

Article

Assessment Validity in the Age of Generative AI: A Natural Experiment

Håvar Brattli ^{1,*} , Alexander Utne ¹  and Matthew Lynch ² ¹ School of Business and Economics, UiT The Arctic University of Norway, 9010 Tromsø, Norway; alexander.utne@uit.no² Faculty of Social Sciences, School of Business, Oslo Metropolitan University, 0166 Oslo, Norway; matthew.lynch@oslomet.no

* Correspondence: havar.brattli@uit.no

Abstract

Universities play a dual role as sites of learning and as institutions that certify student competence through assessment. The rapid diffusion of generative artificial intelligence (GenAI) challenges this certification function by altering the conditions under which assessment evidence is produced. When powerful AI tools are widely available, grades may increasingly reflect a combination of individual understanding and external cognitive support rather than solely independent competence. This study examines how changes in assessment format interact with GenAI availability to reshape observable performance outcomes in higher education. Using exam grade data from a compulsory undergraduate course delivered over five years (2021–2025; $N = 1066$), the study exploits a naturally occurring change in assessment conditions as a natural experiment. From 2021 to 2024, the course was assessed using an AI-permissive take-home examination, while in 2025 the assessment shifted to an AI-restricted, supervised in-person examination. Course content, intended learning outcomes, grading criteria, examiner continuity, and the structural design of the examination tasks remained stable across cohorts. The results reveal a pronounced shift in grade distributions coinciding with the format change. Failure rates increased sharply in 2025, mid-range grades declined, and the proportion of top grades remained largely unchanged. Statistical analysis indicates a significant association between examination period and grade outcomes ($\chi^2(5, N = 1066) = 60.62, p < 0.001$), with a small-to-moderate effect size (Cramér's $V = 0.24$), driven primarily by the increase in failing grades. These findings suggest that AI-permissive and AI-restricted assessment formats may not be measurement-equivalent under conditions of widespread GenAI use. The results raise concerns about construct validity and the credibility of grades as signals of independent competence, while also highlighting tensions between certification credibility and assessment authenticity.



Academic Editors: Amir Ghapanchi,
Reza Ghanbarzadeh and
Purarjomandlangrudi Afrooz

Received: 19 January 2026

Revised: 25 March 2026

Accepted: 2 April 2026

Published: 3 April 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: generative artificial intelligence; assessment validity; grade distributions; higher education assessment; credential signaling

1. Introduction

Universities function not only as sites of learning but also as institutions that certify student competence. Through assessment, universities signal whether students who pass have reached a minimum level of ability required for subsequent study, professional practice, and participation in the labor market. This certification function depends on a core assumption: that assessment outcomes, typically expressed as grades, provide a reasonable

indication of what students can do independently [1,2]. When this assumption is weakened, grades become harder to interpret, undermining both the credibility of academic credentials and the institutional role of the university as a certifying authority [3,4].

The rapid diffusion of generative artificial intelligence (GenAI), particularly large language models such as ChatGPT, has begun to unsettle this premise. Since late 2022, these tools have become widely accessible, increasingly embedded in students' everyday study practices, and difficult to detect reliably in submitted work [5–7]. As a result, it has become easier to produce assessment outputs that meet conventional grading criteria while obscuring the extent to which observed performance reflects individual understanding rather than external cognitive support. This development poses a fundamental challenge to higher education assessment by calling into question what grades actually certify when powerful cognitive assistance is readily available.

Parts of the emerging literature on AI and assessment have shifted the focus away from cheating and misconduct toward concerns about assessment validity and grade credibility. The central issue is the erosion of stable links between observed assessment performance and the underlying competencies institutions claim to measure [3,8]. From this perspective, the challenge is institutional rather than individual: if assessment systems cannot generate defensible inferences about competence under conditions of widespread AI access, universities' claims about the meaning and reliability of grades become increasingly difficult to sustain [4,9].

These validity concerns are not merely theoretical. GenAI is already widely used in assessment-related tasks such as idea generation, drafting, coding, and paraphrasing, often in strategically instrumental ways aimed at successful task completion rather than learning [5,10]. Importantly, the performance effects of AI appear to be uneven. Meta-analytic evidence suggests that lower-performing students tend to benefit more from AI support than higher-performing peers, who already possess the skills required to meet task demands [11]. This asymmetry has distributional consequences. By lowering the cognitive effort required to meet minimum performance standards, AI use can compress grade distributions and reduce differentiation near pass thresholds, precisely where assessment outcomes carry the greatest credentialing significance [11]. Under AI-permissive conditions, some students may therefore cross pass thresholds without demonstrating the level of independent competence that grades are intended to certify, weakening the signal value of academic credentials.

In response to the broader challenges that widespread AI use poses for assessment validity, credential credibility, and institutional certification, universities have experimented with governance strategies including AI detection tools, revised integrity policies, re-designed assessments, and renewed emphasis on AI-restricted in-person examinations [8,9]. Restricting access to AI during assessment increases the likelihood that observed performance reflects unaided competence. At the same time, such restrictions raise concerns about authenticity and relevance. In many disciplines, graduates will be expected to use AI and other digital tools as part of professional practice. Assessments that systematically prohibit these tools risk evaluating students under artificially constrained conditions [6,8].

Despite the intensity of this debate, there is a shortage of empirical studies that isolate how changes in assessment conditions, specifically the availability or restriction of GenAI, affect observable performance outcomes. Much of the existing literature is conceptual, policy-oriented, or based on perceptions and self-reports [6,7,9]. While valuable, such approaches struggle to disentangle the effects of AI access from other factors such as curriculum changes, teaching practices, grading standards, or cohort composition.

The present study addresses this gap by exploiting a naturally occurring change in examination format within a compulsory undergraduate course delivered over multiple

years. From 2021 to 2024, the course was assessed using a take-home, open-resource examination format with access to digital tools, including GenAI. In 2025, the examination was transitioned to an in-person closed-book format conducted under supervised conditions. Crucially, this change occurred while course content, intended learning outcomes, grading criteria, examiner continuity, and the structural design of the examination tasks remained stable across cohorts. While we acknowledge the presence of potentially confounding variables such as exam anxiety or cohort differences, the natural experiment allows an examination of how restricting GenAI access during assessment coincides with changes in grade distributions and failure rates under similar conditions to prior years.

Accordingly, the study addresses the following research question: How does the transition from an AI-accessible assessment format to an AI-restricted in-person assessment format coincide with changes in grade distributions and failure rates in a higher-education course during a period of rapid generative AI adoption? By anchoring concerns about assessment validity, measurement equivalence, and credential signaling in concrete empirical data, the article aims to move the AI-and-assessment debate beyond speculation and toward evidence that can inform institutional decision-making.

2. Theory

Building on the concerns outlined in the Introduction, this section develops a theoretical framework for understanding how GenAI reshapes assessment validity, performance distributions, and the signaling function of grades.

2.1. *Assessment as a Measurement System*

In higher education, assessment functions as a socio-technical measurement system that produces observable indicators such as scores, grades, and classifications, which are used to judge student learning, competence, and progression [3,4]. In contemporary measurement theory, validity is not understood as a fixed property of an assessment tool. Instead, it refers to the extent to which assessment results justify the claims made about students' learning and the decisions that follow from those results. This requires evidence across content, response processes, internal structure, relations to other variables, and consequences [1,2,4]. From this perspective, assessment outcomes are stabilized by an interconnected set of practices: task design, marking and moderation, institutional quality assurance, and regulatory standards [4,12].

The rapid adoption of GenAI represents a structural change in this system rather than a marginal pedagogical innovation. AI-enabled platforms are increasingly embedded in assessment infrastructure through automation, scalable feedback, evaluation, and monitoring [12,13]. At the same time, institutional and regulatory frameworks continue to rely on criteria calibrated for pre-AI conditions, including assumptions about authorship, independence of work, traceability of evidence, and stable interpretation of performance [3,12]. Kaldaras et al. caution that when AI systems mediate or generate assessment evidence, established validity arguments may not hold unless explicit evidence demonstrates that the same constructs are being measured under equivalent conditions [4].

From a measurement standpoint, a central risk is that GenAI introduces systematic sources of variance that are not modeled within existing assessment designs or moderation practices, weakening inferences drawn from assessment results [1,2,4]. Marker uncertainty is a key mechanism. Experimental evidence suggests that markers are generally unable to reliably distinguish between student-authored, AI-modified, and AI-generated submissions [14]. Detection uncertainty can undermine the evidentiary basis of grading decisions and alter marker behavior, increasing second-guessing of authorship and widening inter-

marker variability [14]. These effects introduce additional measurement error and weaken the interpretability of outcomes [4,14].

These risks extend beyond isolated cases of misconduct. Survey evidence indicates broad concern among educators that GenAI challenges the sustainability of conventional written and coding assessments, and that output-based artifacts alone may no longer be reliable indicators of intended constructs [15]. Taken together, the literature suggests that GenAI destabilizes assessment not primarily by increasing integrity breaches, but by altering the relationship between learning activity, observable performance, and evaluative judgment within the measurement system [3,4,12].

2.2. AI as a Low-Cost Cognitive Substitute and Performance Mediator

Building on this measurement perspective, GenAI can also be understood as reshaping the cognitive production layer that links student learning activity to observable assessment artifacts. Rather than functioning only as a productivity aid, GenAI increasingly operates as a low-cost cognitive substitute: it can perform substantial portions of ideation, structuring, language generation, problem decomposition, and synthesis that learners previously executed internally. Empirical evidence suggests that GenAI influences multiple dimensions of cognition, including critical thinking, creative thinking, reflective thinking, computational reasoning, and problem solving, with both positive and negative effects depending on task design and patterns of use [16].

From a cognitive systems perspective, this substitution dynamic is consistent with cognitive offloading, where individuals delegate memory, computation, or reasoning to external artifacts to reduce internal load. Offloading can enhance short-term performance and efficiency, but overuse can reduce deeper encoding, retention, and independent problem-solving capacity [17–19]. Experimental work also shows that individuals adjust offloading strategically in response to performance goals and task demands, but persistent reliance can reshape how cognitive resources are allocated over time [20].

Recent educational evidence suggests that GenAI amplifies the scale and accessibility of cognitive offloading. Students use AI to externalize planning, drafting, revision, and analytical processing, often enabling faster task completion and higher immediate output quality [21]. Structural equation modeling shows that GenAI use predicts cognitive offloading and shared metacognitive regulation, which may mediate short-term academic performance gains [21]. These findings indicate that AI systems become embedded as active cognitive partners in learners' task execution processes.

At the same time, cognitive substitution introduces risks for construct validity, understood here as the extent to which assessment outcomes reflect the underlying competence they are intended to capture. When a significant proportion of processing is delegated to AI systems, the resulting artifact reflects a hybrid of human and machine cognition rather than the learner's internal competence alone. Empirical work links frequent AI use to increased perceived efficiency and confidence alongside strengthened technological dependency, suggesting that competence and tool support can become entangled [22]. Related evidence on technology dependence and learning highlights the risk that reliance on external supports may weaken autonomous reasoning and critical thinking self-efficacy over time [23,24]. Systematic reviews reinforce this dual effect: structured use can support cognitive skills, while habitual reliance is associated with weaker independent engagement and metacognitive monitoring [16].

GenAI also affects the social and regulatory dimensions of learning. AI-supported environments can reshape shared metacognition, collaborative regulation, and strategy selection in group learning contexts, influencing how learners jointly plan, monitor, and evaluate cognitive work [21,25,26]. While these mechanisms can support learning under

intentional design, they add further layers of mediation between individual cognition and assessment evidence, complicating attribution of individual competence.

2.3. *Asymmetric Benefit Distribution, Variance Compression, and Grade Instability*

If GenAI operates as a low-cost cognitive substitute that mediates performance production, its effects on assessment outcomes are unlikely to be evenly distributed across students or tasks. Instead, substitution interacts with baseline competence, self-regulation capacity, and task structure, producing asymmetric gains that reshape grade distributions rather than uniformly improving learning outcomes. Empirical studies show that the impact of GenAI on performance varies across usage patterns and learner profiles, with some engagement associated with improved short-term performance and other patterns associated with stagnation or decline [27–29].

One consistent pattern is that GenAI lowers the cognitive cost required to reach acceptable performance on structured tasks such as summarization, drafting, coding scaffolds, or procedural reasoning. Comparative studies suggest that AI-supported conditions can raise immediate performance and perceived efficiency [27,29,30]. These effects may disproportionately benefit students who previously struggled with organization, language production, or procedural fluency, narrowing observable gaps at the lower end of grade distributions.

However, convergence in observed performance does not necessarily imply convergence in underlying competence. Diaz et al. show that behaviorist patterns of AI use, where AI is treated primarily as a task-completion tool, are associated with weaker learning outcomes [27]. Similarly, Vanacore et al. find that generative feedback can improve performance for certain error types while simultaneously undermining self-regulated learning in more complex contexts [28]. Taken together, these findings suggest that performance gains enabled by AI substitution may mask deterioration in metacognitive regulation, strategic reasoning, and independent problem solving.

At the distributional level, these dynamics can lead to variance compression in observed performance. When AI-supported work reduces differences in structure, language, or problem framing, student submissions may become more similar in form and apparent quality. Under such conditions, small variations in marker judgment or task interpretation can have a disproportionate influence on grade outcomes. Reduced discriminability among performances thus weakens the stability of grade distributions and increases sensitivity to contextual and procedural noise.

These effects also connect to credential signaling. Grades are institutional signals used to communicate competence to employers and downstream institutions, and their credibility depends on differentiating underlying capability in ways that cannot be easily replicated through external mediation [31]. AI-focused evidence suggests grade compression and altered performance distributions may already be emerging in real settings [11], but naturalistic studies tracing changes in outcomes as assessment conditions shift remain limited.

3. Method

This study adopts a natural experiment design based on administrative grade data from a compulsory undergraduate course in Norway, collected over five academic years (2021–2025). Natural experiments arise when an institutional change alters a key condition of interest while other relevant features remain broadly stable, allowing for plausible causal inference without random assignment [32,33]. In the present case, the key change is the examination format, while course content, intended learning outcomes, grading standards, and the structural design of the examination tasks remained substantively unchanged.

The assessment data derive from BED2302 Organizational Theory and Leadership, a compulsory 7.5-credit undergraduate course at UIT School of Business and Economics. The course provides foundational knowledge in organization and leadership and emphasizes analysis, theory application, and academic writing. Examinations require students to identify and analyze organizational problems, apply theoretical concepts to cases, and justify claims in a clear, reasoned academic argumentation.

From 2021 to 2024, the course was assessed using a take-home, open-resource written exam with access to external materials, including digital resources and AI-based tools. In 2025, the exam shifted to a supervised, in-person, closed-book format without access to external aids or internet-enabled tools. This change was driven by institutional assessment policy in response to the widespread availability of GenAI, rather than by curricular revision.

While specific exam questions necessarily varied across years, they were designed to remain comparable in structure, cognitive level, and alignment with learning outcomes. Across all cohorts, the examinations assessed the same core competencies, including the identification and analysis of organizational problems, the application of theoretical concepts to case material, and the construction of reasoned academic arguments. See Table 1 for sample exam questions and grading criteria.

Table 1. Sample questions and main grading criteria from 2024 to 2025 exam.

Sample Questions and Main Grading Criteria	2024 Exam (Take-Home Format)	2025 Exam (In-Person Exam Format)
Sample question	Discuss strategy and efficiency in [case company] in light of the challenges the company faces in its environment. Use theories and models from the course literature to justify your answer.	Explain the concept of learning and discuss what can be done to strengthen learning in [case company], in light of relevant theories and models from the syllabus.
Sample question	[Case company] faces challenges related to the practice of leadership. Identify some of these challenges and discuss how leadership practice can be developed to address them. Use theories and models from the course literature to justify your answer.	Explain the concept of leadership and discuss how leadership can be exercised to ensure continued strong performance in [case company], in light of relevant theories and models from the syllabus.
Sample question	Discuss how [case company] takes care of its employees in terms of motivation and performance, and what can be done to promote motivation and retain employees. Use theories and models from the course literature to justify your answer.	Discuss factors that may promote and hinder employee motivation in [case company], in light of relevant theories and models from the syllabus.
Main grading criteria	For all questions, the case must be actively incorporated and discussed in the response. To pass, students are required to apply theories and models from the syllabus when discussing and analyzing various situations in the organization. It should be clearly evident that they are “consulting” the course literature. Responses in which the presentation of relevant theories and models from the syllabus is entirely absent will be considered a fail.	For all questions, the case must be actively incorporated and discussed in the response. To pass, students are required to apply theories and models from the syllabus when discussing and analyzing various situations in the organization. It should be clearly evident that they are “consulting” the course literature. Responses in which the presentation of relevant theories and models from the syllabus is entirely absent will be considered a fail.

The course was delivered across multiple campuses under a shared syllabus and centrally coordinated materials. The main examiner and lead grader remained constant across the study period, and the same grading scale and general grading standards were applied across cohorts. The examiner team also explicitly discussed the potential risk of subconsciously penalizing the 2025 cohort for not producing the level of polished prose typically associated with take-home exams, and took this into account during the grading process to ensure fair and consistent evaluations.

The dataset comprises final grades on an A–F scale. Cohort sizes ranged from 179 to 263 students.

3.1. Analytical Strategy

To distinguish ordinary year-to-year variation from a potential structural shift associated with the assessment format change, grade distributions from the four take-home examination years (2021–2024) were aggregated to form a historical baseline distribution. The years 2021–2024 were combined because they shared the same examination format and showed broadly similar grade distributions, providing a stable reference for the pre-change assessment regime. Aggregation was performed by summing the number of students receiving each grade (A–F) across these cohorts, yielding a reference distribution representing typical outcomes under the open-resource examination regime.

This baseline distribution was then compared with the grade distribution observed in 2025 following the introduction of the in-person, closed-book examination. The comparison focuses on distributional changes rather than individual-level trajectories, consistent with quasi-experimental approaches that infer effects from discontinuities coinciding with institutional changes while holding other conditions constant [32,33].

3.2. Statistical Analysis

Differences in grade distributions between the aggregated baseline period (2021–2024) and 2025 were examined using chi-square tests of independence, with individual student grades as the unit of analysis. This test evaluates whether grade outcomes are statistically independent of examination period or whether observed frequencies deviate systematically from expected frequencies derived from the marginal totals. The assumptions for the chi-square analyses were met: observations were independent, grade categories were mutually exclusive, and expected cell frequencies were sufficient in all reported comparisons. To complement significance testing, effect sizes were estimated using Cramér's V, providing a standardized measure of the strength of association between examination period and grade distribution.

Although the natural experiment design strengthens internal validity relative to purely cross-sectional comparisons, the study remains observational in nature. In line with established principles of causal inference, findings are therefore interpreted as evidence of association and plausibility rather than definitive causal attribution [34]. For chi-square tests, degrees of freedom were determined by the number of grade categories minus one ($df = 5$), and statistical significance was evaluated using a conventional threshold of $p < 0.05$, with exact p -values reported.

3.3. Ethics and Use of AI

The study was reviewed by a representative for the Research Ethics Committee at UIT School of Business and Economics, which confirmed that formal approval was not required because the study used retrospective anonymized data from ordinary course activities. Data handling was reviewed through [redacted] and deemed not to require a data management plan. Participants' identities were anonymized. The authors used ChatGPT (OpenAI) for language refinement and readability improvements. All decisions

regarding content selection, analysis, interpretation, and conclusions were made by the authors, who retain full responsibility for the article.

4. Results

4.1. Descriptive Grade Distributions

Table 2 presents grade distributions for the course from 2021 to 2025. Grade distributions were relatively stable from 2021 to 2024, with failure rates ranging from approximately 3% to 6% across these years. During this period, the relative proportions of grades A through E exhibited only minor fluctuations.

Table 2. Observed grades 2021–2025.

Grade	2021 (N = 263)	2022 (N = 230)	2023 (N = 211)	2024 (N = 187)	2025 (N = 179)
A	18 (6.8%)	16 (7.0%)	19 (9.0%)	13 (7.0%)	13 (7.3%)
B	56 (21.3%)	53 (23.0%)	53 (25.1%)	46 (24.6%)	30 (16.8%)
C	85 (32.3%)	68 (29.6%)	59 (28.0%)	58 (31.0%)	34 (19.0%)
D	59 (22.4%)	50 (21.7%)	45 (21.3%)	45 (24.1%)	40 (22.4%)
E	34 (12.9%)	32 (13.9%)	23 (10.9%)	20 (10.7%)	29 (16.2%)
F	8 (3.0%)	11 (4.8%)	12 (5.7%)	4 (2.1%)	33 (18.4%)
Total	263 (100%)	230 (100%)	211 (100%)	187 (100%)	179 (100%)

In contrast, the 2025 distribution differs markedly from the preceding years. The failure rate increased sharply to 18.4%, accompanied by a reduction in mid-range grades (B and C) and a modest increase in grade E. The proportion of top grades (A) remained largely unchanged.

4.2. Baseline Versus 2025 Comparison

To test whether the 2025 grade distribution differed systematically from the aggregated 2021–2024 baseline, a chi-square test of independence was conducted. The test indicated a statistically significant association between examination period and grade distribution, $\chi^2(5, N = 1066) = 60.62, p < 0.001$, with a Cramér's V of 0.24. According to conventional benchmarks, this corresponds to a small-to-moderate association [35].

Figure 1 contrasts the observed 2025 grade frequencies with the frequencies expected if the 2025 cohort had followed the same grade distribution as the baseline period. The expected values therefore serve as a reference point for assessing whether the 2025 grade distribution differs from the baseline pattern. The chi-square statistic is based on the extent to which the observed and expected frequencies diverge across grade categories. The largest deviations were observed for failing grades (F), which occurred substantially more frequently than expected, while mid-range grades (B and C) occurred less frequently than expected. Grades A and D showed only minor deviations from expected values.

4.3. Category-Specific Contributions to the Chi-Square Statistic

Table 3 decomposes the overall chi-square statistic into category-specific contributions for the baseline period and 2025. The majority of the chi-square statistic is attributable to failing grades (F), particularly in 2025, indicating that the increase in failures accounts for most of the statistically significant association between examination period and grade distribution. Contributions from other grade categories are comparatively small, suggesting that changes outside the failure category played a more limited role in producing the observed distributional difference.

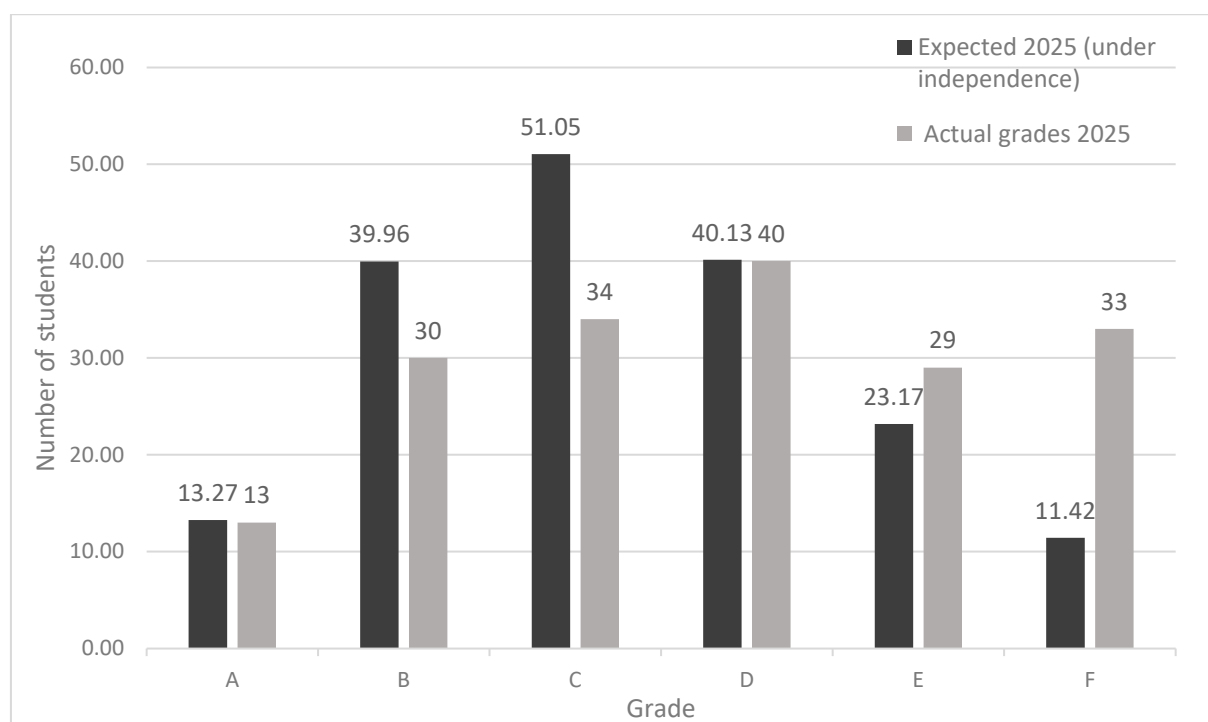


Figure 1. Expected 2025 versus Actual grades 2025.

Table 3. Category-specific contributions to the chi-square test comparing baseline (2021–2024) and 2025 grade distributions.

Grade	Baseline (n)	2025 (n)	Expected 2025	χ^2 Baseline	χ^2 2025
A	66	13	13.3	0.001	0.005
B	208	30	40	0.5	2.48
C	270	34	51.1	1.15	5.69
D	199	40	40.1	0	0
E	109	29	23.2	0.3	1.47
F	35	33	11.4	8.23	40.79
Sum χ^2				10.18	50.44

To ensure that the baseline-versus-2025 result was not driven by any single cohort included in the aggregated baseline, additional chi-square comparisons were conducted between 2025 and each preceding year individually (2021, 2022, 2023, and 2024). As shown in Table 4, the same overall pattern persists across all year-specific comparisons. This supports the interpretation from the baseline table and confirms that 2025 appears as the clear outlier relative to the earlier cohorts.

Cramér's V values ranged from 0.24 to 0.31, indicating small-to-moderate to moderate associations [35]. Because Cramér's V summarizes the association across the full grade distribution, this global effect size should be interpreted alongside the cell-wise chi-square contributions. In the present data, the overall difference is driven predominantly by the failure category (F), which accounts for approximately 81% of the total chi-square statistic, while a smaller contribution is also observed for grade C. Deviations in the remaining grade categories are comparatively modest.

Table 4. Chi-square comparisons of grade distributions between 2025 and earlier cohorts.

Comparison	Chi-Square	df	Cramér's V	p
Baseline 2021–2024 vs. 2025	60.618	5	0.238	<0.001
2021 vs. 2025	36.095	5	0.287	<0.001
2022 vs. 2025	24.294	5	0.244	<0.001
2023 vs. 2025	22.531	5	0.24	<0.001
2024 vs. 2025	34.185	5	0.306	<0.001

5. Discussion

5.1. Credibility and the University's Certification Function

The findings reveal a structural shift in grade distributions that coincides with both rapid student uptake of generative AI and a subsequent restriction of tool access through a change in examination format. Nationally representative survey data indicate that student use of AI-based tools increased sharply between 2023 and 2024, rising from approximately 60% to around 80% during this period [36]. Since publicly accessible large language models were not widely available prior to late 2022, cohorts in 2021–2022 can reasonably be assumed to have completed assessments with limited exposure to contemporary GenAI, with 2023 marking a transitional phase and 2024 a period of widespread use.

Against this timeline, the failure-rate trajectory is substantively meaningful. Failure rates increased modestly between 2021 and 2023, declined in 2024 under the take-home, open-resource format, and then rose sharply in 2025 under a supervised, AI-restricted in-person exam. This sequence is difficult to plausibly attribute to random cohort variation alone. Viewed through the lens of assessment as a socio-technical measurement system [1,2,4], a plausible interpretation is that widespread AI access in 2024 altered the conditions under which assessment evidence was produced, suppressing failure rates by enabling some students to meet criteria through tool-augmented work rather than unaided mastery.

This matters because grades are institutional signals used to certify minimum competence for progression and professional readiness. From a signaling perspective, credibility depends on differentiation of underlying competence in ways that cannot be easily replicated through external mediation. In this course, the results imply that some passing grades under AI-permissive conditions were not robust to AI-restricted conditions. Put plainly, a subset of students who passed under the take-home regime, especially in 2024, may not have passed under AI-restricted conditions. This constitutes a certification credibility problem independent of whether any formal rules were violated.

The distributional structure strengthens this interpretation. The observed changes are not characterized by a uniform downward shift across the grade spectrum. Rather, the largest deviations are concentrated near the threshold: failing grades rise sharply when tool access is restricted, mid-range grades (B and C) decline, and top grades remain stable. This pattern is consistent with the asymmetric benefit and variance-compression mechanisms developed in the theory section [3,4]. In practical terms, GenAI appears to enable marginal students to cross pass thresholds under open conditions, while high-performing students are less affected because their baseline competence already exceeds task requirements.

From a measurement perspective, these patterns point to challenges in maintaining stable and comparable interpretations of assessment results across different assessment contexts. Classical validity theory emphasizes that assessment scores are meaningful only to the extent that they reflect the intended learning outcomes rather than external influences [1,2]. When GenAI contributes to idea generation, reasoning, or linguistic formulation, observed performance can increasingly reflect students' ability to orchestrate

external cognitive resources rather than their independent mastery, introducing construct contamination and weakening comparability across cohorts and formats [3,4]. Finally, the present results align with evidence that GenAI can disproportionately raise outcomes for lower-performing learners and contribute to grade compression, weakening the signal value of grades [11].

5.2. *AI Dependency and the Fragility of Unaided Competence*

The sharp reversion in 2025 under supervised, AI-restricted conditions further refines the interpretation and raises an additional concern: AI dependency. The magnitude of the increase in failing grades suggests that a substantial subset of students, particularly those near the pass threshold, had become reliant on external tool support to produce work meeting grading criteria under open-resource conditions. When external cognitive support was removed, differences in unaided competence became visible in the distributional outcomes, producing substantially higher failure rates and a redistribution of mid-range grades. The stability of top grades across formats is consistent with dependency effects being concentrated in the lower tail of the achievement distribution.

This interpretation should not be framed as a purely individual-level deficit or moral shortcoming. Dependency can be understood as a rational adaptation to the incentive structure of AI-permissive assessment environments. When assessments reward product quality while making authorship and reasoning processes difficult to verify, students have strong incentives to adopt strategies that maximize successful completion, including tool-mediated production. Over repeated cycles, such strategies can reduce the pressure to consolidate foundational knowledge and encourage cognitive offloading. In this sense, dependency is not just a matter of individual behavior; it reflects a structural feature of assessment contexts where powerful cognitive support is widely available but not explicitly accounted for in grading criteria.

From an educational perspective, the dependency hypothesis is important because it suggests a gap between assessment results and lasting learning. When students can repeatedly meet assessment criteria with the help of AI tools, they may advance without developing the independent competence that passing grades are meant to signal. The failure spike observed in 2025 may be interpreted as an empirical manifestation of this gap rather than a simple format effect; it aligns with broader validity concerns that AI-supported performance gains do not necessarily indicate stable learning [1,2,8].

At the same time, the study warrants interpretive caution. The dataset does not include direct measures of individual AI use or the specific ways students integrated tools into their work. Nor does it allow the isolated effect of GenAI restriction to be separated from other changes introduced by the 2025 examination format. The shift from a take-home, open-resource assessment to a supervised, closed-book in-person exam also introduced other conditions that may independently have contributed to lower performance, including stricter time constraints, challenges related to time management, reduced opportunity to look up basic facts or concepts, and higher levels of exam anxiety in a proctored setting. These factors may be especially consequential for students near the pass threshold and could therefore account for part of the observed increase in failure rates. Nonetheless, the combination of prior distributional stability, temporal alignment with national uptake trends, and the immediate reversal under AI-restricted conditions still provides meaningful inferential leverage in a quasi-experimental setting [33,34].

5.3. *Authenticity, Relevance, and the Limits of AI Restriction*

Although the shift to an in-person, AI-restricted examination appears to strengthen fairness and credibility in this course, restricting AI access is not necessarily a universally

desirable solution. A key challenge for higher education is maintaining credible certification while also ensuring that assessments remain authentic and future-relevant. In many disciplines, graduates will be expected to use AI tools in professional practice. If assessment regimes systematically prohibit tools that students will be expected to use responsibly after graduation, assessments may become less authentic indicators of real-world capability [6,8].

This tension suggests that the central issue is not simply whether AI should be “allowed” or “banned,” but whether the assessment system makes explicit, defensible judgments about the competence it intends to certify. If the intended construct is unaided mastery, particularly for threshold competence, then restricting access during examinations may be justified as a way of stabilizing score meaning. Conversely, if effective AI-supported performance is considered a legitimate professional competence, then AI use should be incorporated explicitly into learning outcomes and evaluation criteria. The credibility problem arises most sharply when AI is permitted *de facto* but not acknowledged in construct definitions, grading criteria, or credential claims, creating a mismatch between what assessments appear to measure and what they certify.

From this perspective, supervised, AI-restricted examinations can be understood as a pragmatic corrective that restores interpretability in the short term, especially for threshold decisions. However, a long-term “post-AI” assessment regime likely requires more nuanced designs that are both credible and authentic, for example, assessments that combine evidence of independent competence with structured opportunities for AI use, process documentation, oral components, or staged submissions. While the present study does not adjudicate among design solutions, it provides empirical evidence that treating AI-permissive take-home formats as equivalent to AI-restricted examinations is increasingly difficult to defend [6,8].

6. Limitations and Future Research

Several limitations should be acknowledged. First, the analysis is based on a single course at one institution, which strengthens control over potential confounding factors but limits generalisability. Second, the design is observational; although the format change creates a clear discontinuity, unobserved factors may also contribute. Third, the study does not include direct measures of students’ AI use during the take-home period, so AI-related mechanisms are inferred from timing and distributional patterns rather than directly observed. Fourth, in the final year, the exam removed access to all external resources, not just Gen-AI, while also introducing the psychological pressure of an exam hall.

Future research should combine administrative outcomes with direct measures of AI use and cognitive processes (e.g., surveys, usage traces, drafting histories, oral assessments, or controlled replications) to test mechanisms more directly. Comparative studies across disciplines and task types are needed to identify where AI-permissive formats are most vulnerable to construct instability. Longitudinal studies could examine whether repeated exposure to AI-permissive assessment changes the durability of learning and the stability of performance under AI-restricted conditions, especially around pass–fail thresholds.

7. Conclusions

This study examined how grade distributions changed when an AI-permissive take-home examination was replaced by an AI-restricted in-person examination within the same undergraduate course, holding learning outcomes, course content, the structural design of the examination tasks and grading criteria constant. The results reveal a pronounced and systematic shift following the format change: failure rates increased sharply and mid-range grades were redistributed, while top grades remained stable. The stability of grade

distributions across four years prior to the change suggests a structural break rather than ordinary cohort variation.

The findings imply that take-home and in-person examinations, even when aligned to identical learning objectives, may no longer function as equivalent indicators of student competence under widespread GenAI access. In AI-permissive contexts, outcomes may reflect a hybrid construct combining individual understanding with effective tool use, whereas AI-restricted settings more directly capture independent mastery. As a result, grades awarded under different assessment conditions may carry different meanings, complicating comparisons across cohorts and challenging the credibility of certification when decisions depend on threshold performance.

The study does not suggest that take-home examinations are inherently flawed or should be abandoned. Rather, it indicates that assessment designs must be explicit about the role of AI in what is being measured and certified. As GenAI tools continue to diffuse, evidence on distributional effects of assessment condition changes will be essential for keeping grades valid, interpretable, and aligned with the intended purposes of higher education assessment.

Author Contributions: Conceptualization, H.B., A.U. and M.L.; Methodology, H.B., A.U. and M.L.; Software, H.B., A.U. and M.L.; Validation, H.B., A.U. and M.L.; Formal analysis, H.B., A.U. and M.L.; Investigation, H.B., A.U. and M.L.; Resources, H.B., A.U. and M.L.; Data curation, H.B., A.U. and M.L.; Writing—original draft, H.B., A.U. and M.L.; Writing—review and editing, H.B., A.U. and M.L.; Visualization, H.B., A.U. and M.L.; Supervision, H.B., A.U. and M.L.; Project administration, H.B., A.U. and M.L.; Funding acquisition, H.B., A.U. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors acknowledge the use of ChatGPT-5 (OpenAI) for language refinement and readability improvements.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kane, M.T. Validating the interpretations and uses of test scores. *J. Educ. Meas.* **2013**, *50*, 1–73. [\[CrossRef\]](#)
2. Messick, S. Validity of psychological assessment. *Am. Psychol.* **1995**, *50*, 741–749. [\[CrossRef\]](#)
3. Kabbar, E.; Barmada, B. Assessment validity in the era of generative AI tools. In *Proceedings: CITRENTZ 2023 Conference*; Unitech ePress: Auckland, New Zealand, 2024; pp. 26–33.
4. Kaldaras, L.; Akazeze, H.O.; Reckase, M.D. Developing valid assessments in the era of generative artificial intelligence. *Front. Educ.* **2024**, *9*, 1399377. [\[CrossRef\]](#)
5. Gruenhagen, J.H.; Sinclair, P.M.; Carroll, J.A.; Baker, P.R.; Wilson, A.; Demant, D. The Rapid Rise of Generative AI and Its Implications for Academic Integrity: Students' Perceptions and Use of Chatbots. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100273. [\[CrossRef\]](#)
6. Kizilcec, R.F.; Huber, E.; Papanastasiou, E.C.; Craw, A.; Makridis, C.A.; Smolansky, A.; Zeivots, S.; Radulescu, C. Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100269. [\[CrossRef\]](#)
7. Rudolph, J.; Tan, J.; Tan, E. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *J. Appl. Learn. Teach.* **2023**, *6*, 342–363. [\[CrossRef\]](#)
8. Xia, Q.; Weng, X.; Ouyang, F.; Lin, T.J.; Chiu, T.K.F. A scoping review on how generative artificial intelligence transforms assessment in higher education. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 40. [\[CrossRef\]](#)

9. Bittle, K.; El-Gayar, O. Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information* **2025**, *16*, 296. [\[CrossRef\]](#)
10. Deng, R.; Jiang, M.; Yu, X.; Lu, Y.; Liu, S. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Comput. Educ.* **2025**, *227*, 105224. [\[CrossRef\]](#)
11. Wang, J.; Fan, W. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 621. [\[CrossRef\]](#)
12. Ilieva, G.; Yankova, T.; Ruseva, M.; Kabaivanov, S. A framework for generative AI-driven assessment in higher education. *Information* **2025**, *16*, 472. [\[CrossRef\]](#)
13. Li, Y.; Shan, Z.; Raković, M.; Guan, Q.; Gašević, D.; Chen, G. When AI explains in natural language: Unveiling the impact of generative AI explanations on educators' grading and feedback practices. *Educ. Inf. Technol.* **2025**, *30*, 24931–24964. [\[CrossRef\]](#)
14. Kofinas, A.K.; Tsay, C.H.; Pike, D. The impact of generative AI on the academic integrity of authentic assessments within higher education. *Br. J. Educ. Technol.* **2025**, *56*, 2522–2549. [\[CrossRef\]](#)
15. Smolansky, A.; Cram, A.; Radulescu, C.; Zeivots, S.; Huber, E.; Kizilcec, R.F. Educator and student perspectives on the impact of generative AI on assessments in higher education. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*; ACM: New York, NY, USA, 2023; pp. 378–382.
16. Moongela, H.; Matthee, M.; Turpin, M.; van der Merwe, A. The Effect of Generative Artificial Intelligence on Cognitive Thinking Skills in Higher Education Institutions: A Systematic Literature Review. In *Southern African Conference for Artificial Intelligence Research*; Springer Nature: Cham, Switzerland, 2024; pp. 355–371.
17. Burnett, L.K.; Richmond, L.L. Just write it down: Similarity in the benefit from cognitive offloading in young and older adults. *Mem. Cogn.* **2023**, *51*, 1580–1592. [\[CrossRef\]](#)
18. Grinschgl, S.; Meyerhoff, H.S.; Schwan, S.; Papenmeier, F. From metacognitive beliefs to strategy selection: Does fake performance feedback influence cognitive offloading? *Psychol. Res.* **2021**, *85*, 2654–2666. [\[CrossRef\]](#)
19. Risko, E.F.; Gilbert, S.J. Cognitive offloading. *Trends Cogn. Sci.* **2016**, *20*, 676–688. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Weis, P.P.; Wiese, E. Problem solvers adjust cognitive offloading based on performance goals. *Cogn. Sci.* **2019**, *43*, e12802. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Iqbal, J.; Hashmi, Z.F.; Asghar, M.Z.; Abid, M.N. Generative AI tool use enhances academic achievement in sustainable education through shared metacognition and cognitive offloading among preservice teachers. *Sci. Rep.* **2025**, *15*, 16610. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Zhang, L.; Xu, J. The paradox of self-efficacy and technological dependence: Unraveling generative AI's impact on university students' task completion. *Internet High. Educ.* **2025**, *65*, 100978. [\[CrossRef\]](#)
23. Backfisch, I.; Lachner, A.; Stürmer, K.; Scheiter, K. Variability of teachers' technology integration in the classroom: A matter of utility! *Comput. Educ.* **2021**, *166*, 104159. [\[CrossRef\]](#)
24. Huang, S.; Lai, X.; Ke, L.; Li, Y.; Wang, H.; Zhao, X.; Dai, X.; Wang, Y. AI technology panic—Is AI dependence bad for mental health? A cross-lagged panel model and the mediating roles of motivations for AI use among adolescents. *Psychol. Res. Behav. Manag.* **2024**, *17*, 1087–1102. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Atas, A.H.; Yıldırım, Z. A shared metacognition-focused instructional design model for online collaborative learning environments. *Educ. Technol. Res. Dev.* **2025**, *73*, 567–613. [\[CrossRef\]](#)
26. Singh, C.A.; Muis, K.R. An integrated model of socially shared regulation of learning: The role of metacognition, affect, and motivation. *Educ. Psychol.* **2024**, *59*, 177–194. [\[CrossRef\]](#)
27. Diaz, B.; Chen, G.; Jaselskis, E.; Delgado, C. Supporting Generative AI Literacy: Exploring the Pedagogical Roles Students Assign ChatGPT and Impact on Course Grades. *Comunicar* **2025**, *33*, 46–61.
28. Vanacore, K.; Pankiewicz, M.; Baker, R. Unpacking the Impact of Generative AI Feedback: Divergent Effects on Student Performance and Self-Regulated Learning. Available online: https://osf.io/preprints/edaxiv/tbpn3_v1 (accessed on 1 April 2026).
29. Xie, Y.; Luo, L. The impact of generative AI on learning across grades. In *Proceedings of the 2025 14th International Conference on Educational and Information Technology (ICEIT)*, Guangzhou, China, 14–16 March 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 91–95. [\[CrossRef\]](#)
30. Elmourabit, Z.; Retbi, A.; El Faddouli, N.E. The Impact of Generative Artificial Intelligence on Education: A Comparative Study. In *European Conference on e-Learning*; Academic Conferences International Limited: Reading, UK, 2024; Volume 23, pp. 470–476.
31. Spence, M. Job market signaling. *Q. J. Econ.* **1973**, *87*, 355–374. [\[CrossRef\]](#)
32. Cook, T.D.; Campbell, D.T. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*; Houghton Mifflin: Boston, MA, USA, 1979.
33. Shadish, W.R.; Cook, T.D.; Campbell, D.T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*; Houghton Mifflin: Boston, MA, USA, 2002.
34. Imbens, G.W.; Rubin, D.B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*; Cambridge University Press: Cambridge, UK, 2015.

35. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1988.
36. NOKUT—Nasjonalt Organ for Kvalitet i Utdanningen. *Fire av Fem Studenter Bruker Kunstig Intelligens i Studiene* [News Release]; NOKUT: Oslo, Norway, 10 February 2025. Available online: <https://www.nokut.no/nyheter/fire-av-fem-studenter-bruker-kunstig-intelligens-i-studiene/> (accessed on 1 April 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.