75.1	43.0 (21.6)	23
Medium (50–70%)	58.9	93.9	35.0 (12.4)	37
High (>70%)	79.4	93.0	13.6 (13.4)	13

Note. Gap = quiz percentage minus exam percentage. Bands are defined by examination average across all three examinations. Low versus High: $t(34) = 5.03$, $p < .001$, $d = 1.54$. Band ns reflect all students with examination data ($N = 73$); figure-based and complete-case analyses use the 72 students with complete data across all eleven modules, which reduces the low band to $n = 22$.

Adaptive Quiz Scores Discriminated Poorly Among Higher-Performing Students

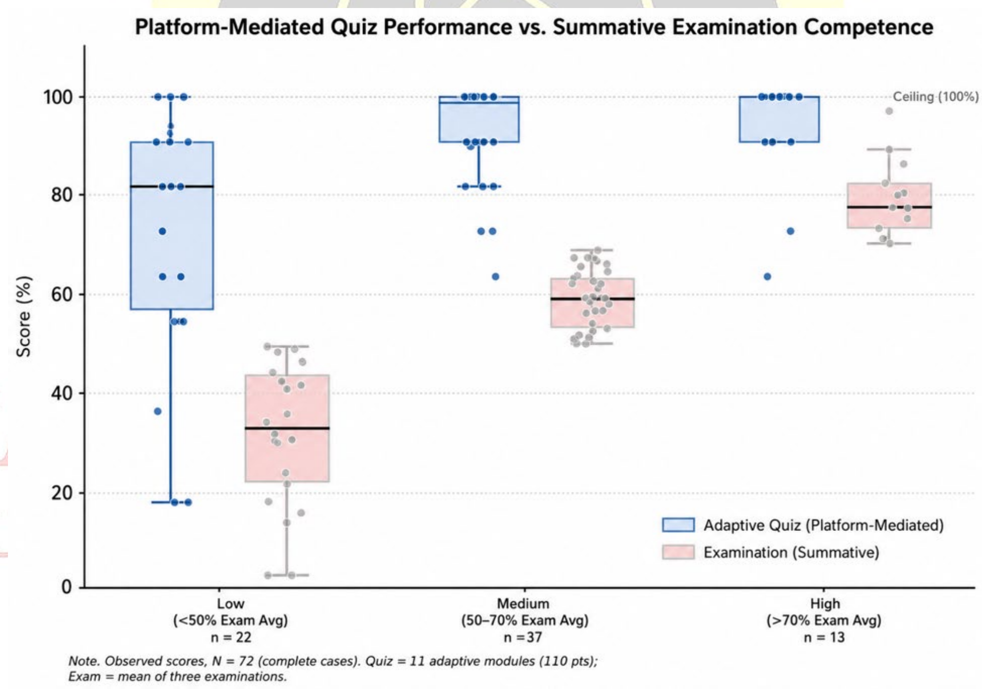
Figure 3 provides an alternative visualization of the calibration gap across achievement groups. Among students in the medium and high achievement bands, adaptive quiz scores were nearly indistinguishable (93.9% and 93.0%, respectively) despite a difference of more than twenty percentage points in examination performance (58.9% and 79.4%). Within this portion of the distribution, adaptive quiz performance provided essentially no information about subsequent examination competence.

This pattern did not hold across the full sample. Students in the low achievement band recorded a substantially lower mean quiz score (75.1%), and adaptive quiz scores across the cohort as a whole spanned a wide range (18–100%, SD = 18.6). The limitation of the adaptive metric is therefore not a uniform ceiling effect but a loss of discriminating power among students performing well within the platform. Because most students clustered in this upper region, adaptive performance was least informative precisely where instructors would most rely on it to identify students who were adequately prepared.

Because examination achievement groups were defined using examination scores, some inflation of between-group differences is expected through regression toward the mean. Nevertheless, the convergence of evidence across cohort-level analyses, individual trajectories, and group comparisons suggests that regression alone cannot account for the magnitude or consistency of the observed calibration gap.

Figure 3

Calibration Gap Across Examination Achievement Groups



Note. Adaptive quiz scores were nearly identical among medium- and high-achieving students, whereas examination performance diverged substantially. The resulting increase in calibration gap among lower-performing students suggests that adaptive learning metrics become progressively less predictive of authentic examination competence as conceptual demands increase.

Transfer-Oriented Examination Tasks Highlighted the Limits of Recognition-Based Performance

The discrepancy between adaptive learning performance and examination performance became especially evident on assessment tasks requiring conceptual transfer rather than factual recognition.

Examination 1: Structural Consequences of Mutation

The first examination required students to explain how a single amino acid substitution alters hemoglobin structure and function in sickle cell disease. Successful performance required students to integrate concepts from molecular genetics, protein chemistry, and structure-function relationships rather than simply recalling isolated facts or definitions. Despite the consistently high adaptive quiz performance preceding this assessment, students achieved a cohort mean of only 50.4%, indicating substantial difficulty applying previously learned concepts to a biologically meaningful problem.

Examination 2: Metabolic Adaptation in Brown Adipose Tissue

A similar pattern emerged on the second examination. Students were required to explain how uncoupling protein 1 (UCP1) alters oxidative phosphorylation during thermogenesis in brown adipose tissue by integrating principles of cellular respiration with physiological regulation. Rather than reproducing memorized metabolic pathways, students were required to reason through an unfamiliar biological context that demanded conceptual transfer. Mean performance declined further to 41.8%, reinforcing the conclusion that adaptive learning success did not consistently translate into independent conceptual reasoning.

Although these examination tasks are presented as illustrative examples rather than independent statistical analyses, they provide concrete demonstrations of the types of higher-order reasoning challenges on which the largest discrepancies between adaptive learning performance and summative assessment became apparent.

Performance Recovered Substantially When Scaffolding Was Reintroduced

Examination 3 was administered as a retake following an instructor-led review session in which the relevant concepts were revisited with substantial scaffolding. Under these conditions, cohort performance rose to 70.1%, compared with 50.4% and 41.8% on the two examinations administered under standard conditions, and the calibration gap narrowed correspondingly from 41.7 to 33.7 percentage points when Examination 3 was included in the average.

This comparison was not experimentally controlled and is subject to several confounds. The retake followed additional exposure to the material, students had previously encountered related content, and both practice effects and increased familiarity with the examination format may have contributed to the improvement. It is therefore not possible to attribute the gain specifically to scaffolding. Nevertheless, the pattern is consistent with the broader argument advanced here: student performance was markedly higher under conditions providing instructional support than under conditions requiring independent retrieval and application of the same concepts.

Summary of Findings

Across four sections of introductory biology, adaptive learning modules consistently produced substantially higher indicators of student success than traditional summative examinations. This discrepancy was evident at multiple levels of analysis, including overall cohort performance, individual student trajectories, and comparisons across achievement groups. Moreover, the divergence between adaptive learning performance and examination performance became increasingly pronounced as assessment tasks required greater conceptual integration and transfer.

Collectively, these findings document a robust and consistent empirical phenomenon. Students demonstrated high levels of success within the adaptive learning environment, yet many struggled to retrieve, integrate, and apply that same knowledge during independent examination conditions. The following section considers whether existing theories of learning adequately explain this discrepancy and examines the Anchoring Hypothesis as one interpretive framework capable of accounting for the observed pattern.

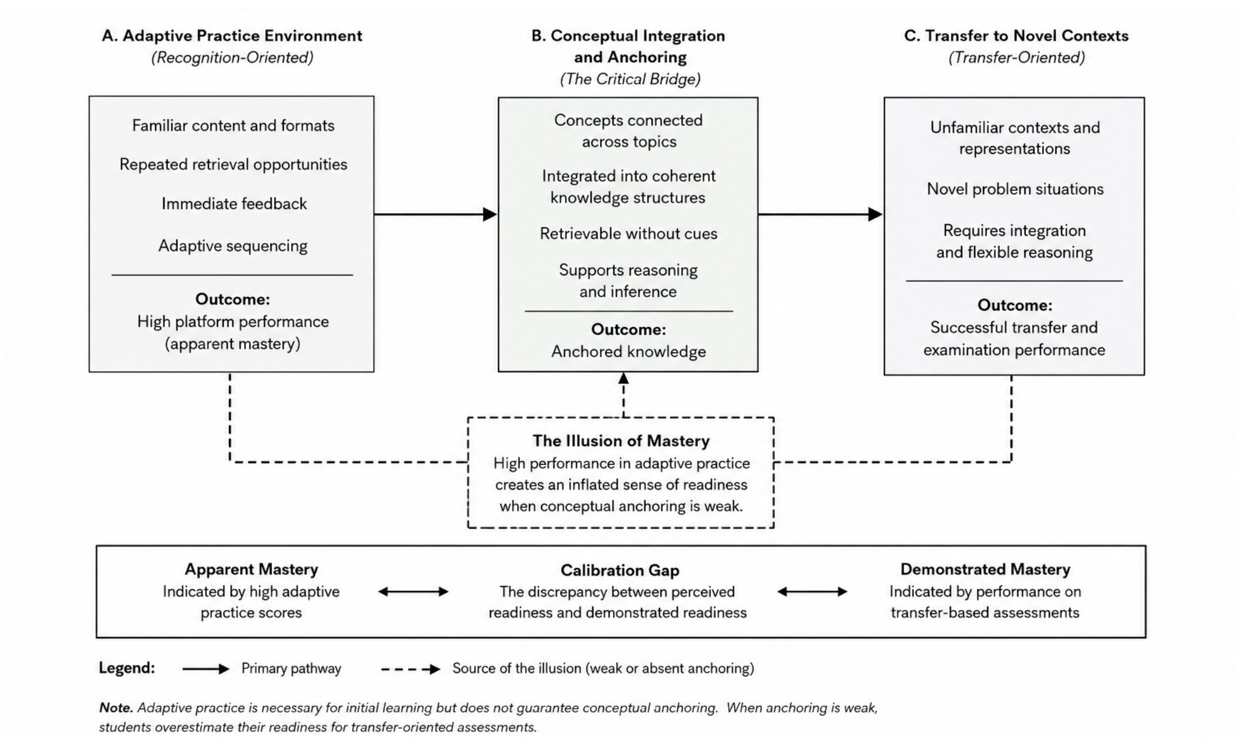
Discussion

The present study documents a robust, consistent, and educationally important discrepancy between students' performance within an adaptive learning environment and their subsequent performance on traditional summative examinations. Across four sections of introductory biology, students consistently achieved high levels of success on adaptive learning modules while performing substantially less well on examinations requiring conceptual application, integration, and transfer. The average calibration gap exceeded thirty percentage points across all three examinations and exceeded forty percentage points when the scaffolded retake was excluded. It was observed across nearly the entire cohort and was largest among students demonstrating the weakest examination performance. Adaptive practice performance was only modestly associated with examination outcomes, accounting for roughly a tenth of the variance in examination scores. Collectively, these findings suggest that adaptive learning environments and traditional examinations measure related, but importantly distinct, dimensions of student learning.

These results are consistent with existing evidence rather than anomalous. In a meta-analytic review of intelligent tutoring systems, Kulik and Fletcher (2016) reported that measured effects were considerably larger on locally developed assessments than on standardized measures, indicating that the apparent benefit of adaptive instruction depends substantially on how learning is assessed. Similarly, Pan and Rickard (2018) demonstrated that the transfer benefits of repeated testing diminish as the distance between practice and assessment conditions increases. More recent experimental work has extended this question to complex educational material, examining how retrieval practice supports not only retention but also the application of concepts to new problems (Corral & Carpenter, 2025). The present findings extend these patterns to an authentic undergraduate biology setting and quantify their magnitude in a single cohort.

Although adaptive technologies have transformed undergraduate instruction by providing individualized practice, immediate feedback, and detailed learning analytics, relatively little attention has been devoted to understanding precisely what adaptive performance represents. High adaptive scores are frequently interpreted by both students and instructors as evidence of mastery. The present findings suggest a more nuanced interpretation. Rather than indicating that students have developed durable conceptual understanding, consistently high adaptive performance may instead reflect successful performance within an instructional environment specifically designed to maximize recognition, retrieval, and repeated success (VanLehn, 2011; Butler, 2010; Roediger & Karpicke, 2006). Consequently, adaptive performance should not automatically be regarded as synonymous with readiness for independent reasoning under authentic assessment conditions.

Figure 4
Conceptual Representation of the Anchoring Hypothesis



Note. Adaptive practice strengthens recognition, retrieval, and procedural fluency, which support performance within the scaffolded environment. Transfer to novel problems is proposed to depend additionally on the integration of these elements into an interconnected conceptual network. The figure is illustrative and does not represent measured variables.

The Anchoring Hypothesis proposes that durable competence depends not simply upon acquiring accurate information, but upon integrating that information into an interconnected conceptual network capable of supporting retrieval, abstraction, reasoning, and transfer. This perspective draws upon foundational theories of meaningful learning (Ausubel, 1968), constructivism (Bruner, 1960), schema development (Chi et al., 1981; Chi, 2009), expertise (Ericsson & Kintsch, 1995), and transfer (Perkins & Salomon, 1989; Barnett & Ceci, 2002; Bransford & Schwartz, 1999). Rather than competing with these established theories, the Anchoring Hypothesis is offered as one way of drawing them together to interpret learning within contemporary adaptive educational environments.

This distinction between performance and learning has long been recognized within cognitive psychology. Bjork and colleagues have argued that instructional conditions producing rapid improvements in performance do not necessarily produce the strongest long-term learning, particularly when learners receive extensive guidance and repeated opportunities for success

(Bjork & Bjork, 2011; Soderstrom & Bjork, 2015). Similarly, Mayer (2002) distinguished rote acquisition from meaningful learning, emphasizing that lasting understanding depends upon organizing knowledge into coherent conceptual structures. The present findings are consistent with these principles and suggest that adaptive learning environments may amplify this distinction, producing highly successful performance during instruction while simultaneously masking incomplete conceptual organization.

Our findings also align with research on metacognitive calibration. Learners frequently mistake familiarity for understanding and fluency for competence (Koriat & Bjork, 2005; Kruger & Dunning, 1999; Dunlosky et al., 2013; Jacoby, 1991). Adaptive learning environments may unintentionally reinforce this illusion because repeated exposure, corrective feedback, and adaptive sequencing progressively increase response accuracy. Consequently, students receive increasingly positive signals regarding their apparent mastery even when their conceptual understanding remains insufficient for successful transfer to novel biological problems.

The present findings may therefore have implications extending beyond adaptive learning itself. Increasingly, educational technologies generate detailed analytics that promise to estimate student mastery, predict future performance, and personalize instruction. These systems are only as valid as the constructs they measure, and learning analytics research has long cautioned that indicators of engagement and activity within a platform should not be equated with evidence of learning (Gašević et al., 2015). If adaptive performance primarily reflects recognition-based success rather than durable conceptual understanding, then learning analytics derived from these systems may systematically overestimate student readiness for authentic assessment. The Anchoring Hypothesis suggests that future learning analytics should incorporate measures of conceptual integration, transfer, and knowledge organization in addition to conventional indicators of accuracy and completion.

The biology education literature further supports this perspective. Scientific understanding depends not merely upon recalling isolated biological facts but upon integrating concepts across multiple levels of biological organization, from molecules and cells to organisms and ecosystems. Conceptual understanding has therefore become a central emphasis within undergraduate biology education (Cary et al., 2019; Freeman et al., 2014; Handelsman et al., 2004). The examination tasks described in the present study were designed to assess these forms of integrative reasoning, requiring students to explain unfamiliar biological phenomena rather than reproduce memorized information. The disproportionately large calibration gaps observed on these tasks suggest that conceptual integration may represent the principal point at which adaptive learning performance and authentic biological reasoning diverge.

Importantly, the present study should not be interpreted as evidence against adaptive learning. On the contrary, retrieval practice, immediate feedback, and individualized instruction remain among the most strongly supported instructional strategies in cognitive science (Roediger & Karpicke, 2006; Karpicke & Roediger, 2008; Butler, 2010; Karpicke & Blunt, 2011). Retrieval practice may also confer benefits beyond the material directly tested, enhancing the learning of subsequently studied content through what has been described as the forward effect of testing (Pastötter & Bäuml, 2014). Adaptive technologies almost certainly improve learning relative to many traditional instructional approaches. Rather, the present findings suggest that success within adaptive learning environments represents one important component of learning rather than its complete expression. Recognition, retrieval, and procedural fluency constitute necessary foundations upon which deeper conceptual understanding is constructed, but they are not themselves sufficient evidence of transferable competence.

Limitations

Several limitations should be acknowledged. Because this investigation was observational, adaptive modules and summative examinations differed simultaneously in multiple characteristics, including assessment stakes, opportunities for repeated practice, immediacy of feedback, and cognitive complexity. Consequently, the present study cannot isolate the independent contribution of any single instructional feature.

Most importantly, the two assessment formats also differed in the kind of processing they required. Research on transfer-appropriate processing has long established that performance is highest when the cognitive operations required at test match those engaged during learning (Morris et al., 1977). A discrepancy between cue-supported practice and unsupported examination is therefore predicted by this account alone, without recourse to any construct of conceptual anchoring. The present data cannot distinguish between these explanations. We note only that transfer-appropriate processing does not obviously predict the graded pattern observed across achievement bands, in which the calibration gap was more than three times larger among low-performing students than among high-performing students; this differential remains to be explained and represents a useful target for future research.

In addition, the examinations used in this study were instructor-developed and were not subjected to formal psychometric evaluation. Item difficulty, discrimination indices, and internal-consistency reliability were not computed, and no independent classification of items by cognitive level was performed. The characterization of the examinations as transfer-oriented and the adaptive modules as recognition-oriented therefore rests on the instructional design of the two environments rather than on verified item-level evidence. Because examination means were low, particularly on Examination 2, the possibility that examination difficulty contributed to the observed gap cannot be excluded. Independent classification of items using established instruments for the cognitive analysis of biology assessments represents an important next step.

Students' motivation, study strategies, and test anxiety were not directly measured and may have contributed to variation in examination performance; the difference in stakes between low-stakes practice and graded examinations is a plausible contributor to the gap in its own right. Because achievement groups were defined using examination scores, regression toward the mean remains a potential source of bias. Finally, the sample was drawn from a single course at a single institution, and the generalizability of these findings to other disciplines, institutions, or adaptive platforms remains to be established. Although these considerations do not appear sufficient to explain the magnitude or consistency of the observed calibration gap, they underscore that the Anchoring Hypothesis should presently be regarded as an interpretive framework rather than a confirmed causal mechanism.

Directions for Future Research

The hypothesis nevertheless suggests several testable directions. First, adaptive learning environments that explicitly require conceptual integration would be expected to produce smaller calibration gaps than systems emphasizing repeated recognition and retrieval. Evidence from physics education indicates that making practice assessments more closely resemble the conditions of the summative examination can alter both study behavior and examination performance (Zhang et al., 2023), suggesting that calibration gaps of the kind reported here may be responsive to instructional design rather than fixed. Second, independent measures of knowledge organization, including concept inventories, concept mapping, and network analyses, would be expected to correlate more strongly with examination performance than adaptive accuracy alone. Third, incorporating indicators of conceptual anchoring into adaptive learning analytics should improve prediction of transfer-oriented assessment performance. Finally, interventions designed to strengthen conceptual connections rather than to increase practice volume would be expected to produce greater improvements in authentic assessment than equivalent increases in adaptive practice alone. This expectation is consistent with evidence that the structure and quality of practice, rather than its sheer quantity, governs the development of expert performance (Ericsson et al., 1993).

Each of these predictions is capable of disconfirming the framework. If independent measures of knowledge organization fail to predict examination performance beyond adaptive accuracy, or if interventions targeting conceptual integration produce no advantage over equivalent additional practice, the Anchoring Hypothesis in its present form would not be supported.

Conclusion

Taken together, the present findings describe a phenomenon that many instructors recognize intuitively but that has received comparatively little systematic documentation in undergraduate biology. Across four sections of introductory biology, adaptive learning environments consistently generated substantially more optimistic estimates of student mastery than were subsequently demonstrated on transfer-oriented examinations. We refer to this discrepancy as the Illusion of Mastery and propose the Anchoring Hypothesis as one framework for interpreting why it occurs. Whether this interpretation ultimately proves correct remains an empirical question. For instructors, the immediate implication is more direct: performance within an adaptive learning platform should be treated as evidence of engagement and practice rather than as a sufficient indicator of readiness for assessments requiring independent reasoning.

References

Akbulut, Y. (2024). Impact of different practice testing methods on learning outcomes. European *Journal of Education, 59*, e12626.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart & Winston.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*, 61–100.

Bruner, J. S. (1960). *The process of education.* Harvard University Press.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118–1133.

Cary, T. L., Wienhold, C. J., & Branchaw, J. (2019). A biology core concept instrument (BCCI) to teach and assess student conceptual understanding. *CBE—Life Sciences Education, 18*(3), ar46.

Chi, M. T. H. (2009). Active–constructive–interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471–533.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Contrino, M. F., Reyes-Millán, M., Vazquez-Villegas, P., & Membrillo-Hernández, J. (2024). Using an adaptive learning tool to improve student performance and satisfaction in online and face-to-face education for a more personalized approach. *Smart Learning Environments, 11*(1).

Corral, D., & Carpenter, S. K. (2025). Effects of retrieval practice on retention and application of complex educational concepts. *Learning and Instruction, 100*, 102219.

Deimeke, E., Teasdale, A., Arthur, R., Gray-Singh, D., & Talpade, M. (2024). Application of growth mindset across disciplines on student academic performance. *Journal of Instructional Pedagogies*, 30.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58.

Du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon, 10*(21), e39630.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211–245.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10*(6), 597–602.

Gray-Singh, D., Arthur, R., Deimeke, E., Mitchell, M., Banks, J., Tolokun, O., Rucker, S., & Talpade, M. (2025). Impact of infusing artificial intelligence in the curriculum across disciplines. *Journal of Instructional Pedagogies*, 31.

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., Gentile, J., Lauffer, S., Stewart, J., Tilghman, S. M., & Wood, W. B. (2004). Scientific teaching. *Science, 304*(5670), 521–522.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513–541.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772–775.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 187–194.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19(*6), 585–592.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78.

Lipnevich, A. A., Janelli, M., Park, M. J., Calistro, T., & Esser, F. J. (2025). The effects of practice testing and feedback on learners' performance and persistence in a massive open online course. *Journal of Research on Technology in Education, 57*(6), 1390–1409.

Mayer, R. E. (2002). Rote versus meaningful learning. *Theory Into Practice, 41*(4), 226–232.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–533.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756.

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology, 5*, 286.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18*(1), 16–25.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Sheaves, C., Froble, D., Morton, K., & Parcell, H. (2022). Evaluating the effect of pre-exam adaptive quizzing on nursing student exam scores. *Nursing Education Perspectives, 43*(6), E100–E102.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199.

Sweller, J. (1988). Cognitive load during problem-solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Wang, H., Tlili, A., Huang, R., Cai, Z., Li, M., Cheng, Z., Yang, D., Li, M., Zhu, X., & Fei, C. (2023). Examining the applications of intelligent tutoring systems in real educational contexts: A systematic literature review from the social experiment perspective. *Education and Information Technologies, 28*, 9113–9148.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441–517.

Zhang, M., Morphew, J., & Stelzer, T. (2023). Impact of more realistic and earlier practice exams on student self-regulated studying behaviors and exam performance. *Physical Review Physics Education Research, 19*(1), 010130.

Appendix A

Author Contributions

This study originated as a collaborative faculty–student research project conducted within an undergraduate biology course. Student authors contributed to data organization, statistical analysis, interpretation of findings, literature review, figure preparation, and manuscript development. They participated actively in discussions that shaped the conceptual development of the Anchoring Hypothesis and contributed to multiple rounds of manuscript revision.

The faculty author conceived and supervised the study, designed the instructional environment, curated and validated the course-performance dataset, developed the theoretical framework, performed the primary interpretation of results, and led manuscript preparation and revision. All authors contributed intellectually to the work, critically reviewed the manuscript for important intellectual content, approved the final version for publication, and agree to be accountable for all aspects of the study.

Ethics Statement and Data Availability

This study analyzed de-identified educational records collected from students enrolled in four sections of an introductory General Biology course. The investigation was conducted in accordance with institutional policies governing research involving human participants. All student records were fully de-identified prior to analysis, and no personally identifiable information is reported. Because the study was conducted using existing educational records collected during normal instructional activities, no additional interventions were performed for research purposes. The analyses reported here did not influence student grades or academic standing. The de-identified dataset and associated analytical code supporting the findings of this study will be made available by the corresponding author upon reasonable request, subject to institutional policies governing student educational records and applicable privacy regulations.